View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Edinburgh Research Archive

THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

Discovering and analysing lexical variation in social media text

Philippa Shoemark

Doctor of Philosophy

Institute for Language, Cognition and Computation
School of Informatics
The University of Edinburgh

2020

Abstract

For many speakers of non-standard or minority language varieties, social media provides an unprecedented opportunity to write in a way which reflects their everyday speech, without censorship or castigation. Social media also functions as a platform for the construction, communication, and consolidation of personal and group identities, and sociolinguistic variation is an important resource that can be put to work in these processes. The ease and efficiency with which vast social media datasets can be collected make them fertile ground for large-scale quantitative sociolinguistic analyses, and this is a growing research area. However, the limited meta-data associated with social media posts often makes it difficult to control for potential confounding factors and to assess the generalisability of results.

The aims of this thesis are to advance methodologies for discovering and analysing patterns of sociolinguistic variation in social media text, and to apply them in order to answer questions about social factors that condition the use of Scots and Scottish English on Twitter. The Anglic language varieties spoken in Scotland are often conceptualised as a continuum extending from Scots at one end to Standard English at the other, with Scottish English in between. There is a large degree of overlap in grammar and vocabulary across the whole continuum, and people fluidly shift up and down it depending on the social context. It can therefore be difficult to classify a short utterance as unequivocally Scots or English. For this reason we focus on the lexical level, using a data-driven method to identify words which are distinctive to tweets from Scotland. These include both centuries-old Scots words attested in dictionaries, and newer forms not yet recorded in dictionaries, including innovative variant spellings, contractions, and acronyms for common Scottish turns of phrase.

We first investigate a hypothesised relationship between support for Scottish independence and distinctively Scottish vocabulary use, revealing that Twitter users who favoured hashtags associated with support for Scottish independence in the lead up to the 2014 Scottish Independence Referendum used distinctively Scottish lexical variants at higher rates than those who favoured anti-independence hashtags. We also test the hypothesis that when specifically discussing the referendum, people might increase their Scots usage in order to project a stronger Scottish identity or to emphasise Scottish cultural distinctiveness, but find no evidence to suggest this is a widespread phenomenon on Twitter. In fact, our results indicate that people are significantly more likely to use distinctively Scottish vocabulary in everyday chitchat on Twitter than when discussing Scottish independence. We build on the methodologies of previous

large-scale studies of style-shifting and lexical variation on social media, taking greater care to avoid confounding form and meaning, to distinguish effects of audience and topic, and to assess whether our findings generalise across different groups of users.

Finally, we develop a system to identify pairs of lexical variants which refer to the same concepts and occur in the same syntactic contexts; but differ in form and signal different things about the speaker or situational context. Our aim is to facilitate the process of curating sociolinguistic variables by providing researchers with a ranked list of candidate variant pairs, which they only have to accept or reject. Data-driven identification of lexical variables is particularly important when studying language varieties which do not have a written standard, and when using social media data where linguistic creativity and innovation is rife, as the most distinctive variables will not necessarily be the same as those that are attested in speech or other written domains. Our proposed system takes as input an unlabelled text corpus containing a mixture of language varieties, and generates pairs of lexical variants which have the same denotation but differential associations with two language varieties of interest. This can considerably speed up the process of identifying pairs of lexical variants with different sociocultural associations, and may reveal pertinent variables that a researcher might not have otherwise considered.

Acknowledgements

I first want to acknowledge how incredibly fortunate I am to have found such competent, conscientious, and compassionate supervisors. Sharon and James both stepped up when I needed them to in difficult times, but also knew when to step back and allow me the time and space to develop my ideas independently. They were always quick to provide clear and constructive feedback when I asked for it, and they both were amazingly adept at understanding the crux of my roughly formulated questions and arguments, and patiently helping me to clarify and refine them. I am especially grateful to Sharon for applying just enough pressure to keep me on track, while always remaining flexible, accommodating, and empathetic. Thank you Sharon and James, you really made all the difference.

Thank you also to my examiners, Brendan O'Connor and Walid Magdy, for taking the time to carefully read this thesis, for helping to improve it with your constructive comments, and for an engaging and insightful viva discussion. Many thanks as well to Lauren Hall-Lew and Dong Nguyen for all your valuable feedback in my annual reviews. Dong, I cannot over-emphasise how grateful I am to you for all the networking and development opportunities you have extended to me over the course of my PhD—you have been a great mentor and an inspiration! I also want to thank Scott Hale and Barbara McGillivray for their great mentorship; I somehow hit the supervision jackpot again with my internship at the Turing Institute, and it really helped to boost my confidence and motivation to complete my PhD.

I'm grateful to the EPSRC for funding my research, and I'm grateful to have been able to benefit from belonging to both the CDT in Data Science and the ILCC. I have learned an enormous amount from these two communities of amazing staff and students, and had a lot of fun along the way. I want to thank the CDT administrators Brittany Bovenzi, Sally Galloway, and Siobhan Carroll for working so hard to make things easy for us.

Thank you to all of the wonderful friends I made throughout my PhD in the CDT, ILCC, elsewhere in Informatics and beyond. I owe so much to so many of you, but special mentions go out to Amy Isard, Arlene Casey, Janie Sinclair, Margaret Ritchie, and Nathalie Dupuy, for having been extremely helpful coffee-and-study-buddies as well as wonderful friends. Thanks to everyone I shared room 3.50 with for being such lovely office mates, and thanks to Adam Lopez and the AGORA group for welcoming me into your midst and offering helpful feedback on my paper drafts. Thanks especially to Sameer Bansal for tirelessly practicing talks with me, and for helping me

with viva preparation. And of course, massive thanks to Alex Robertson for taking the pressure off me to satisfy Sharon's insatiable appetite for Twitter research 😜

Thanks so much to Connor King, Hannah Daly, Raymah Tariq, Samantha Myers, and Sarah Cee for providing much needed moral support, and huge thanks to my family for your heroic efforts in not asking too often about how the writing was going. Thanks most of all to Samuel Abreu, for being so patient, so kind, and so supportive all the way through.

Lay summary

Social media provides a new platform for the construction of personal and group identities. In particular, it affords an unprecedented opportunity for speakers of non-standard or minority language varieties to write in a way which reflects their everyday speech. The ease and efficiency with which vast social media datasets can be collected make them convenient for use in large-scale sociolinguistic analyses. While this is a growing area of research with many technological applications, the limited meta-data associated with social media posts often makes it difficult to assess the generalisability of results inferred from these analyses.

The main aim of this thesis is to contribute to overcome this shortcoming and advance methodologies for discovering and analysing patterns of sociolinguistic variation in social media text. As an application of these methods, we study how social factors condition the use of Scots and Scottish English on Twitter. For instance, by investigating the relationship between support for Scottish independence and distinctively Scottish vocabulary use on Twitter, we find that users who favoured hashtags associated with support for Scottish independence in the lead up to the 2014 Scottish Independence Referendum used Scottish lexical variants at higher rates than those who favoured anti-independence hashtags. Interestingly, however, we also find the use of Scottish vocabulary to be more prevalent in everyday chitchat on Twitter than when discussing Scottish independence. Finally, we develop an automatic computational tool that takes as input text that mixes different language varieties, and generates associations betweens words from each language that have the same meaning. This tool can considerably speed up the process of identifying pairs of lexical variants with different sociocultural associations, and reveal pertinent variables that a researcher might not have otherwise considered.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Philippa Shoemark*)

Table of Contents

Chapter 1

Introduction

This thesis is concerned with patterns of lexical variation on Twitter, and attempts to account for them in terms of factors relating to the situational context. Large social media datasets can be used to identify linguistic features which are statistically associated with particular geographical regions, demographic traits, ideological stances, or other sociocultural factors. Once features that index particular sociocultural groups or ideologies have been identified, we can ask: in what situational contexts are individuals more likely to use them? The scale and richness of social media data enables us to discover sociolinguistic associations in a bottom-up, data-driven fashion, rather than relying on pre-conceived ideas about which features or groups to consider. It also enables us to test predictions of sociolinguistic theories on a large scale.

Social media writing is often stylistically distinct both from other written genres, and from speech. Nevertheless, similar effects to those established in traditional sociolinguistic studies of speech have been reported in studies of social media writing. The use of non-standard words and spellings on Twitter mirrors patterns of variation in speech, both in terms of their geographical and demographical distributions (Eisenstein et al., 2014; Huang et al., 2016), and in terms of the phonological and syntactic contexts in which they occur (Eisenstein, 2015). Twitter users appear to be cognisant of who is likely to see their tweets and sensitive to the social meaning of linguistic variables, since they appear to modulate their linguistic choices accordingly. For example, individuals with a greater proportion of same-gender ties in their social networks make greater use of gender-marked variables (Bamman et al., 2014b), and there is evidence for effects of audience size on usage rates of non-standard and minority language varieties in tweets from the USA (Pavalanathan and Eisenstein, 2015a) and the Netherlands Nguyen et al. (2015).

This type of research is often complicated by the lack of detailed meta-data associated with social media posts, which can make it hard to control for different explanatory factors and to know whether results obtained on a particular user sample generalise to another sample. Another outstanding methodological challenge in this area is the bottom-up discovery of sociolinguistic variables.

Traditionally, a linguistic variable is any linguistic item than can be realised in different ways: the set of alternative realisations are the variants. A *socio*linguistic variable is one whose different variants are associated with different social identities or ideologies. In contrast with traditional sociolinguistic studies, most large-scale social media studies to date have studied differences across contexts in the frequencies of individual terms (Doyle, 2014; Bamman et al., 2014b; Jones, 2015; Pavalanathan and Eisenstein, 2015a). If we instead analyse the *relative* frequencies of different realisations of the same variable, we can be more confident that any effects we observe are effects on *how* people are choosing to refer to things, and not on *which* things they are choosing to refer to.

In this thesis we present two large-scale studies of factors which condition lexical variation in Scottish tweets, and a system to facilitate efficient, data-driven curation of lexical sociolinguistic variables. In the first of these, presented in Chapter 4, we investigated the relationship between support for Scottish independence and the use of distinctively Scottish vocabulary on Twitter, and found that distinctively Scottish lexical variants were used at a higher rate by users of pro-independence hashtags than by users of anti-independence hashtags. However, some questions remained about how rates of Scottish variant usage are affected by the topic of a tweet and/or the size of its expected audience. We addressed these open questions in our second study, presented in Chapter 5. We used a more sophisticated method of analysis to model effects of audience size and topic on the use of Scotland-specific variants, whilst controlling for variation in the base rate of Scottish variant usage across different users and variables. We looked at two groups of users with different overall rates of Scottish usage, and found audience size and topic to have independent effects on Scottish variant usage in both groups. The qualitative effects of topic were similar across the two user groups, demonstrating a clear relationship between the topic of discussion and the odds of choosing Scottish variants. However, the sizes and directions of the audience effects were inconsistent across the two groups.

Both of these studies improved upon prior related work by defining the dependent variable in terms of lexical alternations, and thus better controlling for differences in

what is being referred to, as opposed to *how*. We selected the lexical alternations to analyse by first using a data-driven method to identify distinctively Scottish terms, and then manually pairing these with Standard English equivalents. The manual pairing process was labour intensive and required a high degree of familiarity with both language varieties, which motivated us to devise a system to facilitate this process.

The task of identifying lexical alternations is similar to bilingual lexicon induction, but more challenging than the typical case where separate monolingual corpora are available for the two languages in question, as we want to be able to identify variables whose variants belong to closely related language varieties including minority languages, dialects, or sociolects for which monolectal corpora are not necesarily available and are far from straightforward to create. We therefore developed a system to identify lexical variables from a single code-mixed corpus, without any pre-specified labels indicating which linguistic variety a given span of text belongs to. Our system, described in full in Chapter 6, takes as input a code-mixed corpus and a small set of seed variables, and returns a ranked list of additional candidate variables. This facilitates the process of curating sociolinguistic variables considerably, as researchers only have to accept or reject candidate variant pairs from a ranked list. In experiments on three different pairs of English dialects or related language varities, we demonstrate useful results over a range of hyperparameter settings, with precision@100 of over 70% in some cases using as few as five seed pairs.

1.1 List of main research questions and contributions

For convenience, we list here the primary research questions addressed in Chapters 4 and 5, and summarise our main contributions below.

RQ1: Were Twitter users who supported independence more likely to use distinctively Scottish variants in their tweets than those who opposed it?

RQ2: Are there independent effects of tweet topic and audience on the use of distinctively Scottish variants?

RQ3: Are topic and audience effects consistent across users sampled on the basis of having used hashtags relating to Scottish independence, and users sampled on the basis of having tweeted with a Scottish geotag?

The main contributions of this thesis are as follows:

- We establish that both centuries-old Scots words attested in dictionaries, and Scottish lexical innovations not yet recorded in dictionaries, are in use on Twitter.
- We show that while users who supported independence used distinctively Scottish lexis at higher rates than those who opposed it; it was used by both groups, and both groups were more likely to use it in everyday chitchat than when discussing politics.
- We build on the methodologies of previous large-scale studies of lexical variation on social media, taking greater care to:
  - Avoid confounding form and meaning
  - Distinguish effects of audience and topic
  - Assess whether findings generalise across different user samples
- We develop a new data-driven, computational method to facilitate the identification of lexical variables from code-mixed text.

1.2 Structure of the thesis

– Chapter 2 provides background on social media text, Scots and Scottish English, and our motivations for analysing sociolinguistic variation at the lexical level, as well as the challenges involved.

– Chapter 3 provides background about the Twitter social media platform, ethical issues surrounding the use of Twitter data in research studies, and methodological challenges involved in the use of Twitter data in social scientific analyses. Chapter 3 also includes an overview of all the datasets we collected and used in our studies.

– Chapters 4-6 are each centered on a published paper. Each includes the paper as published, along with an additional introduction and commentary section to motivate the work and evaluate its contributions in the context of the thesis, as well as proposing specific areas for future work pertaining to the topic of the chapter.

- Chapter 4 is about Scots and Scottish English use on Twitter in the context of the 2014 Scotland independence referendum, and is based on the following paper:

  Shoemark, P., Sur, D., Shrimpton, L., Murray, I., & Goldwater, S. (2017, April). Aye or naw, whit dae ye hink? Scottish independence and linguistic identity on social media. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers* (pp. 1239-1248).

- Chapter 5 is about effects of topic and prospective audience size on Scots and Scottish English use on Twitter, and is based on the following paper:

  Shoemark, P., Kirby, J., & Goldwater, S. (2017, September). Topic and audience effects on distinctively Scottish vocabulary usage in Twitter data. In *Proceedings of the Workshop on Stylistic Variation* (pp. 59-68).

- Chapter 6 is about an automatic method to facilitate the curation of lexical alternation variables for use in sociolinguistic studies, and is based on the following paper:

  Shoemark, P., Kirby, J., & Goldwater, S. (2018, November). Inducing a lexicon of sociolinguistic variables from code-mixed text. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text* (pp. 1-6).

– Chapter 7 concludes this thesis with a more detailed summary of our contributions and a discussion of directions for future work which could build on these.

Chapter 2

Background

In this thesis we develop and apply methodologies for conducting quantitative variational sociolinguistic analyses of social media text. We focus in particular on variation in the use of lexical items which are distinctive to Scots and Scottish English, though our approach is broadly applicable to other language varieties. In Section 2.1, we discuss the unqiue constellation of properties which sets social media text apart from both speech and other written domains, and argue that these properties make sociolinguistic variation a particularly important resource for signalling attitudes and identity on social media.

In Section 2.2 we discuss the complex relationship between the English and Scots languages, historical and ongoing changes in attitudes towards the Scots language, and how these lead to difficulties in categorizing texts as wholly Scots or English. In Section 2.3 we discuss some methodological challenges in analysing linguistic variation on social media, particularly in the use of minority languages which do not have established written norms.

2.1 Social media text

Social media platforms with public APIs provide efficient access to high volumes of spontaneously produced, informal, conversational text from large and diverse user bases. The ability to efficiently collect such massive amounts of natural language data enables the study of infrequent linguistic phenomena, whilst avoiding the Observer's Paradox.

The OBSERVER'S PARADOX is a term coined by William Labov to describe the predicament of trying to systematically observe how people use language when they

are not being systematically observed. Labov (1978b, p.209) argues that people tend to pay greater attention to their speech when aware they are being observed by researchers; and when paying greater attention to their speech they are more likely to style-shift towards more formal or prestigious linguistic varieties. Hence he advocates collecting 'rapid and anonymous observations' where possible, which requires creative approaches when vernacular speech is the object of study, but can more easily be done on a vast scale with social media text.

While not always complete or verifiable, the meta-data that often accompanies social media text—such as time-stamps, locations, social network connections, names, and photographs—makes it possible to conduct fine-grained analyses of socio-cultural correlates of linguistic variation and change, using large-scale, aggregated data sets (Hovy et al., 2015; Blodgett et al., 2016; Jones, 2015; Grieve et al., 2018). These large-scale data sets can also be used to identify social categories and sociolinguistic variables in a transparent, bottom-up manner, rather than relying on researcher intuition or pre-conceived notions about which features or groups to consider (Eisenstein et al., 2011; Bamman et al., 2014b; Huang et al., 2016)

From an engineering perspective, better understanding the dynamics of linguistic variation in social media text can help to improve performance in various tasks such as document classification (Hovy, 2015), sentiment analysis (Sánchez-Rada and Iglesias, 2019), and hate-speech detection (Mishra et al., 2019).

Social media text is also an interesting object of study in its own right, since distinctive characteristics of social media interactions may influence users' choices about what to say and how to say it in distinctive ways, resulting in different patterns than those which are found in other modes of language.

2.1.1 Comparison with speech and other written mediums

Social media differs from speech and other written media in several ways, first discussed in depth by boyd (2008), who characterised 'networked publics' (i.e. social media) as having four properties—persistence, replicability, searchability, and scalabity—which complicate and shape the ways in which their users interact.

While writing in general is inherently persistent, and recording devices allow people to make persistent recordings of speech, we do not typically record our everyday spoken conversations, whereas social media automatically archives each and every casual interaction. Not only is every act on social media persistently recorded by default,

due to their digital nature these records are also easily replicable and searchable, and scalable in that they can be disemminated to vast numbers of people all around the world in an instant.

Though television and radio share the property of scalability, their content is usually designed specifically for public consumption, whereas social media blurs the boundaries between public and private interactions. Moreover, while the content that is distibuted on TV and radio is centrally controlled by broadcasting organisations, social media is decentralised, such that "the property of scalability does not necessarily scale what individuals want to have scaled or what they think should be scaled, but what the collective chooses to amplify" (boyd, 2008, p.33).

Together, boyd argues, the properties of persistence, replicability, searchability, and scalabity result in a lack of spatial, social, and temporal boundaries on social media. While in the offline world, one would typically present oneself differently depending on the social context, on social media it is often difficult to control or anticipate the context in which a message will be seen or interpeted, and social contexts which are typically segmented in the offline world are collapsed together in networked publics.

Although we might have a good idea of who is likely to read and engage with the content we post on social media, we often have no way of knowing exactly who will see a particular post, or when they will see it. For example, on Twitter, posts are public by default. While there are affordances to target posts towards particular groups or individuals, unless a user has actively set their profile to private so that only those who have explicitly 'followed' them can see their tweets, then in principle anyone on the internet could potentially stumble across any of that user's tweets, at any time after it has been posted. However, the fact that a tweet *could* be seen by anyone does not mean that it *will* be seen by everyone. The sheer volume of content in social media streams "creates an attention economy in which people must compete for visibility." (boyd, 2008, p.32)

To investigate how Twitter users navigate this attention economy with its collapsed contexts and invisible audiences, Marwick and boyd (2011) interviewed a sample of Twitter users about how they imagine the audience of their tweets, and how that factors into what they choose to write. Participants indicated that while they have no way of knowing exactly who comprises the audience of a given tweet, they do have a mental picture of who they're addressing, and they modulate the content of their tweets according to the imagined judgement of this imagined audience. Based on these interviews, Marwick and boyd argued that the lack of control over context on social media

creates a tension between the desire to maintain positive impressions (keeping the most sensitive potential readers in mind, such as parents, partners, and bosses)—and the desire to seem true or authentic to others. This pressure to appear authentic is reinforced by the social norms of the Twitter platform, which encourage the sharing of personal information in order to build and maintain social connections.

The indeterminacy of the audience of a given post, and the fact that it can be replicated and retrieved at any time and divorced from its original context, may give users an additional impetus to index information about their identity, attitudes, and communicative goals within the content of the post itself. Since brevity is enforced on Twitter by a character limit (and encouraged on social media more generally by the attention economy), lexical choice (as opposed to more expositional strategies) is arguably a key resource with which users can efficiently convey (or obscure) such information within their posts. Moreover, relaxed written norms on social media give users a license to experiment with ways of representing sociophonetic variation orthographically (Eisenstein, 2015; Jones, 2015; Tatman, 2015), and the replicability and scalibility properties of social media have lead to some minority language varieties becoming more visible to people outside of their offline communities of practise. For instance, Florini (2014) has analysed how in the absence of reliable physical cues to racial identity on Twitter, many Black American Twitter users align themselves with Black cultural identities through the use of particular kinds of verbal dexterity, wit, and wordplay which mirror the practise of SIGNIFYIN' in African-American oral culture. This includes the use of phonetic re-spellings, slang terms, and textual representations of gesture (e.g. emoticons) to convey an elaborate oral delivery style. Florini (2014, p. 225) notes that

> "The activity of Black Twitter users has not been lost on many bloggers and journalists, who have generated much discussion and debate about the existence and nature of 'Black Twitter."

In summary, social media is different from other written domains because it is informal but not typically private; and different from informal speech because it is inherently persistent, replicable and searchable, with an 'invisible' audience which likely comprises a mixture of people with whom one would traditionally interact in very different social contexts. Social media can be conceived of as a stage for the construction, presentation, reinforcement, and renegotiation of personal and group identities, and the ability to mine vast quantitities of social media data in real-time provides unprecedented opportunities to observe how linguistic variation is put to work in these practices.

2.1.2 Linguistic variation as a means to negotiate networked audiences

Whereas Marwick and boyd (2011) focused mainly on the choices users make about *what* to say in their tweets, this thesis is concerned with the choices that users make about *how* to say it: the choices they make about linguistic *style*, or specifically, their choice of language variety or register. In a closely related ethnographic study, Androutsopoulos (2014) examined how multingual users of the Facebook social media platform use their choice of language to 'maximise' or 'partition' the audience of their posts, as well as how audience members respond to these choices. On analogy with Marwick and boyd's observation that some social media users limit the topics of their posts to those which they believe will be non-offensive to the broadest possible audience, Androutsopolous noted that posts can also be *styled* in a way that makes them accessible to as many members of a user's network as possible. Similarly, posts can be tacitly directed towards narrower audiences not only by carefully choosing particular types of *content* to include, but also by tailoring the choice of *language*: "language style, and more specifically language choice, becomes a key resource by which to bring together or separate various parts of the networked audience" (Androutsopoulos, 2014, p.71).

In a large-scale quantitative analysis of the language choices of Twitter users in Friesland and Limburg (two provinces in the Netherlands where minority languages are spoken as well as Dutch), Nguyen et al. (2015) provided further evidence that social media users tailor their language choices with respect to their audiences. They found that while most tweets from these provinces were written in Dutch, users were more likely to use a minority language when replying directly to other tweets in which it had been used, when using hashtags which had previously been used in lots of other minority language tweets, and when addressing another user who had frequently posted in the minority language.

While Androutsopoulos (2014) and Nguyen et al. (2015) looked at language choices on the level of utterances or posts, Pavalanathan and Eisenstein (2015a) focused on the lexical level. They looked at English language tweets in the US, and found that Twitter users were more likely to use non-standard terms—both those which are geographically specific (e.g. *jawns*, *ole*, *dang*) and those which are frequently used across the USA (e.g.*lol*, *omg*, *pics*)—in tweets directed at smaller, geographically closer audiences.

In a similar vein to Nguyen et al. (2015) and Pavalanathan and Eisenstein (2015a), we present in Chapters 4 and 5 two large-scale statistical analyses of factors that influence minority language use on social media. We are interested in methodologies which can be applied to cases where it may be difficult to classify a post as belonging entirely to one language variety or another, such as when there is intra-sentential code-switching, or when the varieties in question share a lot of syntax and vocabulary. We have therefore chosen to follow Pavalanathan and Eisenstein (2015a) in analysing variation at the lexical level. We focus on Scots and Scottish English, and consider effects of topic and political stance, in addition to audience.

2.2 Scots and Scottish English

This thesis is mostly concerned with questions about factors which influence people's use of Scots vocabulary on social media, though the methods we apply and develop here are more broadly applicable to lexical variation across other linguistic codes. Scotland has been described as something of a 'paradise' for sociolinguists (Görlach, 1985), owing to its rich history and diversity of interconnected dialects and sociolects. There are three historical indigenous languages that are spoken in modern-day Scotland: English, Scots, and Gaelic. While Gaelic is a Celtic language, Scots has Anglo-Scandinavian origins in common with English. The standard variety of English in Scotland differs from that spoken South of the Scottish border mostly in pronunciation, but also has some distinctive vocabulary, idioms, and syntactic constructions.

2.2.1 Evolving attitudes towards Scots

At different periods in Scotland's history, both Scots and Gaelic have enjoyed a status as the language of prestige, but both were subsequently marginalised, and throughout the 18th and 19th centuries even strongly and explicitly stigmatised in favour of English.

In the case of Scots, not only did it lose its prestige after the poltical union of Scotland and England, but it came to be seen by its own native speakers as a corrupted form of English, rather than a language in its own right (Purves, 2002). This attitude persisted throughout the 20th century: for example the Scottish Education Department's official language policy on Scots in the 1940s was that it was "not the language of 'educated' people anywhere, and could not be described as a suitable medium of edu-

cation or culture" (quoted by Aitken (1982,2015)). As recently as thirty years ago, the Scottish writer and broadcaster Billy Kay (1988, p.151) observed:

> "I have heard many Scots speakers say that they are only comfortable talking Scots to someone from the same locality. With everyone conditioned to some extent by official disdain for the tongue, it takes a strong person to speak Scots in a formal situation where people may classify them according to one or other stereotype as coarse or uneducated; it is also much simpler to speak English and save yourself the hassle."

However, the cultural climate in Scotland has begun to evolve since then, with the devolution of the Scottish parliament in 1997 and the 2014 independence referendum bolstering political engagement and a sense of national pride. In tandem with the growth of the Scottish Nationalist movement, momentum has grown behind efforts to recognise, preserve, and promote the Scots and Gaelic languages. For the first time, the census of Scotland in 2011 asked about Scots language skills, and over 1.5 million people in Scotland (30% of the population) reported that they could speak it (Scots Language Centre, 2013). The previous year, the Scottish Goverment had surveyed a representative sample of the adult population in Scotland about their attitudes towards the Scots language, and more than 80% agreed that it plays an important and valuable part in their heritage, culture, and identity (Scottish Government, 2010).

Against this backdrop of cultural heritage, historical repression, and resurgence lead by Scottish nationalists, it seems almost inevitable that choices regarding the use of Scotland's languages carry political and/or cultural connotations. This inspired us to investigate usage rates of Scots vocabulary on social media, and their interactions with stances on Scottish nationalism, as well as the type of audience, and the topic under discussion.

2.2.2 Difficulties in delimiting distinctively Scottish vocabulary

Defining 'use of Scots vocabulary' is complicated by the fact that Scots shares a large stock of core vocabulary with English, due in one part to their common ancestry, and also due to having undergone similar patterns of language contact in the Medieval Period (Macafee, 2003). Given that non-standard spellings which reflect regional pronunciations are a common feature in social media text (Eisenstein, 2015), it can be difficult to conclusively say whether a particular token such as *baw* should be classified as an instance of a Scots word which is cognate with English *ball*, or as a phonetic re-spelling representing a Scottish pronunciation of the English word itself.

Because of the complex intersections among the various Anglo-Scandinavian-derived language varieries in current use in Scotland, sociolinguists often describe them as a continuum extending from Scots at one end to Standard English at the other. Most native Scots are able to code-mix and style-shift up and down this continuum depending on the situational context (Stuart-Smith, 2003). In the work presented in this thesis, rather than trying to discrestise this continuum or place tweets or word tokens at specific points along it, we instead focus on investigating how speakers of Scots and/or Scottish English modulate their use of *identifiably Scottish* vocabulary, i.e. lexical variants which are associated with some variety of Scots or Scottish English, and *not* shared with Standard British English.

A similar approach was taken by the makers of the Scottish National Dictionary (SND) (Grant and Murison, 1931), which documents modern Scots vocabulary and is available online as part of the Dictionary of the Scots Language (Scottish Language Dictionaries, 2004). However, as it was last updated in 2005 and is primarily based upon literary sources, we cannot rely upon the SND as a reference point for identifing distinctively Scottish lexis in the novel domain of social media text (bearing in mind that Facebook and Twitter were only launched globally in 2006.)

It is important to note that like English, Scots consists of many different local dialects, but unlike English it has no established standard orthography. Moreover, while Scots has remained a living language in its spoken form, before the advent of social media the average Scottish person's exposure to written Scots would have been largely confined to literary domains such as poetry, folk songs, fictional dialogue, or comic narrative (Grant, 1931, §18). Scots usage in these domains is often romanticised or stereotyped and not truly representative of any variety of contemporary spoken Scots. For many people in Scotland, then, social media may be providing an arena to read and write in a way that reflects the way they speak for the first time.

In summary, given (1) the wide range of regional variation in spoken Scots, (2) the lack of any single, widely-recognised orthographic standard, and (3) the fact that creative and non-standard use of orthography is pervasive on social media; if we want to gain an accurate and thorough understanding of how distinctively Scottish lexis is used on social media, we should not base our analysis on outdated or out-of-domain dictionaries or surveys. Instead, the analyses in this thesis are based on bottom-up, data driven approaches, which reveal innovative variant spellings, derived forms, and recent neologisms that would have been systematically excluded had we relied on pre-existing lexicons.

2.3 Lexical variation

Broadly speaking, sociolinguistics is concerned with relating linguistic structure to social structure, or understanding how social and stylistic factors constrain linguistic variation. A sociolinguistic variable is a set of alternate linguistic forms which have the same denotational meaning, but differ with respect to the social identities or attitudes they connote. For example, in his famous department store study, Labov (1978b, Ch. 2) analysed the social significance of variation in rhotic and non-rhotic pronunciations in New York City, and concluded (1) that rhotic variants were used more frequently by speakers with a higher socieconomic status, and (2) that speakers in general were more likely to use rhotic variants when paying greater attention to their speech.

In principle, the construct of the sociolinguistic variable can be applied at any level of linguistic analysis (the variants could be phonemes, morphemes, lexemes, syntactic constructions, etc.). However, sociolinguistic variation at the lexical level has been relatively under-studied, despite it arguably being that which "is most accessible and most salient for a non-specialist audience" (Durkin, 2012). This relative lack of attention is due to various methodological challenges which are more pertinent at the lexical level than at the level of phonology or morphology.

2.3.1 Challenges of analysing variation at the lexical level

Firstly, it is more straightforward to delineate a complete set of the phonological or morphological units that are available in a given language variety than it is to enumerate an entire lexicon. Durkin (2012, p.6) argues:

> "The lexicon is not a fixed entity, but is almost infinitely extendible, and the lexicon of each individual is likely to be unique. Thus, in any situation where we want to measure variation, we cannot make any easy assumptions about what the available pool of variants will be."

Secondly, even once assumptions have been made about the pool of available lexical items, it is by no means straightforward to decide which ones should be conceptualised as alternative variants of the same variable, since "a set of words that can sometimes realize the same basic concept are highly unlikely to be full synonyms" (Durkin, 2012, p.6). Two words may be semantically equivalent in some contexts but not in others, or may appear equivalent to a researcher lacking in familiarity with the language variety in question, but in fact have subtle differences.

In addition to the problem of appropriately delineating our variables, it is important that we have enough observations of them to be able to obtain good estimates of their relative frequencies across the different contexts we are interested in. Moreover, if we want to make generalisable claims about the factors governing variation at the lexical level (such as claims about usage rates of distinctively Scottish lexis, as opposed to claims pertaining specifically to the use of one particular lexical variable), we need to consider large numbers of lexical items in aggregate (Ruette et al., 2014).

For these reasons, defining lexical variables is an arduous task, and is inherently subjective, requiring a high degree of familiarity with the language varieties in question. Due to these obstacles, it may be tempting to limit the analysis to those variables which are most salient to the researcher, or else rely on existing catalogues of lexical variables which have been used in previous studies that did not necesarily deal with the same domains, communities, or time periods. This may introduce undesired biases and preclude the full picture of variation from being revealed.

In our work, we attempt to minimise such biases by identifying variables from the data. In Chapters 4 and 5, we identify terms that are distinctive to tweets from Scotland and then manually pair these with Standard English equivalents. Using Twitter data to identify distinctively-Scottish terms reveals spelling variants as well as novel lexical items not found in existing Scots dictionaries or word-lists from older studies. However, the manual process of pairing these with Standard English equivalents is still labour-intensive and dependent on intuition and experience with both language varieties. Many distinctively-Scottish terms do not have single-word Standard Enlish equivalents, some have equivalents which work only in certain contexts, and some have multiple equivalent forms in Standard-English, which could easily be missed. To address this problem, in Chapter 6 we introduce a data-driven method to identify pairs of terms, one of which is distinctive to a particular language variety, and the other of which is an equivalent term in a second language variety. Familiartiy with both varieties is still needed to accept or reject the resulting candidate pairs, but manually checking candidate variables is easier than thinking them up in the first place.

2.3.2 Alternation variables vs. frequency variables

So far I have focused on the traditional kind of variable used in quantitative sociolinguistics. Grieve (2015) refers to these as ALTERNATION VARIABLES and contrast them with FREQUENCY VARIABLES, which are commonly used in the field of corpus lin-

guistics. Whereas an alternation variable consists of multiple semantically-equivalent forms and is quantified in terms of their relative frequencies to one another, a frequency variable is concerned with only a single form, and is quantified in terms of its absolute frequency (usually normalised with respect to an appropriate measure of sample size).

Studies using frequency variables have tended to focus on the level of syntax, which shares the issues outlined above with the lexical level and has likewise received relatively little attention from the field of variationist sociolinguistics. Frequency variables circumvent the issue of precisely defining semantic equivalence between complex linguistic units, while large text corpora make it possible to quantitatively analyse large numbers of variables. Hence Grieve (2015) advocates the use of frequency variables in large corpus-based dialectometry studies, arguing that they enable the exploration of grammatical dialect variation "in a much more systematic and comprehensive way than had previously been possible".

Frequency variables have also been used in large scale corpus-based analyses of lexical variation across different geographical regions, demographic groups, and social contexts (Doyle, 2014; Bamman et al., 2014b; Jones, 2015; Pavalanathan and Eisenstein, 2015a). While these have revealed interesting patterns, great care is required when attempting to *explain* these patterns. It is important to consider that differences in word frequencies can reflect differences in *what* people are saying, and not only in *how* they are choosing to say it. While variation in the prevalance of certain topics or communicative functions across different social groups or contexts is a worthy object of study in its own right, we subscribe to Weiner and Labov's (1983, p.31) view that

> "the sharpest analytical conclusions on the conditioning factors that constrain linguistic change and variation can be made when form varies but meaning is constant, rather than when both are varying together."

A handful of large scale corpus-based lexical variation studies have used alternation variables (e.g. Gonçalves and Sánchez, 2014; Jørgensen et al., 2015; Huang et al., 2016), and have both replicated some known patterns and provided new insights. We have used alternations variables in our studies in Chapters 4 and 5, and in Chapter 6 we present a method to facilitate the curation of such variables.

Chapter 3

Data

The studies presented in this thesis use data collected from Twitter using its public APIs. These are relatively less restrictive and easier to use than those of other popular social media platforms, which has arguably led to an over-representation of Twitter in academic studies on social media. That being said, there have been few large scale studies about usage of minority language varieties on any social media platform. While ease of access was naturally a strong motivating factor in our choice of data source, we consider Twitter a particularly compelling domain in which to investigate the role of social factors in linguistic variation.

While all tweets are public[1], and thus have a potential audience which is boundless and inscrutable, Twitter does provide affordances to manipulate the *likely* composition of a tweet's audience, or to indicate the author's *intended* audience. Twitter users must learn to strategically employ these affordances in order to target the relevant audience for each tweet. Linguistic style and language choice are also resources that Twitter users can draw on in order to target particular sections of their potential audience. Indeed, Irvine (2002, pp.23-24) conceptualises linguistic style as "the ways speakers as agents in social (and sociolinguistc) space negotiate their positions and goals within a system of distinctions and possibilities". Since Twitter is itself a social space in which users are constantly negotiating their positions and goals, we consider it an auspicious stage on which we can observe how stylistic variation and code switching are put to work in this negotiation process.

Authenticity is highly valued on Twitter (Marwick and boyd, 2011), which may sometimes provide a license and impetus to emphasise aspects of one's personal style or identity through the use of distinctive linguistic variants. Creative and non-standard

[1] except those posted from protected accounts (see §3.1)

use of language is further encouraged by design features of the platform itself, such as the restrictive character limit for posts and the high-volume, real-time feeds which engender fast-paced interactions. On the other hand, the fact that Twitter provides unprecedented opportunities to interact with strangers from diverse backgrounds all around the world can create a competing impetus to use language that is inoffensive and accessible to as wide an audience as possible. As Pavalanathan and Eisenstein (2015a, p.188) note, social media platforms "play host to a diverse array of interactional situations, from high school gossip to political debate, from career networking to intense music fandom", such that "even though platforms such as Twitter are completely public, they capture language use in natural contexts with real social stakes."

The rest of this chapter is structured as follows: in §3.1 we provide a brief overview of Twitter's user interface, highlighting key features that enable users to influence a tweet's audience. In §3.2 we review ethical issues regarding the collection, analysis, and sharing of Twitter data. In §3.3 we review what little is known about the demographics of Twitter's user base, and discuss issues around assessing and addressing unintended biases in datasets sampled from Twitter. Finally, §3.4 is a quick reference guide to all of the datasets we have collected and used across our three studies.

3.1 What is Twitter and how do users manipulate their audience?

Twitter is a popular social network and microblogging platform, on which users post short[2] messages called TWEETS. While tweets can also include photos and video, this thesis is concerned only with the text. We will now provide an overview of some of the key features of the platform, focusing particularly on those which enable users to target their tweets towards particular sections of their potential audience. We note that Twitter regularly introduces new features and tweaks existing ones, so while we hope that this overview will be useful to other researchers, it is always important to verify what does and doesn't apply to the time periods relevant to one's own research.

A user's tweets appear in reverse chronological order on their own profile page, or USER TIMELINE. To FOLLOW a user is to subscribe to receive their tweets in your HOME TIMELINE: the personalised stream of tweets you see when you log into Twit-

[2]Tweets were originally limited to 140 characters; this character limit was extended to 280 in November 2017 (i.e. after the period covered by any of the datasets used in this thesis).

ter[3]. Following relationships can be asymmetrical on Twitter. The users who Follow you and receive your tweets on their Home Timelines are called your FOLLOWERS; the users who you Follow and whose tweets appear on your own Home Timeline are called your FRIENDS.

By default, tweets are public, so while only your Followers receive them in their Home Timeline streams, anyone (including people without a Twitter account) can view your User Timeline or find your tweet through Twitter's search facility. However, Twitter users do have the option to PROTECT their accounts, which makes their tweets viewable and searchable *only* by their Followers. Naturally, our studies are limited to the analysis of public tweets. In 2014, the year from which most of our data was drawn, the social media analytics platform Twopcharts[4] estimated that just 5.3 percent of all Twitter users were choosing to protect their accounts.

There are several ways in which users can curate additional timelines not limited to tweets by their Friends: they can create LISTS of users and see streams consisting of tweets by those specific users; they can perform a query using Twitter's search facility; or they can click on a HASHTAG (see Figure 3.1b). A Hashtag is any sequence of alphanumeric characters prepended with the 'hash' (a.k.a 'pound') symbol, e.g. *#PHDchat*, *#WritingWoes*, *#7thCupOfCoffee*. Twitter automatically makes hashtags into hyperlinks, connecting to streams of other tweets that contain the same hashtag. The original intended function of hashtags is to highlight relevant keywords or phrases, so that these can be used to categorise tweets and thereby help people to follow the conversation around topics that interest them. Including a hashtag in a tweet therefore has the potential to extend its audience beyond the sender's Follower group, increasing the probability that it will be seen by others with an interest in its subject matter.

The size of a tweet's audience can also be extended by users other than its author, through the RETWEET and QUOTE features. To Retweet is to forward a tweet to one's own Followers (and to one's own user Timeline), while the Quote facility enables users to append a comment of their own before Retweeting.

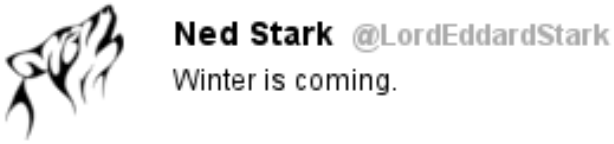
As well as directing their own tweets towards topic-based audiences through the use of hashtags, users can also direct tweets towards specific individuals through the use of MENTIONS. Twitter users are identified by unique usernames called HANDLES,

[3] Inititally this stream was updated in real time and presented in reverse chronological order, but in 2016 Twitter began to display TOP TWEETS, ranked algorithmly according to how likely the user is to care about them, nearer the top of the feed, above more recent but lower-ranked tweets. As of 2019 users can choose between a curated or strictly chronological Home Timeline.

[4] https://web.archive.org/web/20140625055055/http://twopcharts.com/Twitteractivitymonitor

Figure 3.1: Fictionalised example tweets illustrating various features that tweet authors can use to manipulate the likely composition of their audience (NB: while these affordances affect which streams a tweet is pushed to and thus who is more *likely* to see it, anybody potentially *could* view any of these tweets by visiting the author's profile or using Twitter's search facility)

(a) An original tweet without any HASHTAGS or MENTIONS. This tweet would be pushed to the Home Timeline streams of Ned Stark's Followers. Throughout this thesis we will refer to this kind of tweet as a BROADCAST.



(b) A tweet which includes a HASHTAG. This would be streamed to Olenna Tyrell's Followers and **additionally** to users following the *#Westeros* hashtag.

(c) A tweet which *begins* with a MENTION. Jon Snow would receive a special notification about this Mention, and the tweet would be pushed to the Home Timeline streams of **only** those users who follow both Ygritte and Jon Snow.

(d) A tweet which includes a MENTION but does *not* begin with it. Like a broadcast, this tweet would be pushed to the Home Timeline streams of all of Tyrion Lannister's Followers, but additionally Jamie would receive a notification about it.

which consist of alphanumeric strings prepended by '@', e.g. '@justinbieber', '@taylorswift13, '@TheEllenShow'. To Mention a user in a tweet is to include that user's handle in the tweet text. When a user is Mentioned they receive a special notification drawing their attention to the tweet.

If a tweet begins with a Mention, that is to say, if the very first token in the tweet is a user's Handle (as in Figure 3.1c), then that tweet is pushed to the Home Timeline streams of only those users who Follow both the sender and the mentioned user.[5] Since April 2014[6], beginning a tweet with a Mention has also had the effect of excluding it from the default view of the author's User Timeline, though anyone visiting the author's profile can still reveal such tweets by selecting the alternative 'Tweets and Replies' view[7]. Including a Mention later on in the tweet text (as in Figure 3.1d) does not restrict its audience in these ways, but does still generate a special notification for the Mentioned user.

To summarise, while their audience is never entirely within Twitter users' control, they can indicate who individual tweets are intended for, or manipulate who will be more likely to see them, through the use of Hashtags and Mentions. Other users can subsequently extend a tweet's audience by Retweeting or Quoting it, and optionally introducing Mentions or Hashtags of their own. In this thesis we focus on the use of Mentions and Hashtags in users' own original tweets, since we are interested in how their linguistic choices interact with the audiences they conceive as they are composing them.

[5]NB: this was the case throughout the time period covered by all of our datasets, and is still the case at the time of writing (see 'Note' here: https://web.archive.org/web/20200127211855/https://help.twitter.com/en/using-twitter/types-of-tweets), but Twitter has considered changing this and may do so in the future (see second footnote here: https://web.archive.org/web/20200405163231/https://blog.twitter.com/en_us/a/express-even-more-in-140-characters.html.

[6]More specifically, while the 'Tweets and Replies' tab was introduced to Twitter's web interface in April 2014 (see https://blog.twitter.com/official/en_us/a/2014/your-new-web-profile-is-here.html), on Twitter's iOS app tweets beginning with Mentions were still included in the main view of a User's Timeline until March 2017, with the change happening on the Android app some time in between (see https://www.socialmediatoday.com/social-networks/facebook-and-twitter-trying-out-new-looks-tweetscomments).

[7]Tweets with initial Mentions are sometimes refered to as 'Replies' because their original purpose was to address other users when replying to their tweets. It is important to note, however, that tweet-initial Mentions can also be used to *initiate* conversations (and also for other purposes; see §5.4.4), so not all tweets which begin with Mentions are actually Replies—though they are nevertheless restricted to the 'Tweets and Replies' tab. It is also worth noting that in March 2017 (which is after the period our datasets cover) Twitter introduced a mechanism for replying directly to a tweet without using Mentions; so since then it has not been the case that all Replies are tweets with initial Mentions, either.

3.2 Twitter data: ethical issues

Over the course of conducting the studies included in this thesis, the conversation in the wider research community around ethical issues with the use of Twitter data has evolved. While there is still not yet a well-established ethical framework for the use of Twitter data in sociological research, some guidelines for ethical decision-making have recently been proposed on the basis of exploratory surveys of Twitter users' expectations around the use of their tweets in academic research (Williams et al., 2017; Fiesler and Proferes, 2018).

Altough Twitter makes tweets freely and publicly available via its APIs, and users legally consent to this by agreeing to Twitter's terms of service, we cannot realistically assume all users to have read and understood these terms. Indeed, survey results indicate that many users are quite naïve about how their tweets may be collected and used by people beyond their followers. Thirty-four percent of the 564 UK Twitter users surveyed by Williams et al. (2017) reported being unaware that by accepting Twitter's terms of service they had provided consent for some of their information to be accessed by third parties, and almost two thirds of respondents to a similar survey by Fiesler and Proferes (2018) indicated that prior to the survey they had been unaware that researchers sometimes use tweets in academic research.

Reassuringly, 84% of those surveyed by Williams et al. were not at all or only slightly concerned about their tweets being used for research in university settings (vs. 16% who were quite or very concerned), and likewise 80% of Fiesler and Proferes's respondents indicated they were not uncomfortable with the idea of their tweets being used in academic research. Importantly however, both studies suggested that a majority of users would wish to be asked for their consent beforehand, particularly with regard to their tweets being *quoted* in academic research outputs. Both studies also indicate that a majority of users would be more comfortable with their tweets being quoted if their identity were to be anonymised.

When quoting tweets in the papers included in Chapters 4 and 6, the decision we took at the time was to reproduce only the text of the tweets, omitting the user handles and other metadata. However, this does not actually provide complete anonymity to the authors, since it is often possible to identify them by looking up the quoted text on Twitter's online search facility (importantly though, if the authors have since deleted the tweets in question or protected their accounts, then their identity will *not* be recoverable in this manner). In light of the recent survey findings and their accompanying

recommendations, if we were to do something similar in future we would instead seek prior consent from the users concerned to quote their tweets with attribution, or else use fictionalised examples, as we have done in Figure 3.1.

Apart from the issue of quoting tweets in publications, there are also important ethical considerations to be made around the collection and analysis of Twitter data. Fiesler and Proferes's respondents tended to feel more comfortable with the idea of their tweets being used in research if they were part of a large dataset being analysed computationally, as opposed to a small collection of tweets being read by humans. On the other hand, they tended to be less comfortable with the idea of their entire tweet histories being used than with individual tweets, and they also indicated discomfort with other profile or geolocation information being analysed along with their tweets—both of which are things we have done in our studies, albeit on a large-scale, aggregate level. While a majority of Fiesler and Proferes's respondents felt that researchers should not be able to use tweets in research without user permission, when dealing with large volumes of tweets, it is infeasible to obtain informed consent from all of the users concerned. Williams et al. (2017) condone the quantitative anaylsis of large-scale tweet datasets without seeking user consent, provided findings are presented only in aggregate form.

Another ethical challenge raised by Williams et al. (2017) is the potential for harm that could arise as a result of classifying content and users with sensitive labels without their knowledge. In chapter 4 we infer users' political stances with respect to Scottish independence. These could be construed as sensitive labels, though the classification is made entirely on the basis of hashtags associated with the pro- and anti- independence campaigns, so arguably in using these hashtags the users we are labelling have already chosen to publicly associate themselves with a stance on the issue. That being said, in applying a somewhat crude classification heuristic to a large dataset we are inevitably missing out on a great deal of context and nuance, which is likely to result in inaccurate labels for some users. While a small number of mislabeled datapoints may not substantively affect the outcomes of our quantitative analyses, misrepresenting their political stance could well cause harm to the users concerned. It would therefore be unethical to store or disseminate any inferred labels that we have associated with identifiable users, and indeed this is prohibted by Twitter's Developer Terms[8].

When it comes to archiving and sharing data collected from Twitter, there is a ten-

[8] https://developer.Twitter.com/en/developer-terms/more-on-restricted-use-cases.html – Accessed 2019-05-03.

sion between data protection and user privacy concerns on the one hand, and research transparency and reproducibility on the other. In its Developer Terms, Twitter advises that the best way to disseminate content obtained using its APIs to other parties is by sharing only the tweet or user IDs, such that the other party must then re-request the content itself using the APIs. This ensures that if a user decides to delete or 'protect' a tweet after the intitial dataset is collected, their decision will be respected when it is subsequently re-used. The downside of this is that if a significant proportion of tweets in the original dataset become unavailable, it may not be possible to verify results through reproduction studies.

3.3 Twitter data: demographics and sampling issues

The popularity of Twitter data in social scientific analyses is largely due to the fact that Twitter makes high volumes of data freely available to the public via APIs. This makes it possible to quickly and cheaply gather data from large numbers of individuals (orders of magnitude more than can be reached by traditional sociolinguistic surveys), and in some cases it is possible to obtain large numbers of datapoints per individual, covering long time periods and a variety of topics and interlocutors. However, a major disadvantage of Twitter data is that unlike with datasets obtained using more traditional methods, we usually have very little information about the individuals in our dataset, often lacking even coarse demographic categories such as gender, age, ethnicity, nationality, or socioeconomic class. Put succintly by Sloan (2017, p.2): "social scientific analysis is based on the investigation of group differences, but we can't easily identify the groups".

One thing we do have some ground truth about is the demographics of Twitter users as a whole. According to data collected by telephone surveys in 2015, adult Twitter users in the UK skew disproportionately young and male with respect to the the UK population at large, and are more likely to be in managerial and professional occupations (Sloan, 2017). Similarly, Twitter users in the USA skewed disprortionately male, young, affluent and educated in 2014 and 2015, but annual surveys show that US Twitter user demographics are changing over time: between 2012 and 2016, representation on Twitter increased significantly amongst older people, Whites and Hispanics, those in rural ares, and those with higher incomes (Duggan and Brenner, 2013; Duggan and Smith, 2013; Duggan et al., 2015; Duggan, 2015; Greenwood et al., 2016). The difference in the demographic distribution of Twitter users and that of the wider population

(plus the difficulty of estimating it, and the fact that it is ever changing) is an issue when Twitter data is used to make predictions about the offline world (e.g. election outcomes; Huberty 2015). However, as this thesis is specifically concerned with language use *on Twitter* as an object of study in its own right, it does not matter for us how representative Twitter is of the general population.

What does matter more is that we are interested in characterising linguistic choices for a particular *subpopulation* of Twitter's user base, and the lack of ground truth information available means it is often difficult to infer who belongs to a particular subpopulation. Moreover, it is difficult to determine whether certain sociodemographic groups may be systematically excluded by the indicators or inference methods we use to try to sample from that subpopulation. In sections §3.3.1 and §3.3.2, we discuss potential sources of bias in data sampled from Twitter, the steps we have taken to minimise them, and what implications they have for the interpretation of our findings.

3.3.1 Defining the sampling universe

A foundational challenge in sociolinguistic research is deciding how to define the SAMPLING UNIVERSE, that is, how to circumscribe the subsection of the population that you wish to study (Tagliamonte, 2006, chap. 2). With traditional sociolinguistic fieldwork methods such as surveys and interviews, it can be difficult to gain access to members of the target community. In a large-scale corpus study using data from Twitter, where most users' data is publicly accessible via APIs, the challenge does not lie in gaining access to data from people who belong to the target community, but in *identifying* who they are.

User profiles on Twitter are quite minimal: users are invited to provide a brief public summary about themselves, a location, a website URL, and their date of birth, all of which are optional and can be left blank. Unlike Facebook, Twitter has never required users to provide their real names, and users are free to choose creative display names and change them at will. Users can also upload a profile picture, but again this is optional and there is no obligation for the picture to be a true likeness of the user.

In our studies we're interested in regionally specific language varieties, so our target subpopulations can be defined on the basis of geographical location. Even though Twitter does provide facilities for users to disclose their location both on their profiles and in the metadata of individual tweets, only a minority of users opt to do so.

Geotags Twitter offers its users the facility to geotag their tweets, that is, to label them with a precise or approximate location. If a tweet is tagged with a precise location, the exact latitude and longitude is included in the tweet object returned by the APIs. Alternatively, users can label their tweets with a broader location label (such as the name of a city or neighbourhood) or, in select locations only, with the name of a specific landmark or business.

However, only a small proportion of tweets are geotagged: Sloan and Morgan (2015) reported only 3.1% of users had any geotagged tweets in the 1% Spritzer sample (see below) for April 2015. Moreover, while for some analyses it may be very useful to know precisely where a user was at the moment they posted a tweet, individual tweet-level labels can be misleading when what we really want to know is where the user lives. Ideally we would identify where a user spends most of their time by considering multiple geolocated tweets by the same user over a reasonably long time period, but in a 1% sample we may have very few geotagged tweets per user. An alternative heuristic employed by Pavalanathan and Eisenstein (2015a) is to infer a user's home region based on the geotags of tweets in which they are mentioned by other users. If a user is only ever mentioned in tweets from one particular locality, it is reasonable to assume they have strong ties to that area. However, this still requires a sufficiently large dataset such that there will be many users for whom the dataset contains mentions by several other users.

Location field Another manner in which Twitter users can disclose geographical information is through the optional location field in the user profile, though users can write whatever they wish in this field and there is no requirement that it be accurate, or even anything to do with location. While we may be able to find many users who do report locations that we can parse and reverse geocode, this would bias our sample in favour of people who choose to provide their location in a standard format, and may systematically exclude those with a greater tendency to use language in non-standard or locally-specific ways. Clearly, this would be a problem when non-standard and locally-specific language use is the object of study! Also, as with geotags this will be noisy data if we want to use it to identify where people are 'from', as not everyone fills it in sincerely, and even if it is sincere, we still don't know whether it's people's current location, where they grew up, or a location they identify with for some other reason.

Who discloses their location? Several studies have tried to profile the demographics of people who share their location via geotags or their user profile.

Pavalanathan and Eisenstein (2015b) linked self-reported first names with aggregate statistics from United States Censuses in order to induce approximate distributions over ages and genders for users who geotag their tweets to locations in the USA, versus users who report US city names in the location field on their profile. Compared with those who geotagged their tweets, a greater proportion of those who provided city names in their profiles were estimated to be middle aged or older, and the proportion who had male-associated names was slightly but significantly greater.

Wood-Doughty et al. (2017) used a variety of approaches to infer demographic attributes in a separate dataset of US tweets, and likewise found that male-tagged users were significantly more like to fill out their location fields, and significantly less likely to enable geotagging. They also used census data to predict users' ethnicity, and found that Asian- and Hispanic/Latino-tagged users were more likely than White and African American users to disclose their location, both via geotags, and via the location in their profile. However, they caution that this could be an artefact of the way they classify ethnicity: users with a greater propensity for self-disclosure may be easier to classify as Asian or Hispanic, which then presents a biased view of associations between behaviour and ethnicity.

Hashtags Rather than using explicit location information, another way to sample users with ties to a particular region is to query the APIs for tweets which contain hashtags relating to that region. More generally, sampling on the basis of hashtags is a common approach when the population of interest consists of tweets relating to (or users with some connection to) a particular topic or event (e.g. González-Bailón et al. 2011, Conover et al. 2013; Schrading et al. 2015, Sobhani et al. 2017).

Just as the subpopulation of users who disclose their location are not necesarily represenative of the wider Twitter population, frequent users of hashtags may not be either. Bruns and Stieglitz (2014) suggest that "accounts that use hashtags regularly may be 'Twitter experts', and different from other users in their behaviour and activity." This is particularly important to bear in mind given that one question we are specifically interested in is whether people modulate their linguistic behaviour when including hashtags in their tweets. The answer may be different for people who frequently use hashtags than for those who don't.

Another issue to keep in mind when sampling on the basis of hashtag use is that people use hashtags for a variety of reasons. Although hashtags were originally devised as a way to categorise tweets by keyword and extend their audience to potentially interested users outside the author's follower circle, the goal of bringing one's message or opinion to a broader audience and addressing people interested in a specific topic is only one of ten motivations for hashtagging identified in a multimethod study by Rauschnabel et al. (2019). While the audience-extending function was the most frequent motivation for hashtag use among 141 Twitter users they surveyed, other motivations include amusing other social media users, conforming with the posting style of peers, being seen to be engaged with trending topics, and even strenghthening bonds with an inner circle of friends, e.g. by using hashtags to refer to shared experiences that only they would understand. Hence the intention behind the use of a hashtag may not always be transparent, and it may not always be valid to assume that including a hashtag in a tweet indicates a genuine interest in or association with the topic it appears to index, or endorsement of the stance it appears to represent.

Summary and implications In summary, limiting our sampling universe to those who do share their location may bias it towards certain sectors of our target population, and systematically exclude others. Any other feature set we might conceive of to define our sampling universe would bring its own inherent biases. As noted at the beginning of this section, defining the sampling universe in such a way that it does not exclude any sectors of society who are of interest to the study, or include any who are not, is also a difficult issue for sociolinguistic studies which use traditional survey or interview-based methods. However, when using Twitter data we have the additional problem of it also being very difficult to characterise the demographic composition of the sample we ultimately obtain.

This predicament, which Gerlitz and Rieder (2013) describe as "the fundamental problem of how to go fishing in a pond that is big, opaque, and full of quickly evolving populations of fish", has different implications for different kinds of studies. We discuss the implications for our own studies in detail in the relevant commentary sections in Chapters 4-6, but generally speaking, the upshot is we need to be careful about making quantitative claims, and should not assume that our findings are generalisable beyond the samples we have.

3.3.2 Obtaining an unbiased sample

Supposing we did have a well-defined target population whose members could be reliably identified using tweet metadata, how could we retrieve an unbiased sample of this population?

Streaming APIs The Twitter Streaming APIs offer samples of the Twitter datastream via two public endpoints: the Sampling endpoint (aka the Spritzer) returns a small random sample (initially advertised as approximately 1%[9]) of all public statuses, while the Filtering endpoint returns public statuses that match one or more filter predicates.

Morstatter and colleagues have assessed the extent to which samples returned by the Streaming APIs are unbiased representations of the overall population of relevant tweets, by comparing data collected using these APIs with data collected over the same time period using the proprietary Firehose stream (i.e. the full stream of publicly available statuses). Their results indicate that the Sample endpoint is indeed unbiased, since characteristics of the data it yields closely resemble those of unbiased random samples drawn from the full Firehose stream (Morstatter et al., 2014).

On the other hand, their analyses did reveal some bias in the way data is provided by the Filtering endpoint. Specifically, data obtained using the Filtering endpoint is biased when the volume of tweets which match the query parameters exceeds 1% of the volume of all tweets. When this happens, the API returns an incomplete sample from the pool of tweets which match the query. Morstatter et al. (2013) used top hashtag lists and topics obtained via LDA to characterise the pool of all tweets in the Firehouse data that match a given set of parameters. They then compared these with the top hashtags and topics in data drawn from the Filtering endpoint over the same time period with the same parameters, and in unbiased random samples that they drew themselves from the Firehouse pool. They found that the topics and top hashtags diverged more from those of the Firehouse pool in the data from the Filtering endpoint than in their own samples. These results indicate bias in the sampling mechanism used by the Filtering endpoint, a conclusion which was corroborated in a similar investigation by Tromble et al. (2017).

Search API Tweets can also be retrieved using Twitter's Search API, which searches against a pool of so-called 'top' tweets published in the past 7 days. Twitter's docu-

[9] https://web.archive.org/web/20120708200616/https://dev.Twitter.com/docs/api/1/get/statuses/sample

mentation does not specify how 'top' tweets are chosen.

Analyses by González-Bailón et al. (2014) and Tromble et al. (2017) indicate that given the same query parameters, the Search API consistently yields a smaller volume of tweets than the Filtering endpoint of the Streaming API.

Tromble et al. (2017) used logistic regression models to gain insight into what features Twitter may use to sample data for the Streaming and Search APIs. As input they used a dataset of tweets from the Firehouse that matched certain parameters. The dependent variable was a binary variable indicating whether or not each tweet was returned by an API query using those same parameters. As independent variables they used a variety of features relating both to the tweet itself (e.g. whether or not it was a retweet, how many hashtags it contained, whether it contained embedded media) and to the user who posted it (e.g. the user's follower count, or the total number of tweets they had posted).

Tromble et al.'s results suggest that the sampling mechanism used by the Filtering endpoint of the Streaming API differs from that of the Search API. There were also some striking differences in effect directions and sizes between models built with different datasets collected a few months apart using the *same* API. This, they argue, points to the fundamental problem that "we simply do not know what we do not know" (Tromble et al., 2017, p.25). While careful comparisons can provide some insights into the extent to which particular attributes are systematically over- or under-represented in samples returned by different APIs, they can't be relied upon to inform future research because "Twitter's API algorithms can and do change on a regular basis" (Tromble et al., 2017, p.7).

Summary and recommendations Table 3.1 summarises the pros and cons of each of the APIs discussed above. The Sample endpoint of the Streaming API is the best option for collecting an unbiased, representative sample of all activity on Twitter. Since the 1% sample can be prohibitively small, the Filtering endpoint of the Streaming API may be more suitable for sampling a dataset with high coverage of tweets which meet specific criteria, particularly if these criteria are unlikely to ever match much more than 1% of the total volume of Twitter traffic (e.g. tweets from a reasonably small geographical area, or tweets which contain niche keywords or hashtags). However, the Filtering endpoint must be used with caution, as it is far from straightfoward to determine *when* and *how* the samples it provides are biased (Morstatter et al., 2014; Tromble et al., 2017). As for the Search API, since it is also subject to unknown biases

Streaming API: Sampling endpoint	Streaming API: Filtering endpoint	Search API
✓ Unbiased sample of all Twitter traffic	✓ Can yield larger samples vs post-filtering the Sample endpoint	✓ Requires less work and resources to use
✓ Repurposable: full real-time sample can be archived and post-filtered	× May be biased when >1% of Twitter traffic matches parameters	✓ Can retrieve historical tweets (up to 1 week)
× Small sample size (∼1%). Post-filtering on infrequent phenomena may yield prohibitively small samples	× Requires a substantial amount of additional work and resources to identify *when* it is biased	× Rate limits result in low data volumes
× Technical expertise and resources needed to maintain connection	× Technical expertise and resources needed to maintain connection	× Search results biased to favour 'top' tweets

Table 3.1: Pros (✓) and cons (×) of Twitter's APIs for sampling

but yields lower volumes of data than the Filtering endpoint of the Streaming API, we recommend it only for sampling Tweets which relate to unanticipated events that occured within the last 7 days, if post-filtering an archive of the Spritzer stream is not an option.

3.4 Overview of datasets collected

The analyses in this thesis are based on the following six datasets[10], all of which are derived from an archive of Twitter's 'Spritzer' stream (see §3.3.2). The relationships among these datasets are visualised in Figure 3.2.

[10]Various filtering/pre-processing steps were applied to each dataset before use; see Chapters 4 to 6 for full details.

Geotagged-UK (GU) Tweets from the 'Spritzer' sample which were posted between September 1st 2013 and September 30th 2014, and are geotagged to locations within the UK.

Geotagged-Scotland (GS) The subset of tweets in the GU dataset which are geotagged to locations within Scotland, specifically.

Indyref Tweets (IT) Tweets from the 'Spritzer' sample which were posted between September 1st 2013 and September 30th 2014, and contain hashtags relating to the 2014 Scottish Independence Referendum

Scottish Geotag Users' Autumn 2014 Timelines (SG-Users) Tweets which were posted in August, September, or October 2014 by users from the GS dataset. This dataset is not restricted to tweets which appear in the 'Spritzer' sample; instead it consists of complete User Timelines for the months concerned, retrieved using the `statuses/user_timeline` endpoint of Twitter's REST API in March 2017. The API returns up to 3200 of a user's most recent tweets, so the dataset is restricted to users who, as of March 2017, had not posted more than 3200 tweets since autumn 2014. For other users, their autumn 2014 tweets were no longer available at the time we collected this dataset.

Indyref Hashtag Users' Autumn 2014 Timelines (IH-Users) Tweets which were posted in August, September, or October 2014 by users from the IT dataset. Collected in the same way as the SG-Users dataset.

US Geotags (G-USA) Tweets from the 'Spritzer' sample which were posted between June 30th 2013 to July 1st 2016, and are geotagged to locations within the USA.

While Twitter's Terms of Service and the ethical considerations discussed in §3.2 preclude us from sharing these datasets in full, we have made each dataset available in the form of lists of user and tweet IDs, such that the tweets themselves can be re-collected for use in future studies via Twitter's APIs. The lists of IDs for each of these six datasets are publically available at https://doi.org/10.5281/zenodo.3517244.

In Chapter 4, the GU and GS datasets were used to identify distinctively Scottish words. We then analysed relationships between support for Scottish independence and usage rates of these distinctively Scottish words in the IH dataset.

In Chapter 5, we wanted to tease apart effects of topic and audience on the use of distinctively Scottish lexis. Because these analyses were concerned with intraspeaker variation, we needed more tweets per user. The authors of the tweets in the GS and IT datasets constitute two samples of users with a connection to Scotland: we know the users in GS to have been there, while the users in IT care enough about the Scottish independence referendum to have tweeted about it. So, we collected the SG-Users and IH-Users datasets, and applied mixed effects models to each of these.

In Chapter 6, we developed an automatic method to facilitate the identification of lexical variables in code-mixed text. We used the GU dataset to try to identify Scottish/BrEng variables, and the G-USA for AAVE/GenAm variables. We combined GU and G-USA to make a dataset for identifying GenAm/BrEng variables.

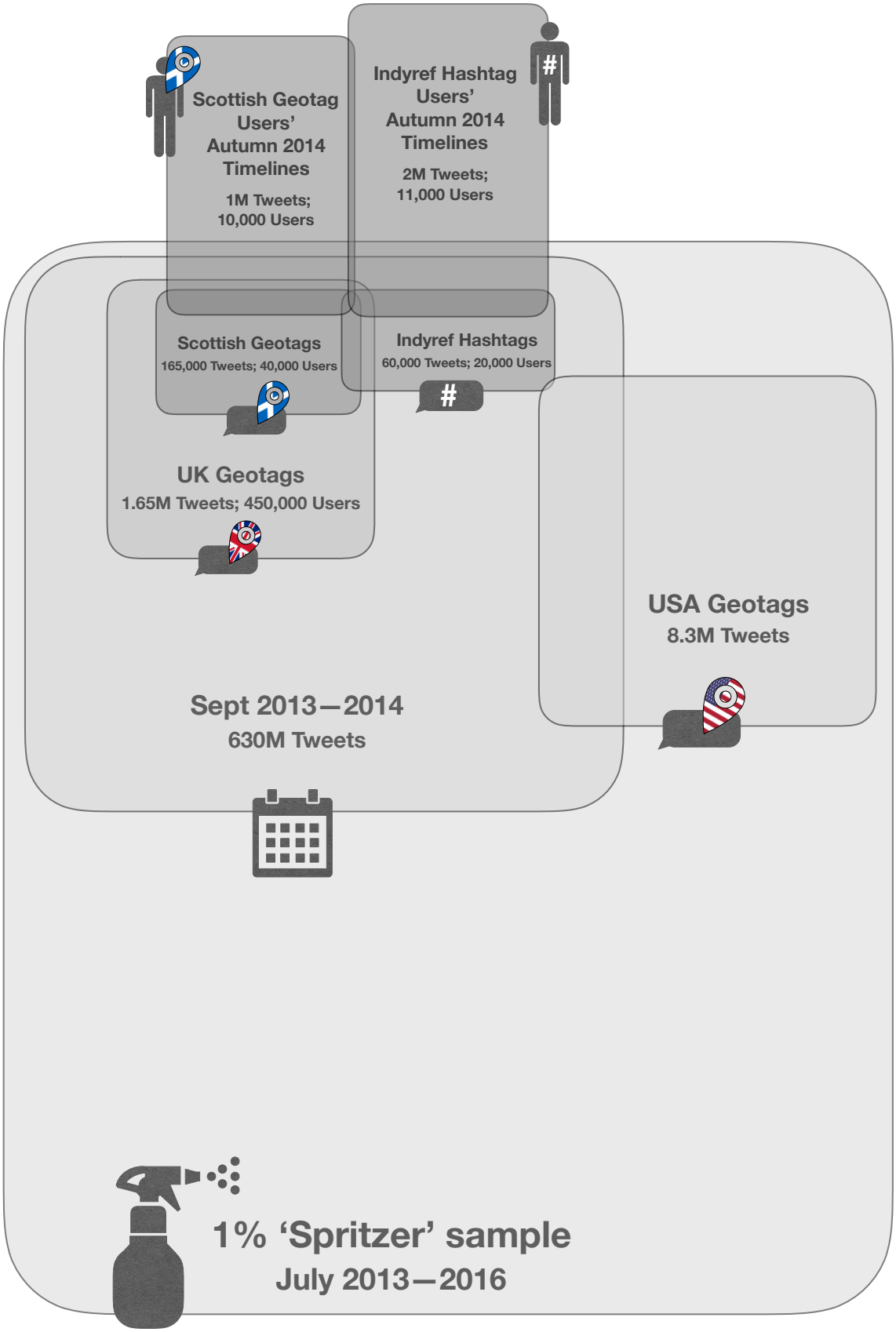


Figure 3.2: Venn diagram of datasets.

Chapter 4

Distinctively Scottish vocabulary and Scottish independence

4.1 Introduction

In this paper we investigate how support for Scottish independence correlates with distinctively Scottish vocabulary usage on Twitter. The extent to which people feel a sense of Scottish identity has been suggested to have been an important factor in voting decisions in the 2014 Scottish Independence Referendum. While not all supporters of Scottish independence identify as Scottish (and many who oppose it do), the 2013 Scottish Social Attitudes survey did indicate a clear correlation between people's sense of national identity and their voting intentions in the referendum (ScotCen, 2013). Moreover, a 2010 Scottish Government survey indiciated that the Scots language is an important part of Scottish identity (Scottish Government, 2010).

In Labov's (1978a) famous study of diphthong centralisation on the island of Martha's Vineyard, he found that inter-speaker variation in the use of this distinctive phonological feature was best explained by the degree to which speakers identified themselves with the island and the island way of life. A strong sense of Scottish identity may similarly be indexed through the use of distinctively Scottish vocabulary.

We hypothesised (1) that Twitter users who supported Scottish independence would be more likely to use distinctively Scottish lexis in their tweets, and (2) that they would increase this usage when specifically discussing the referendum. Using a large dataset of tweets covering a one-year period around the 2014 referendum, we identified terms which are statistically associated with tweets composed in Scotland, as opposed to the rest of the UK. We paired these with Standard English equivalents, and compared usage

rates of the Scottish variants across users who had used pro- or anti- independence hashtags.

This is a somewhat sensitive area, as many people feel agrieved that the Scots language is often construed as being 'just for nationalists', and argue that politicisation of its use is counterproductive to efforts to promote and destigmatise it (e.g. Uri 2018). Our quantitative analysis shows that distinctively Scottish lexical variants are used on Twitter both by pro- and anti-independence users. While those users we infer to be supporters of independence do indeed use them at higher rates, our analysis suggests that they are no more likely to use distinctively Scottish lexical variants when tweeting about the independence referendum than in their general Twitter activity.

4.2 Author contributions

The paper is co-authored by me, Debnil Sur, Luke Shrimpton, Iain Murray, and Sharon Goldwater. Luke Shrimpton extracted the initial dataset from the Spritzer archive, and provided guidance with managemdent and processing of the data. Debnil Sur helped to formulate the research questions, process the data, and conduct preliminary analyses. As the leading author, I co-supervised Debnil Sur in the initital stages of the project, produced the final datasets, conducted the final analyses, and wrote the paper. Iain Murray advised on experimental design and helped to revise the final manuscript. Sharon Goldwater supervised the project, offered suggestions, and helped to revise the final manuscript.

4.3 The paper

The paper was accepted for publication at the 2017 EACL conference in Valencia, where it was featured as an oral presentation. The publication reference is as follows:

> Shoemark, P., Sur, D., Shrimpton, L., Murray, I., & Goldwater, S. (2017, April). Aye or naw, whit dae ye hink? Scottish independence and linguistic identity on social media. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers* (pp. 1239-1248).

Aye or naw, whit dae ye hink?
Scottish independence and linguistic identity on social media

Philippa Shoemark[*]
p.j.shoemark@ed.ac.uk

Debnil Sur[†]
debnil@stanford.edu

Luke Shrimpton[*]
luke.shrimpton@ed.ac.uk

Iain Murray[*]
i.murray@ed.ac.uk

Sharon Goldwater[*]
sgwater@inf.ed.ac.uk

[*]School of Informatics
University of Edinburgh

[†]Department of Computer Science
Stanford University

Abstract

Political surveys have indicated a relationship between a sense of Scottish identity and voting decisions in the 2014 Scottish Independence Referendum. Identity is often reflected in language use, suggesting the intuitive hypothesis that individuals who support Scottish independence are more likely to use distinctively Scottish words than those who oppose it. In the first large-scale study of sociolinguistic variation on social media in the UK, we identify distinctively Scottish terms in a data-driven way, and find that these terms are indeed used at a higher rate by users of pro-independence hashtags than by users of anti-independence hashtags. However, we also find that in general people are *less* likely to use distinctively Scottish words in tweets with referendum-related hashtags than in their general Twitter activity. We attribute this difference to style-shifting relative to audience, aligning with previous work showing that Twitter users tend to use fewer local variants when addressing a broader audience.

1 Introduction

A central idea from sociolinguistics is that people's social identity is reflected in their use of language, and that people modulate their use of language in order to present particular identities in different situations. The recent availability of social media data has raised interest in confirming and extending these results using large scale datasets. For example, Twitter data has been used to examine patterns of regional variation in general US English (Doyle, 2014; Huang et al., 2015), African American English (Jones, 2015), and global Spanish (Gonçalves and Sánchez, 2014), and to study variation associated with factors such as race/ethnicity (Jones, 2015; Blodgett et al., 2016; Jørgensen et al., 2015) and gender (Bamman et al., 2014). These studies have shown that tweets mirror spoken language in many ways, such as displaying dialect variation not only in the use of distinct lexical items, but also in the use of non-standard spellings to indicate non-standard pronunciation—in fact, these spellings even reflect the phonological processes found in spoken language (Eisenstein, 2015). There is also evidence that, as in spoken language, individuals may shift their style of language in response to the audience. In particular, studies have found that when the expected audience of a tweet is larger, Americans use fewer non-standard and local words (Pavalanathan and Eisenstein, 2015) and Dutch bilingual speakers of a minority language are more likely to use Dutch rather than their other language (Nguyen et al., 2015). A small-scale case study of a single Scottish Twitter user also provides preliminary evidence that users may modulate their production of regional variants according to the topic of the tweet (Tatman, 2015).

Here we present the first large-scale sociolinguistic study of British tweets, and the first to examine the relationship between sociolinguistic variation and political views using social media data. We use a large corpus of tweets to examine the relationship between users' linguistic choices and their views about the 2014 Scottish independence referendum. The referendum (on whether Scotland should leave the UK) generated considerable political discussion and an unprecedented turnout of 84.6% of the

electorate, with the 'No' (anti-independence) side taking 55.3% of the vote. The 2013 Scottish Social Attitudes Survey (ScotCen, 2013) showed a clear correlation between national identity and voting intentions (53% of those who identified as 'Scottish not British' said they intended to vote 'Yes' to independence, vs. just 5% of those who identified as 'British not Scottish'), and there was much discussion in the popular press about the relationship between a sense of Scottish identity and support for Scottish sovereignty.

Although this recent discussion was not centered on language, there is a long history of scholarly discourse connecting the use of the Scots language[1] and sociolinguistic and political identity (Grant, 1931; Mcafee, 1985; Corbett et al., 2003). If this connection still holds today, then we might expect to find that those on the 'Yes' side of the debate use more identifiably Scottish language than those on the 'No' side. We might also expect to find some modulation of Scottish language use depending on whether users are discussing the referendum or not.

To examine these questions, we used a data-driven approach to identify linguistic terms that are used more in Scotland than in the rest of the UK. The identified terms include uniquely Scots words that are attested in Scots literature dating back to the 1600s and earlier, contemporary regional colloquialisms, spelling variants of Standard English words which reflect Scottish pronunciations, and acronyms used as shorthand for distinctive Scottish phrases. From these, we selected variables for which users can produce either a Standard English or Scottish variant (e.g., DO vs. DAE). We then classified users as pro- or anti-independence based on the referendum-related hashtags they used and asked whether these two groups use Scottish variants at different rates. We found that the pro-independence group did use Scottish variants significantly more than the anti-independence group, although the overall rate of Scottish variants is very low amongst all users.

Next, we compared the use of Scottish variants in tweets containing referendum-related hashtags to their use in other tweets. If users are aiming to project their Scottish identity as part of political discourse, then we might expect greater use of Scottish variants in referendum tweets than in non-referendum tweets. However, previous studies have suggested that non-standard and local variants are used *less* frequently in tweets containing hashtags, which typically have a larger audience than other tweets (Pavalanathan and Eisenstein, 2015). This effect would predict the opposite result—a lower use of Scottish variants in tweets with referendum hashtags—and indeed this is the result we found. So it appears that although pro-independence users do make greater use of Scottish variants overall, they do not increase their Scottish usage when engaging in broad-audience political discourse.

To summarize, the contributions of our paper are: (1) The first large-scale study of dialect variation on twitter in the UK. We show that in addition to using Scots in speech and some literary genres such as poetry, people are using Scots in informal public writing. The data-driven approach enables us to identify Scotland-specific lexical items without relying on pre-conceived notions of which variables to look for (cf. Tatman, 2015), and reveals that in addition to using attested Scots vocabulary, Twitter users appear to be creatively adapting to the medium with their use of acronyms for distinctly Scottish turns of phrase. (2) The first study connecting sociolinguistic variables to political stance using social media data, showing that pro-independence users have a higher rate of Scottish usage. (3) Further evidence of Pavalanathan and Eisenstein's (2015) claim that Twitter users modulate their language according to the audience, with local variants being less likely in tweets directed to larger audiences.

2 Context

'Scots' refers to the group of dialects historically spoken in the Lowlands of Scotland. While Scots has Anglo-Scandinavian origins in common with English, by the 16th century its pronunciation, vocabulary, and literary norms had considerably diverged from those of English, and Scots had become established as the prestige language in Scotland (Kay, 1988).[2] However, following the Union of Crowns in 1603, when King James VI of Scotland acceded to the thrones of England and Ireland,

[1] Historically, Scots has been considered a different language than English (see §2), though with many cognates and overlapping vocabulary. Most native Scottish people today speak some variety of Scottish English, which retains a few uniquely Scots words but is mainly distinguished from other varieties of English by its pronunciation.

[2] Previously, Gaelic had been the dominant spoken and literary language in Scotland. Note that while in medieval times non-Gaelic speakers referred to the Gaels as 'Scots', what we now refer to as 'Scots' is the Anglo-Scandinavian language which spread at the expense of Scottish Gaelic (a Celtic language) in the 15th & 16th centuries.

he and his court began to adopt English norms in their writing. After the Union of Parliaments in 1707, English firmly replaced Scots as the language of serious or elevated discourse in Scotland (Grant, 1931). While some people still use distinctive elements of Scots in their speech, until recently the average Scottish person's exposure to written Scots would have been largely confined to a select few literary domains such as poetry and comic narrative (Corbett et al., 2003). However, social media has given rise to a new genre of casual, communicative writing that is potentially visible to large and diverse audiences, providing both a platform and an impetus to express one's identity through the use of written language. Below, we provide three example tweets (each from a different user) which contain orthographic representations of Scots vocabulary and/or Scottish English pronunciation. Standard English variants of Scottish terms are provided in italics.

(1) No matter how shite *[shit]* a day you've had just remember there's always good biscuits in yer *[your]* grannies hoose *[house]*

(2) "Absolute carnage" at polling station earlier. Bairns *[kids]* playing, polite grannies, Yessers and Nos blethering *[blathering]* to each other. #VoteYesScotland

(3) #fuckoffscotland hud on we will fuck off but afore we dae eh challenge ye tae a square go ya queen loving DIDDY doughnut Sasijs YUP-TAE
#fuckoffscotland hold on we will fuck off but before we do I challenge you to a fair fight you queen loving fools. What are you doing!?

3 Data

Our data was drawn from the Sample endpoint of Twitter's Streaming API (a.k.a. the 'Spritzer'), which provides a random 1% sample of all public tweets in near real-time. We started with all tweets streamed from the Spritzer between 1st September 2013 and 30th September 2014. These dates cover a year of activity leading up to the referendum, as well as the day the vote took place (18 September 2014), and immediate reactions. We used a language classifier (Lui and Baldwin, 2012) to filter out non-English tweets, yielding an initial dataset of 629,431,509 tweets.[3] Because we are interested in the linguistic choices that individuals make in various contexts, we took steps to remove tweets which were not originally authored by the individual who posted them. Retweets (tweets which are verbatim copies of other tweets) were identified by a case-insensitive search for the token 'RT', and discarded. Quote tweets (tweets which contain verbatim copies of other tweets, but are augmented with original comments) were dealt with by discarding any text between double quotation marks, but retaining the remainder of the tweet.

From this initial dataset we extracted three overlapping subsets:

The Geotagged-UK (GU) dataset contains all tweets geotagged to a location in the United Kingdom (1,654,204 tweets by 446,923 distinct users).

The Geotagged-Scotland (GS) dataset contains all tweets geotagged to a location in Scotland (166,992 tweets by 40,861 distinct users).

The Indyref Tweets (IT) dataset consists of tweets containing hashtags relating to the 2014 Scottish Independence Referendum.

To construct the IT dataset, we first created a list of relevant hashtags, starting with the following five seed hashtags: *#IndyRef, #VoteYes, #VoteNo, #YesScotland, #BetterTogether*.[4] For each of these five seeds, we extracted from our initial filtered dataset a list of *all* tweets by any user who used the seed hashtag. We identified the 100 most frequent hashtags in each of these five lists of tweets, and manually discarded all hashtags which were unrelated to the referendum, as well as those which were highly ambiguous (e.g., *#Indy*, which sometimes refers to the referendum, but also commonly refers to a genre of music). The resulting list of referendum-related hashtags is given in Table 1.

Next, we extracted all tweets from our initial dataset which contain at least one of the hashtags on this list, yielding 77,708 tweets by 26,019 distinct users. We then applied a heuristic to filter out tweets produced by bots and spammers: for

[3] One might be concerned that an automatic language filter could remove some of the heavily Scottish tweets. However, even tweets such as example (3) in §2 are assigned a very high probability of being English by the filter. Perhaps other tweets with many Scottish terms were filtered out, in which case we will underestimate the probability that users choose Scottish variants. However this issue should not cause us to find differences in use between different groups where there are none.

[4] 'Yes Scotland' and 'Better Together' are the names of the principal organisations representing the Yes and No vote campaigns, respectively.

each user in the IT dataset for whom we had at least 5 tweets in the initial dataset, we computed the proportion of their tweets that contain URLs, and discarded users for whom this proportion was in the 90th percentile. This step filtered out 11,443 tweets by 1389 users.

Note that seven of the hashtags in Table 1 (*#voteyes, #bettertogether, #nothanks, #voteno, #yes2014, #letsstaytogether,* and *#yesvote*) are occasionally used in contexts unrelated to the Scottish Independence Referendum (e.g. *#bettertogether* can also refer to interpersonal relationships). However, they are distinctive enough that if a user has also used hashtags which are unambiguously related to the referendum, then it seems reasonable to assume that their usage of these potentially-ambiguous hashtags relates to the referendum too. Therefore, in order for a tweet containing one of these seven hashtags to be retained in the Indyref dataset, we required that its author had also used at least one other hashtag from Table 1. This criterion filtered out a further 6601 tweets by 6041 distinct users, such that the final IT dataset contains 59,664 tweets by 18,589 distinct users.

4 Identifying distinctively Scottish vocabulary on Twitter

We wish to identify terms that are more likely to be used by Twitter users in Scotland than in the rest of the UK. We follow the method of Pavalanathan and Eisenstein (2015), who used the Sparse Additive Generative Model of Text (SAGE) framework (Eisenstein et al., 2011) to identify tweet terms associated with metropolitan areas in the United States. SAGE models deviations in the log-frequencies of terms in a corpus of interest (here, the GS dataset) with respect to their log-frequencies in some "background" corpus (here, the GU dataset). The estimated deviations are regularized to avoid overstating the importance of deviations in the frequencies of rare words. Here, we use a publicly available implementation of SAGE[5] to obtain log-frequency deviation estimates for all terms which occur at least fifty times in the GU dataset, excluding hashtags, mentions, URLs, and stopwords. The terms with the highest estimates are those which are most distinctive to tweets geo-located in Scotland.

[5]https://github.com/jacobeisenstein/jos-gender-2014/

4.1 Scotland-specific terms

Unsurprisingly, many of the Scotland-specific terms are proper nouns which are topically associated with Scotland, such as Scottish placenames, political figures, and sports personalities. There are also several common nouns (e.g. 'devolution', 'bagpipes') and verbs (e.g. 'canvass', 'invade') which are strongly associated with the political or cultural climate in Scotland. These terms occur with greater relative frequency in the GS dataset simply because their referents are discussed with greater relative frequency; not because they are distinct from the terms that people in the rest of the UK use to index those referents. However, there are also many terms with high log-frequency deviations that *are* linguistically distinctive. To isolate such terms, we began with the 400 terms with the highest estimated deviations, and then manually filtered this list, discarding Standard English words, proper nouns, numerals, and non-standard terms which had clear topical associations (e.g. 'devo': an abbreviation for 'devolution'; 'hh': an acronym for 'Hail Hail', a football chant used by supporters of Celtic F.C.). The remaining 113 distinctively Scottish terms are listed in Table 2.

Almost three fourths of these terms are attested in the Scottish National Dictionary (SND) (Grant and Murison, 1931) or its online supplement (Scottish Language Dictionaries, 2004), which catalogue words that are distinctive to Scots (i.e. those which are not used, or are used differently, in Standard English), covering the period from the 1700s up to the present day. Many are also attested in the Dictionary of the Older Scottish Tongue (Aitken et al., 1990), which catalogues the entire vocabulary of Scots from the 1100s to the late 1600s. Of the attested Scots words, some are unique to Scots, e.g. BAIRNS ('sons/daughters'), GREETIN ('weeping'); some are cognates with English words that have fallen out of common usage, e.g. CRABBIT ('crabbed'; 'ill-tempered'), FEART ('feared'; 'frightened/timid'); some are cognates with English words but have a wider range of senses, e.g. HUNNERS is cognate with'hundreds', but used more generally to mean 'lots' as in "love you hunners", "there was hunners to do"; and many differ only in form from their English cognates, e.g. AFF ('off') and BAW ('ball').

Of the 29 terms that are not attested in SND, 9 are spelling variants or derived forms of attested

Neutral hashtags: *#IndyRef* (46,491) *#ScotlandDecides* (2552) *#BBCIndyref* (1591) *#ScotDecides* (934) *#BigBigDebate* (676) *#ScottishIndependence* (583) *#IndyPlan* (296) *#ScottishReferendum* (239) *#IndyReasons* (180) *#IndependentScotland* (26)
Yes hashtags: *#VoteYes* (8463) *#YesScotland* (1453) *#YesBecause* (1312) *#The45* (908) *#YouYesYet* (827) *#YesScot* (670) *#ActiveYes* (508) *#HopeOverFear* (325) *#Yes2014* (321) *#VoteYesScotland* (256) *#GoForItScotland* (153) *#The45Plus* (138) *#YesFlash* (114) *#GenYes* (92) *#YesVote* (76) *#1Year2Yes* (56) *#VoteAye* (53) *#FreeScotland* (52) *#SaorAlba* (45) *#YesGenerations* (39) *#RIPBetterTogether* (36) *#NHSForYes* (24) *#AnotherScotlandIsPossible* (23) *#EndLondonRule* (13)
No hashtags: *#BetterTogether* (2342) *#NoThanks* (1103) *#VoteNo* (867) *#LabourNo* (333) *#LetsStayTogether* (145) *#VoteNo2014* (92) *#UKOK* (86) *#VoteNoScotland* (45) *#JustSayNaw* (43) *#VoteNaw* (42) *#NoScotland* (34) *#DayOfUnity* (30) *#MaintainTheUnion* (9)

Table 1: Hashtags related to the Scottish Independence Referendum and their frequencies in the IT dataset

Scots words, e.g. CANA, CANNY, and CANI are alternative spellings of the attested CANNAE, and WANTY is a contracted form of 'want to', analogous to the attested GONNAE and GONY. A further 5 are orthographic representations of distinctively Scottish pronunciations, e.g. ANO ('I know'), HING ('thing'); and 2 are acronyms for distinctively Scottish turns of phrase: GTF ('Get Tae Fuck') and MWI ('Mad Wae It'). The final 13 could be described as contemporary Scottish slang, and include abbreviations: BEVY ('beverage')[6], DEFOS ('definitely'); drug-related lexis: WHITEY, ECCIES; profanities: BOABY, FANNYS; and everyday affective and descriptive words: DYNO ('amazing'), ROASTER ('idiot').

4.2 Lexical variables

Our goal is to measure the rate at which people index their Scottishness (either consciously or subconsciously) through the use of distinctively Scottish words, and to find out whether this rate varies across different groups of users (Yes hashtag users vs. No hashtag users), or across different contexts (tweets which contain referendum-related hashtags vs. tweets that don't).

Were we to directly compare the frequencies of our Scottish terms across different sets of tweets, it would be difficult to untangle differences in the rate at which users are indexing the *referents* of those terms from differences in the rate at which they are indexing their Scottishness. For example, if people use the term MASEL ('myself') with a lower frequency in one context than in another, this could be because they are modulating their use of distinctively Scottish terms in response to the context, but it could also be because they are modulating the rate at which they talk about themselves. To avoid this confound, we instead compare the *conditional* probabilities with which Scottish terms are used, given that their referents are being indexed at all.

We therefore consider only those Scottish terms for which we can identify semantically equivalent Standard English variants. We require that each variant of a given variable indexes the same set of senses and can occur in the same set of contexts, so for example we do not include YOUS as a variant of **YOU**, since while Scottish YI and Standard English YOU can index both the singular and plural second person pronouns, YOUS is only used for the plural. We also did not include variants of **YES** and **NO** since their use could be influenced by campaign slogans (e.g., the hashtags *#VoteAye* and *#JustSayNaw*). Our variables are listed in Table 3.

5 Study 1: Scotland-specific vocabulary usage on either side of the debate

Do tweeters who use Yes hashtags use Scottish variants at a higher rate than tweeters who use No hashtags, either when using these hashtags, or in general?

5.1 Method

We assign users in the IT dataset to two groups, ***Yes*** and ***No***, based on the quantity $\frac{n_{u,yes}}{n_{u,yes}+n_{u,no}}$, where $n_{u,yes}$ is the number of tweets in which user u has used at least one of the Yes hashtags and none of the No hashtags in Table 1; and $n_{u,no}$ is the number of tweets in which u has used at least one No hashtag and none of the Yes hashtags. The ***Yes*** group consists of all users for whom this quantity is greater than or equal to 0.75, while the ***No*** group consists of all users for whom it is less than or equal to 0.25. Users for whom the value lies between 0.25 and 0.75 (as well as those for whom our dataset does not contain any tweets with Yes or No hashtags), are not assigned to either group. The ***Yes*** group

[6]While'bevy' is also used colloquially for 'beverage' in other parts of the UK, in Scotland it is more frequent and can additionally be used as a mass noun ("I had so much bevy I couldn't even carry it"), and as a verb ("I'd bevy with him every weekend").

Acronyms:	GTF MWI
Closed Class Words:	ABOOT AE AFF ATS DAE FAE HAE MASEL MASELF OAN OOR OOT TAE WAE WAN WI WIS YERSEL YI YIN YOUS
Contractions	CANNAE CANNI CANY CANA DEH DINI DINNY DIDNY DOESNY GONNAE GONY ISNY WANTY YER YIR
Discourse Markers:	ACH ANAW ANO AWRIGHT AWRITE AWRYT AYE EH NAE NAW OOFT YASS YASSS YASSSS YASSSSS YIP
Open Class Words:	AULD AWFY BAIRNS BAW BAWS BELTER BELTERS BEVY BOABY BOKE BRAW BURD BURDS CRABBIT DAFTY DAIN DEFOS DOON DUGS DYNO ECCIES FANNYS FEART FITBA FUD GAD GAWN GEES GID GRANDA GREETIN HAME HAW HING HINK HOOSE HOWLIN HUNNERS JIST LADDIE LASSIE LASSIES MANKY MAW MAWS MORRA MONGO PISH PISHED PISHING RAGIN ROASTER SARE SHITE SHITEY STEAMIN SUHIN WEANS WHITEY

Table 2: Scotland-specific vocabulary. Standard English equivalents of many words are shown in Table 3.

contains 4,513 users, while the ***No*** group contains 1,356 users, which is consistent with the general perception at the time that the Yes campaign was much more vocal than the No campaign. To test our hypothesis that the probability of choosing Scottish variants is, on average, greater for users in the ***Yes*** group than for users in the ***No*** group, we estimate the difference between the two groups in the average probability of choosing Scottish variants, and conduct a permutation test to approximate the distribution of this difference under the null hypothesis. We first test whether the ***Yes*** group are more likely than the ***No*** group to use Scottish variants in tweets which contain hashtags that indicate a stance on the referendum. Subsequently, we test whether the ***Yes*** group are more likely than the ***No*** group to use Scottish variants in general across all of their tweets.

5.1.1 Test statistic

Let U_g be the set of all users in group $g \in \{yes, no\}$ who have used at least one of the variables in Table 3. For a given user $u \in U_g$, let V be the set of all variables that u has used in at least one tweet. We estimate the probability of user u choosing a Scottish variant of variable $v \in V$ as $\hat{p}_{u,v} = \frac{n_{u,vscot}}{n_{u,v}}$, where $n_{u,vscot}$ is the token count of Scottish variants of v in user u's tweets, and $n_{u,v}$ is the token count of all variants of v in user u's tweets. Averaging across variables, we obtain $\hat{p}_u = \frac{1}{V}\sum_{v \in V} \hat{p}_{u,v}$. We then average across users to obtain the group mean, $\hat{p}_g = \frac{1}{U}\sum_{u \in U_g} \hat{p}_u$. Our test statistic is the difference between the two group means, $d = \hat{p}_{yes} - \hat{p}_{no}$.

5.1.2 Permutation test

We randomly shuffle users between the two groups (maintaining each group's original number of users), and re-compute the value of d using these permuted groups. We repeat this procedure 100,000 times in order to approximate the distribution of differences in group means that would be observable were the difference independent of the assignment of users to groups. The proportion of permuted differences which are greater than or equal to the observed difference between the original group means provides an approximate p-value.

	Tweets w/ Yes or No hashtags		All tweets	
Group	***Yes***	***No***	***Yes***	***No***
# Users	3776	1121	4352	1322
# Tweets	10,436	2411	173,171	80,736

Table 4: Number of users and tweets included per group in the two analyses in Study 1

5.2 Results

For a tweet to be included in the analysis, it must contain at least one of the variables in Table 3. Hence not all users contribute data to the test statistic, as some have not used any of the variables in their tweets. The number of tweets and users included in each analysis are shown in Table 4.

The results for the first analysis are shown in the left column of Table 5. The difference between the two groups in their average probability of choosing Scottish variants in tweets that contain polarised referendum hashtags is statistically significant ($p < 0.002$). Results for the second analysis are shown in the right column of Table 5. Once again, the difference between the two groups is statistically significant ($p < 0.001$).

5.3 Discussion

The results show that the ***Yes*** group do use Scottish variants at a significantly higher rate than the ***No*** group, both when using Yes or No hashtags, and in general. The stronger significance level for the 'All tweets' dataset is partly due to its larger size (see Table 4), which enables better estimates of the

Variable	Scottish variants (freq. per million words)	Standard English variants (freq. per million words)
ABOUT	ABOOT (50)	ABOUT (2562)
ALRIGHT	AWRIGHT (10), AWRITE (17), AWRYT (17)	ALRIGHT (77), ALL RIGHT (4)
BALL	BAW (11)	BALL (116)
BALLS	BAWS (17)	BALLS (47)
BIRD	BURD (35)	BIRD (78)
BIRDS	BURDS (31)	BIRDS (44)
DEFINITELY	DEFOS (27)	DEFINITIELY (217)
DIDNT	DIDNY (26)	DIDNT (563), DID NOT (31)
DO	DAE (61)	DO (2712)
DOESNT	DOESNY (18)	DOESNT (433), DOES NOT (33)
DOGS	DUGS (11)	DOGS (69)
DOING	DAIN (17)	DOING (590)
DONT	DEH (12), DINI (12), DINNY (62)	DONT (2880), DO NOT (92)
DOWN	DOON (49)	DOWN (786)
FOOTBALL	FITBA (13)	FOOTBALL (289)
FROM	FAE (77)	FROM (2485)
GIVES	GEES (14)	GIMME (5), GIVE ME (108), GIVE US (21), GIVES (75)
GOING	GAWN (15)	GOING (1884)
GOOD	GID (82)	GOOD (2602)
GRANDAD	GRANDA (7)	GRANDAD (19), GRANDFATHER (5), GRANDPA (9)
HAVE	HAE (9)	HAVE (4549)
HOME	HAME (22)	HOME (832)
HOUSE	HOOSE (20)	HOUSE (463)
I KNOW	ANO (42)	I KNOW (556)
ISNT	ISNY (16)	ISNT (342), IS NOT (151)
JUST	JIST (7)	JUST (5550)
MYSELF	MASEL (14), MASELF (15)	MYSELF (553)
OF	AE (75)	OF (9186)
OFF	AFF (82)	OFF (1567)
OLD	AULD (28)	OLD (526)
ON	OAN (38)	ON (7782)
ONE	WAN (33), YIN (28)	ONE(2537)
OUR	OOR (14)	OUR (790)
OUT	OOT (181)	OUT (3053)
PISSED	PISHED (19)	PISSED (66)
PISSING	PISHING (12)	PISSING (32)
SHIT	SHITE (428)	SHIT (764)
SHITTY	SHITEY (25)	SHITTY (52)
SOMETHING	SUHIN (17)	SOMETHING (614)
SORE	SARE (13)	SORE (140)
THATS	ATS (9)	THATS (1405)
THING	HING (11)	THING (749)
THINK	HINK (34)	THINK (1939)
TO	TAE (186)	TO (19996), TOO (1629)
TOMORROW	MORRA (27)	TOMORROW (1183)
WANT TO	WANTY (52)	WANNA (284), WANT TO (940)
WAS	WIS (33)	WAS (4197)
WITH	WI (85), WAE (116)	WITH (4774)
YOU	YI (26)	YOU (10891)
YOUR	YER (237), YIR (11)	YOUR (3094), YOURE (915), YOU ARE (342)
YOURSELF	YERSEL (11)	YOURSELF (193)

Table 3: Variables used in our studies, with each variant's frequency per million tokens in the GS dataset

	Tweets w/ Yes or No hashtags	All tweets
$\hat{p}_{yes}$	0.00766	0.01443
$\hat{p}_{no}$	0.00211	0.00734
d	0.00555	0.00709
p-value	0.00103	0.00001

Table 5: Results of the two analyses in Study 1

usage rates. While the rates are very low overall, the relative differences are large: the ***Yes*** group rate is more than three times the ***No*** group rate when we include only tweets with Yes or No hashtags, and approximately twice as big when we include all tweets. The higher rates in the 'All Tweets' dataset suggest that both groups of users chose Scottish variants less often when discussing the referendum than in their other tweets. However, the test we used does not provide a significance value for the difference in usage rates across the two datasets. To establish whether users do modulate their usage of Scottish variants when discussing the referendum, we will need a more careful paired design.

6 Study 2: Effects of topic and audience on Scotland-specific vocabulary usage

Do tweeters choose Scottish variants at a different rate when using referendum-related hashtags than in their other tweets?

6.1 Method

We need a statistic that corrects for the fact that some variables might have higher rates of Scottish variants than others. For example if users tend to produce Scottish variants of variable v_1 at a higher rate than for v_2, and use v_1 more in tweets that don't contain referendum-related hashtags, then it could appear that users are suppressing their Scottish usage in referendum-related tweets when in fact this is a lexical effect.

Let U be the set of all users who have used at least one of the variables in Table 3 in both a tweet that contains a referendum-related hashtag (i.e. a tweet that belongs to the IT dataset, referred to hereafter as an Indyref tweet) and in a tweet that does not contain a referendum-related hashtag (referred to hereafter as a Control tweet). For a given user $u \in U$, let V be the set of all variables that u has used in at least one Indyref tweet, and in at least one Control tweet. Let $\hat{p}_{I,v}$ for user u be the estimated probability that u chooses a Scottish variant of variable $v \in V$, conditioned on the fact that she is using variable v in an Indyref tweet. Analogously, let $\hat{p}_{C,v}$ be the estimated probability that u chooses a Scottish variant of variable v, conditioned on the fact that she is using variable v in a Control tweet. The difference in user u's probability of choosing a Scottish variant of variable v in an Indyref tweet and in a Control tweet is then $d_v = \hat{p}_{I,v} - \hat{p}_{C,v}$. Averaging across all variables, we define $d_u = \frac{1}{V} \sum_{v \in V} d_v$.

The null hypothesis is that on average, users are no more or less likely to choose Scottish variants in Indyref tweets than in Control tweets. Therefore, under the null hypothesis, the mean value of d_u across all users, $\bar{d}_u = \frac{1}{U} \sum_{u \in U} d_u$, would be zero. We perform a one-sample t-test to determine whether $\bar{d}_u$ is significantly different than zero.

We use this method to conduct two separate analyses. In the first analysis, our pool of Control tweets is the set of *all* tweets from the original filtered dataset that do not contain any of the hashtags in Table 1. In the second analysis, we limit our pool of Control tweets to those which do not contain any of the hashtags from Table 1, but *do* contain at least one other hashtag. This second analysis is designed to test whether the recent finding that US Twitter users are less likely to use regionally-specific words in tweets which contain hashtags (Pavalanathan and Eisenstein, 2015) applies to Scottish users as well.

6.2 Results

The number of tweets and users that were included in each analysis are shown in Table 6.

Results for the first analysis are shown in the left column of Table 7. The difference is statistically significant ($p < 0.01$), indicating that on average, individuals are less likely to choose Scottish variants when using referendum-related hashtags than in their other tweets. Results for the second analysis are shown in the right column of Table 7. In this case, the difference is not statistically significant.

6.3 Discussion

In light of (a) the apparent relationship between national identity and constitutional preference, (b) the history of Scots as the prestige language of a previously-independent Scotland, supplanted by English in large part due to the birth of the United Kingdom, and (c) the results of Study 1, which indicate that pro-independence users choose Scottish variants at a significantly higher rate than anti-

	All Controls	Controls w/ Hashtags
# Users	11,011	7429
# Indyref Tweets	41,924	35,241
# Control Tweets	693,815	195,145

Table 6: Number of users and tweets included in the two analyses in Study 2

	All Controls	Controls w/ Hashtags
$\bar{d}_u$	-0.0015	-0.0010
std error	0.0005	0.0006
t-statistic	-2.996	-1.758
p-value	0.0027	0.0788

Table 7: Results of the two analyses in Study 2

independence users—it may at first appear surprising that people are *less* likely to choose Scottish variants in tweets containing referendum-related hashtags than in their other tweets.

It is conceivable that ***Yes*** users increase their rate of Scottish variants in Indyref tweets whilst ***No*** users decrease it, such that their effects cancel out; but since ***Yes*** users are more prolific in the IT dataset, if anything we would expect this imbalance to make the effect even more positive. The fact that we see a significant negative effect in spite of the greater number of ***Yes*** tweets means we can be reasonably confident that even if ***Yes*** users aren't significantly reducing their usage of Scottish variants in Indyref tweets, they certainly aren't increasing it.

It is also worth noting that we did not exhaustively identify every hashtag that has been used in relation to the referendum, so inevitably there will be some tweets with referendum-related hashtags in the Control set (such as example tweet (3) in §2), and there may also be some non-referendum tweets in the Indyref set. However, if anything this would dilute any differences between the two lists, yet we still find an effect.

The fact that this effect does not reach significance when we remove Control tweets without hashtags suggests that the primary reason users are reducing their rate of Scottish variants in Indyref tweets is not because of the *topic* under discussion, but because the use of hashtags broadens the potential audience. This explanation accords with Pavalanathan and Eisenstein's (2015) finding that amongst Twitter users in the US, non-standard and regional variants are less likely to be used in tweets that target larger audiences. Of course, it is possible that topic has an effect as well, but the present study does not provide evidence for that conclusion.

7 Conclusion

We presented the first large-scale study of distinctively Scottish language use on social media, showing that this use includes a mixture of traditional Scots vocabulary, newer Scottish slang, and alternative spellings that reflect Scottish pronunciation. We also studied how users' language might reflect their political views and discourse. We showed that ***Yes*** users use Scottish variants at a higher rate than ***No*** users, whether discussing the independence referendum or not. But overall, users tend to decrease their use of Scottish variants when discussing the referendum. This result suggests that although ***Yes*** users generally express a stronger Scottish linguistic identity than ***No*** users, they are not choosing to express this identity strongly in political discourse aimed at a broad audience. Due to the very low rates of Scottish variants overall, our data set is too small to study differences between individual variables or even conclusively say whether there may be effects of both topic and audience size on the use of Scottish language. However, we hope to be able to answer these questions in future by collecting a more complete set of data for the particular users studied here.

8 Acknowledgements

This work was supported in part by the EPSRC Centre for Doctoral Training in Data Science, funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016427/1) and the University of Edinburgh.

References

Adam J. Aitken, James A.C. Stevenson, Sir William Alexander Craigie, and Margaret G. Dareau. 1990. *A Dictionary of the Older Scottish Tongue Vols. 1-7: From the Twelfth Century to the End of the Seventeenth*. MacMillan Publishing Company.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

Lin Su Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social

media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130. Association for Computational Linguistics.

John Corbett, J. Derrick McClure, and Jane Stuart-Smith. 2003. A brief history of Scots. In John Corbett, J. Derrick McClure, and Jane Stuart-Smith, editors, *The Edinburgh Companion to Scots*, pages 1–16. Edinburgh, Edinburgh University Press.

Gabriel Doyle. 2014. Mapping dialectal variation by querying social media. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 98–106. Association for Computational Linguistics.

Jacob Eisenstein, Amr Ahmed, and Eric P. Xing. 2011. Sparse additive generative models of text. In *Proceedings of the International Conference on Machine Learning*, pages 1041–1048.

Jacob Eisenstein. 2015. Systematic patterning in phonologically-motivated orthographic variation. *Journal of Sociolinguistics*, 19(2):161–188.

Bruno Gonçalves and David Sánchez. 2014. Crowdsourcing dialect characterization through Twitter. *PloS one*, 9(11):e112074.

William Grant and David D. Murison. 1931. *The Scottish National Dictionary*. Scottish National Dictionary Association.

William Grant. 1931. Phonetic description of Scottish language and dialects. In *The Scottish National Dictionary*, volume 1, pages 9–41. Online: `http://www.dsl.ac.uk/about-scots/history-of-scots/`.

Yuan Huang, Diansheng Guo, Alice Kasakoff, and Jack Grieve. 2015. Understanding US regional linguistic variation with Twitter data analysis. *Computers, environment and urban systems*.

Taylor Jones. 2015. Toward a description of African American Vernacular English dialect regions using "Black Twitter". *American Speech*, 90(4):403–440.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2015. Challenges of studying and processing dialects in social media. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 9–18. Association for Computational Linguistics.

Billy Kay. 1988. *Scots: The Mither Tongue*. Grafton, first edition.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30. Association for Computational Linguistics.

Caroline Mcafee. 1985. Nationalism and the Scots renaissance now. In Manfred Görlach, editor, *Focus on: Scotland (Varieties of English around the world, V.5)*, pages 7–16. Amsterdam/Philadelphia, John Benjamins Publishing Company.

Dong-Phuong Nguyen, RB Trieschnigg, and Leonie Cornips. 2015. Audience and the use of minority languages on Twitter. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media*, pages 666–669.

Umashanthi Pavalanathan and Jacob Eisenstein. 2015. Audience-modulated variation in online social media. *American Speech*, 90(2):187–213.

ScotCen. 2013. Should Scotland be an independent country? (combined responses of those who have and those who haven't decided yet) broken down by 'Moreno' national identity. Retrieved from: `http://whatscotlandthinks.org/`. Accessed: 2016-09-30.

Scottish Language Dictionaries. 2004. Dictionary of the Scots language. `http://www.dsl.ac.uk/`. Accessed: 2016-12-20.

Rachael Tatman. 2015. #go awn: Sociophonetic variation in variant spellings on Twitter. *Working Papers of the Linguistics Circle of the University of Victoria*, 25(2):97–108.

4.4 Comments on the paper

In this paper we have established that use of the Scots language and regionally-specific terms and spellings is prevalent on Twitter and corresponds to features known in the linguistic literature about Scots and Scottish English, while we have also identified some new distinctively Scottish terms which are specific to social media text.

In Study 1 we found that distinctively Scottish variants were used at a higher rate by users who predominantly used pro-independence hashtags (the ***Yes*** group) than by those who predominantly used anti-independence hashtags (the ***No*** group). A limitation of this study is that the IT dataset did not contain enough geotags to characterise the geographic distribution of the tweets within it (while the GS dataset did not contain enough tweets with indyref hashtags to be used for the main analyses). Since we lack geolocation information, we can't rule out the possibility of an imbalance in the proportions of users who are from Scotland in the ***Yes*** and ***No*** groups; for example, it could be that a greater proportion of users in the ***No*** group are from outwith Scotland, and hence less likely to even have the potential to index a Scottish identity through the use of distinctively Scottish variants.

While the low absolute rates of Scottish variant usage suggest that the IT dataset does contain many users who do not have these variants in their verbal repertoires at all, this is less of a concern for Study 2, where we looked at *intraspeaker* variation. Study 2 is concerned with relative usage rates by the same set of users across different contexts, so if many users do not use distinctively Scottish variants at all, this may affect the size of the effect, but not its direction. However, this study is limited by the fact that its design does not enable us to distinguish the effect of a tweet being about the Scottish independence referendum from the effect of the audience-broadening function of a hashtag.

We address these limitations in Chapter 5, wherein we use a topic model rather than relying on hashtags to classify the topics of tweets, and we use a more sophisticated method of analysis to model effects of both audience and topic on the use of Scotland-specific variants, whilst controlling for variation in the base rate of Scottish variant usage across different users and variables.

4.5 Follow-ups and future work

The studies presented in this chapter were replicated by Stewart et al. (2018) on a dataset of tweets related to the Catalonian referendum. Like us, they found more minority language use among those who used pro-independence hashtags. However, contrary to our Study 2 results regarding usage of distinctively Scottish lexis, they found that usage of Catalan was *more* likely in tweets with hashtags relating to the referendum than in other tweets by the same users. A potential explanation for this difference is the fact that, as Stewart et al. note, the Catalan language plays an integral part in the Catalonian nationalist narrative; while conversely the Scottish National Party explicitly promoted a narrative of *civic* nationalism during their referendum campaign, deliberately seeking to distance themselves from *ethnic* nationalism which would base the case for Scottish independence on claims to a distinctive cultural and ancestral heritage (McAnulla and Crines, 2017). Civic nationalism focuses more on distinctive political and economic institutions and practices, and enables any individual resident in Scotland to legitimately identify as Scottish, regardless of their birthplace, ancestry, or the language they speak (Soule et al., 2012)[1]. Furthermore, Catalan enjoys much more prevalence and prestige within Catalonia than the Scots language does in Scotland: Catalan is the official language of Catalonia and is very widely spoken there, including as a medium of education, while Scots is still quite stigmatised even within Scotland, and only fairly recently gained recognition as a *minority* language in Scotland.

It would be interesting to conduct similar studies with respect to Scotland's other indigenous minority language, Scottish Gaelic. Another potential direction for future work (which would require a larger collection of geolocated tweets) would be to analyse the geographical dispersion of the different distinctively Scottish variants we have identified. In the studies presented here we have treated all of these variants as if they belong to one homogeneous language variety, but in reality the Scots language consists of multiple dialects, and Scottish English accents also vary regionally. For example, the variants HING, HINK, and SUHIN all reflect th-debuccalisation, a phonological feature traditionally associated with working-class speech in Glasgow (Stuart-Smith et al., 2007). It would be interesting to see whether usage of these variants on Twitter is also localised, and to what extent dialect levelling has occured. It would also be interesting

[1] That being said, Mycock (2012) has argued that the SNP's claim to progressive civic nationalism "is an aspiration rather than a reality", noting that they do sometimes opportunistically flag up grievances around the historical oppression of Scotland's indigenous languages, and have asserted that the Scots language 'is part of our identity and our heritage as a nation' (Mycock, 2012)

to investigate whether there are differences in usage patterns between the variant forms that are attested in Scots dictionaries, and those that are not. Finally, the time dimension could be exploited to investigate whether usage patterns changed throughout the referendum campaign and its aftermath.

Chapter 5

Teasing apart topic and audience

5.1 Introduction

In Chapter 4 we found that users tended to reduce their rates of distinctively Scottish vocabulary usage in tweets which contained hashtags relating to the Scottish Independence Referendum. Although our intent had been to test whether users modulate their usage of Scottish variants when tweeting about the referendum, what we actually measured in that study was the effect of including referendum-related *hashtags* in a tweet, which can function not only as markers of topic but *also* to broaden the potential audience. In this chapter we present a follow-up study, in which we infer tweet topics using topic models as opposed to hashtags, and use mixed effects logistic regression to tease apart the effects of tweet topic and expected audience size (operationalised in terms of hashtag and mention use) on usage rates of distinctively Scottish lexis in two datasets of tweets by distinct user samples.

These two datasets cover the same time period but are sampled from two distinct (though slightly overlapping) populations: users who have posted tweets with Scottish geotags, and users who have posted tweets containing referendum-related hashtags. Both of these are populations which we could reasonably expect to include users with distinctively Scottish variants in their verbal repertoires, but studies suggest that geotags are more likely to be used by younger users (see §3.3.1), while we may assume that the demographic profile of users of political hashtags is likely to skew older and diverge from that of the general Scottish Twitter population in other ways. Analysing these two samples side-by-side enables us to examine how biases implicit in the construction of datasets can substantively affect results.

We find that in both user groups, topic and audience have independent effects on

rates of Scottish variant usage, providing stronger evidence than in previous work that users are indeed sensitive to their audience. While supporters of Scottish independence are sometimes accused of ramping up their use of the Scots language in order to accentuate perceived sociocultural differences between Scotland the rest of the UK (Clark, 2018), we observe no evidence for this on a broad scale. In fact, our results indicate that people are significantly more likely to use distinctively Scottish vocabulary in everyday chitchat on Twitter than when discussing Scottish indepedence. While the effects of topic are quantitatively similar across our two user groups, the effects of audience diverge. For the geotag users, rates of Scottish variant usage follow the pattern predicted by previous research: lowest among tweets with the largest expected audience, and rising as the expected audience size shrinks. In contrast, the indyref hashtag group shows a less consistent and less pronounced pattern which does not align cleanly with audience size. This highlights the difficulty of sampling representative groups from social media data, and underscores the importance of replicating studies on distinct user samples before drawing strong conclusions.

5.2 Author contributions

The paper is co-authored by me, James Kirby, and Sharon Goldwater. As the leading author, I wrote the code, collected the data, performed the analysis, and wrote the paper. James Kirby and Sharon Goldwater supervised the project, offered suggestions, and helped to revise the final manuscript.

5.3 The paper

The paper was accepted for publication at the Workshop on Stylistic Variation at ACL 2017 in Berlin. The publication reference is as follows:

> Shoemark, P., Kirby, J., & Goldwater, S. (2017, September). Topic and audience effects on distinctively Scottish vocabulary usage in Twitter data. In *Proceedings of the Workshop on Stylistic Variation* (pp. 59-68).

Topic and audience effects on distinctively Scottish vocabulary usage in Twitter data

Philippa Shoemark[*]
p.j.shoemark@ed.ac.uk

James Kirby[†]
j.kirby@ed.ac.uk

Sharon Goldwater[*]
sgwater@inf.ed.ac.uk

[*]School of Informatics
University of Edinburgh

[†]Dept. of Linguistics and English Language
University of Edinburgh

Abstract

Sociolinguistic research suggests that speakers modulate their language style in response to their audience. Similar effects have recently been claimed to occur in the informal written context of Twitter, with users choosing less region-specific and non-standard vocabulary when addressing larger audiences. However, these studies have not carefully controlled for the possible confound of topic: that is, tweets addressed to a broad audience might also tend towards topics that engender a more formal style. In addition, it is not clear to what extent previous results generalize to different samples of users. Using mixed-effects models, we show that audience and topic have independent effects on the rate of distinctively Scottish usage in two demographically distinct Twitter user samples. However, not all effects are consistent between the two groups, underscoring the importance of replicating studies on distinct user samples before drawing strong conclusions from social media data.

1 Introduction

Linguistic variation in social media is a growing research area, with interest stemming both from the engineering goal of developing tools that work well across different styles and dialects (Hovy, 2015; Stoop and van den Bosch, 2014; Vyas et al., 2014; Huang and Yates, 2014), and from the social science goal of studying user behaviour (Bamman et al., 2014; Eisenstein, 2015; Huang et al., 2016; Nguyen et al., 2015). However, this type of research is often complicated by the messy nature of social media data, which can make it hard to control for different explanatory factors and to know whether results obtained on a particular user sample generalize to another sample.

For example, previous studies have suggested that Twitter users modulate their use of regional and non-standard language depending on the expected size of the audience (operationalized as whether a Tweet contains hashtags, @-mentions, or neither) (Pavalanathan and Eisenstein, 2015a; Shoemark et al., 2017). However, these studies did not sufficiently control for possible effects of topic, which may be confounded with audience size: e.g., users may use more hashtags when discussing political events than when discussing daily routines. These studies also did not look at the degree to which their results generalize across different populations of users.

In this work we study two largely disjoint groups of (mainly) Scottish Twitter users: one group sent tweets geotagged within Scotland, while the other used hashtags related to the 2014 Scottish independence referendum. We use mixed-effects models to tease apart the effects of audience and topic on their choice of Scottish-specific terms. We find that in both user groups, topic and audience have independent effects on the rate of Scottish usage, providing stronger evidence than in previous work that users are indeed sensitive to their audience.

Nevertheless, our study does not confirm all aspects of previous work. When comparing our two user groups, the effect of topic is qualitatively similar: tweets about lifestyle or politics have lower rates of Scottish usage than "chitchat" tweets. However, the effects of audience differ between the two groups. For the geotagged users, rates of Scottish usage follow the pattern predicted by previous research: lowest among tweets with the largest expected audience, and rising as the expected audience size shrinks. In contrast, the independence referendum group showed a less consistent and less pronounced pattern which does not align cleanly

with audience size. We were unable to find a clear explanation of this difference. Nevertheless, it highlights the difficulty of sampling representative groups from social media data and the need to interpret results with caution until they are shown to generalize across several different populations.

2 Background

Bell's (1984) Audience Design theory posits that intra-speaker stylistic variation is primarily conditioned by the audience of the interaction. Bell argues that stylistic variation across topics derives from so-called 'reference groups' whom the speaker associates with the topics in question, and predicts that effects of topic on style variation will be weaker than direct effects of audience. However, later studies of spoken conversation (e.g. Rickford and McNair-Knox, 1994) have suggested that both topic and audience affect a speaker's style, and that topic may even have a greater effect. Topic also appears to influence stylistic variation in computer-mediated communication—for example, statistical associations between lexical features and author attributes such as gender are often mediated by the topic of discourse (Herring and Paolillo, 2006; Bamman et al., 2014).

Our work is primarily inspired by two previous studies of Twitter users and how their use of regional lexical variants is influenced by either audience (Pavalanathan and Eisenstein, 2015a) or topic (Shoemark et al., 2017). In the first of these, Pavalanathan and Eisenstein (2015a) studied lexical items that were strongly associated with tweets from specific regions of the US, as determined by a data-driven approach (Eisenstein et al., 2011). They found that users were less likely to use these regional terms, as well as other nonstandard terms, in tweets containing hashtags, and more likely to do so in tweets containing @-mentions (i.e., other users' IDs). They attributed these findings to style-shifting in relation to audience size, since tweets with hashtags are more likely to be viewed by users outside of the author's follower group, while by default tweets which begin with a mention are shown only to the author, the mentioned user, and their mutual followers.

While suggestive, there are alternative explanations for this finding. For example, in their study of Scottish tweets, Shoemark et al. (2017) pointed out that if users use the word 'masel' (a Scottish variant of standard English 'myself') less frequently in tweets with hashtags, it could be simply because people talk about themselves less in tweets with hashtags, not because they are modulating the use of a regionally specific variant.

Shoemark et al. (2017) focused mainly on effects of topic rather than audience, but to avoid similar confounds, they measured the frequencies of regional variants of lexical variables[1] *relative* to their standard variants. They found that, amongst users who tweeted about the Scottish independence referendum, both pro- and anti-independence users decreased their use of Scottish-specific terms in tweets containing referendum-related hashtags, compared to other tweets. A follow-up analysis suggested that this effect might be due to the larger audience obtained by using referendum-related hashtags, but the evidence was indirect as the original study was not designed to test that hypothesis.

Our work extends these two previous studies by building models that include factors for both topic and audience. We follow Shoemark et al. (2017) in focusing on variables that alternate between Scottish English and Standard English variants, but use a wider range of topics identified with a topic model rather than just hashtags. We use mixed-effects logistic regression in order to establish whether there are independent effects of audience and topic, whilst controlling for variation in the base rate of Scottish-variant usage across different users and variables. In addition, we explicitly examine how different methods of sampling users might affect results, by performing the same study on two user groups gathered in different ways.

3 Data

3.1 Lexical variables

We use 50 of the 51 lexical variables identified by Shoemark et al. (2017). Each variable consists of one or more distinctively Scottish variants and one or more Standard English variants, all of which are referentially and syntactically equivalent; examples are shown in Table 1. From the original 51 variables, we discard **SHIT**, since the variant identified as Scottish-specifc, SHITE, is used at a higher rate than the Scottish-specific forms of the other variables (e.g. 27% of **SHIT** occurences in Shoemark et al.'s Indyref-Tweets dataset are realized as SHITE; more than twice the rate of Scottish variant use for any other variable), and for many users **SHIT** is the

[1] A *variable* is any linguistic item than can be produced in different ways; the *variants* are the different realizations.

Variable	Scottish variants	Std variants
DONT	DEH, DINI , DINNY	DONT, DO NOT
FOOTBALL	FITBA	FOOTBALL
MYSELF	MASEL, MASELF	MYSELF
SOMETHING	SUHIN	SOMETHING
TO	TAE	TO, TOO

Table 1: Examples of lexical variables.

only variable for which any Scottish variant use is observed. This suggests that SHITE is less marked as 'distinctively Scottish' than the Scottish-specific variants of the other 50 variables.

3.2 Dataset construction

We aim to study Scottish language use, but only a small proportion of Twitter users disclose their location, either by including it in their user profile or by opting to automatically tag their tweets with geographic coordinates when using a GPS-enabled device. Moreover, studies have indicated that those who do share their location are not representative of the wider Twitter user base (Pavalanathan and Eisenstein, 2015b; Sloan and Morgan, 2015).

To help assess the generalizability of our findings, we therefore consider two datasets, both covering the same time period but sampled from distinct (though slightly overlapping) populations: 'Scottish Geotag Users', who have tagged their tweets with locations in Scotland; and 'Indyref Hashtag Users', who have used hashtags relating to the 2014 Scottish Independence Referendum. As we will demonstrate, users in the two samples do differ in some aspects of their behaviour, emphasizing how biases implicit in the construction of datasets can affect results.

Our two groups of users are taken from the Geotagged-Scotland (GS) and Indyref-Tweets (IT) datasets collected by Shoemark et al. (2017). Both of these datasets were drawn from an archive of Twitter's 'Spritzer' stream, which provides a 1% sample of the public data flowing through Twitter, covering the period from September 2013 to September 2014. The GS dataset consists of tweets by users for whom the archive contained at least one tweet which was geotagged with a location in Scotland, while the IT dataset consists of users for whom it contained at least one tweet with a hashtag relating to the 2014 Scottish Independence referendum (see Table 3 in Shoemark et al. (2017) for a list of hashtags).

As a heuristic to filter out bots and spammers,

		IH Users	SG Users
(a)	N Users	14,572	17,942
	N Tweets	4,703,040	1,750,343
	N Variables	10,482,683	3,733,133
	% Scottish	0.5	1.8
(b)	N Users	12,101	11,307
	N Tweets	4,674,251	1,678,498
	N Variables	10,424,067	3,594,659
	% Scottish	0.5	1.8
(c)	N Users	10,786	10,103
	N Tweets	3,456,277	1,371,694
	N Variables	7,689,621	2,878,352
	% Scottish	0.7	2.3
(d)	N Users	10,784	10,103
	N Tweets	2,165,320	1,112,931
	N Variables	4,934,186	2,365,496
	% Scottish	0.8	2.3

Table 2: Dataset statistics for Indyref Hashtag Users and Scottish Geotag Users (a) after basic preprocessing, (b) after discarding users with <50 variable instances, (c) after discarding users for which there is strong evidence of non-use of Scottish variants and (d) after labelling audience & topic. '% Scottish' is the percentage of variables realized as the Scottish variant.

we computed the proportion of tweets for each user in the GS and IT datasets which contained URLs, and discarded users for whom this proportion was in the 90th percentile. For the remaining users, we then retrieved a more complete set of their tweets: for each user we attempted to retrieve all the tweets they posted in August, September, or October 2014 (excluding retweets), using Twitter's REST API. The API allows us to retrieve up to 3200 of a user's most recent tweets, so if a user had posted more than 3200 tweets since autumn 2014, we were unable to retrieve their tweet histories for this period. We obtained complete histories for at least one of the three months for a total of 18,370 Scottish Geotag (SG) Users, and 14,832 Indyref Hashtag (IH) Users. We then applied some simple ad-hoc text filters to remove tweets produced by apps which automatically share user's horoscopes or track users' follower counts, as well as some particularly prevalent types of marketing tweets. See Table 2a for summary statistics after this filtering step. Note that there are 363 users who are in both datasets.

Next, we removed all users for whom the total number of observed variable instances was less than 50 (see Table 2b), as with so few observations it would be difficult to make reliable inferences about these users' usage rates of distinctively Scot-

tish variants.

Finally, since our population of interest is those who vary between Scottish and standard variants, we discard individuals for whom we had enough observed variable instances to conclude that they probably *never* used distinctively Scottish variants of any of our variables. For SG Users, we chose the threshold of 'enough observed variable instances' to be 298, since this is the smallest value n such that the cumulative binomial probability of seeing at least one Scottish variant in n variable instances is ≥ 0.99 (assuming a constant usage rate of Scottish variants of 0.0184, as listed in Table 2b). That is, if we assume that any user who does use Scottish variants will do so 1.84% of the time, then in 99% of cases where we have observed at least 298 variable instances from such a user, we would expect a Scottish variant to have been used in at least one of those instances. For IH Users, we assumed a constant usage rate of distinctively-Scottish variants of 0.05, and discarded all those for whom we had observed at least 870 variable instances and no Scottish variants. Table 2c provides summary statistics for the two resulting datasets.

When considering the differences in average rates of Scottish variant usage across the two groups, it is important to note that Shoemark et al. (2017) identified these Scottish variants using the GS dataset, i.e. the same dataset from which we drew our Scottish Geotag Users. It is therefore to be expected that that the Scottish Geotag Users would use these variants at a higher rate, and it is important to bear in mind that the Indyref Hashtag Users may be more frequent users of other distinctively Scottish variants.

4 Topic & Audience

4.1 Audience labelling

We follow Pavalanathan and Eisenstein (2015a) in assuming that tweets containing hashtags (any token prepended with the '#' character) typically have a wider audience than other tweets, since anyone interested in a particular topic or event can browse the stream of Tweets which contain associated hashtags. Conversely, tweets beginning with @-mentions typically have a narrow audience since by default they only appear in the feeds of the author, the mentionee, and users who follow both the author and the mentionee. Any user @-mentioned in a tweet (whether at the beginning, or elsewhere within the tweet) will by default receive a special notification drawing their attention to it.

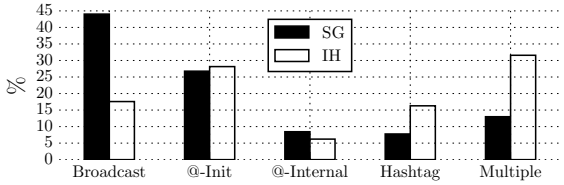


Figure 1: Distribution of tweets with each audience label in the two datasets.

Pavalanathan and Eisenstein hypothesise that both kinds of mention serve to narrow the intended audience, whilst hashtags serve to widen it, relative to broadcast tweets (i.e., those without hashtags or mentions, which appear on the feeds of all the author's followers). The grounds for hypothesising a narrowing function for tweet-internal mentions are less evident than those for tweet-initial mentions, since tweets which do not begin with a mention are *not* limited by default to the feeds of the author and mentionee's mutual followers.

We label each variable instance in our two datasets with three binary variables indicating whether or not they contain hashtags, initial mentions, and/or internal mentions. We then discard any tweets for which two or more of these indicators are activated, since we do not have intuitive a priori hypotheses about how combining more than one of these variables within a single tweet would affect its intended audience.

Figure 1 shows the proportion of tweets in each dataset which have each audience label (or which had multiple labels and were subsequently discarded), and reveals qualitative differences in the two groups' behaviour: SG Users post relatively more 'broadcast' tweets, whilst IH Users use relatively more hashtags (which is unsurprising given that they were selected on the basis of their hashtag use).

4.2 Topic labelling

We assign topics to tweets using a Latent Dirichlet Allocation (LDA) topic model (Blei et al., 2003) estimated with collapsed Gibbs sampling (Griffiths and Steyvers, 2004) from both datasets combined. Following Hong and Davison (2010) and others, we create 'documents' by concatenating together tweets by the same author. To account for possible topic drift within individuals over time, we group each user's tweets by month and model each per-user-per-month document as a distinct mixture of

topics. We use the inferred topic model parameters to label each tweet with a topic, as described below.

The corpus was preprocessed as follows: tweets were tokenised using the Twokenize program[2], a tokeniser designed specifically for Twitter text, and all non-alphabetic tokens, except for those which begin with hashtags, were discarded. The vocabulary was then pruned to the 100,000 most frequent terms across the two datasets. We set the number of topics, T, to 30, and used symmetric Dirichlet priors of $\alpha = \frac{50}{T}$ and $\beta = 0.01$ on the multinomial distributions over topics and terms, respectively[3]. The Gibbs sampler was run for 750 iterations.

Upon inspection of the most probable words and documents for each topic, we deemed that twenty of the topics could be grouped into three broader themes, which we describe as 'chatter' , 'lifestyle' , and 'politics' . Later, we consider a different grouping, where we split off a 'sports' theme from the 'lifestyle' theme, and an 'indyref' theme from the 'politics' theme. Table 3 shows the most probable words (excluding stopwords) for each topic within these three/five themes. Of the ten topics that we did not assign to these themes, four could be described as spam topics, four as foreign language, and two as relating to purely stylistic dimensions as opposed to any particular topic of discussion: one for distinctively Scottish terms, and the other for 'netspeak'-style spellings and abbreviations.

To assign topic labels to individual tweets, we take a Gibbs sample and then for a given tweet, each topic t is assigned a weight, defined as

$$\text{weight}_t = \sum_{w \in \boldsymbol{w}} \hat{p}(t|w)$$

where $\boldsymbol{w}$ is the bag of words which occur in the tweet (excluding stopwords and any variant of any of our variables of interest), and $\hat{p}(t|w)$ is obtained by maximum likelihood estimation from the Gibbs-sampled topic-token assignments. Finally, we take the topic with the highest weight, and label the tweet with its broader theme. If the topic with the highest weight is one of the two 'stylistic' topics, we defer to the topic with the next highest weight. We discard tweets labelled as 'spam' or 'foreign language', as well as those for which the highest weight is not unique, if the topics which share this weight belong to different themes.

[2]https://github.com/myleott/ark-twokenize-py

[3]During development we experimented with values for T between 10 and 100, and α between 0.015 and 1.5, and saw little qualitative difference in the themes that emerged, based on manual inspection of topic keywords.

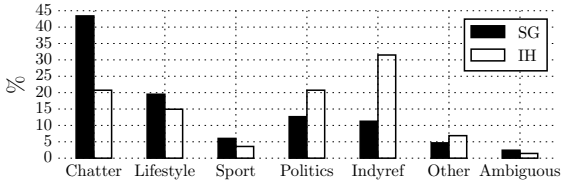


Figure 2: Distribution of tweets with each topic label in the two datasets.

Using this method, we obtain 2.3m broad-topic-labeled variable instances from SG Users, and 4.9m from IH Users. Figure 2 shows the distribution of topics in each data set, and Table 4 gives a breakdown of variable instances by audience-type and broad-topic-label. IH Users have a much larger proportion of tweets with 'indyref' or 'politics' labels than SG Users, which once again is unsurprising, given how these users were sampled.

5 Method

We use the glmer() function from the lme4 package (Bates et al., 2015) for R (R Core Team, 2013) to estimate mixed effects logistic regression models, predicting Scottish variant usage (yes = 1, no = 0) from the intended audience size and topic of the tweet in which a lexical variable occurs. Our four-level categorical audience factor (initial mention, internal mention, broadcast, hashtag) is dummy coded into three binary variables, with broadcasts as the reference level. Our tweet topic labels are also dummy coded, taking the 'chatter' topic as the reference level. By specifying random effects for users and variables, we control for the influence of different baseline rates of Scottish variant usage across different users and variables. Hence our models are of the form

$$\text{logit}\{E(\mathbf{y})\} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \mathbf{y} \sim Bernoulli$$

where $\mathbf{y}$ is the $n \times 1$ vector of responses from a Bernoulli distribution, $\mathbf{X}$ is an $n \times p$ design matrix for the fixed effects $\boldsymbol{\beta}$, and $\mathbf{Z}$ is an $n \times q$ design matrix for the random effects $\mathbf{u}$. We do not include random slopes in our models, since we do not have enough observations per group to provide stable estimates of the variances. Our models are fit by Laplacian approximation to Maximum Likelihood estimation.

Topic theme	Keywords
Chatter	love feel life fucking fuck people shit actually hate omg school gonna time excited oh
	time yeah bit oh probably actually maybe seen lot pretty hope haha bad getting stuff
	lol love thank xx thanks hope day oh happy lovely xxx ha haha morning beautiful
	night happy birthday haha day wait tonight tomorrow hahaha bed getting wee weekend days week
Lifestyle	love song music album world amazing god top white black girl watch band ice looks baby life listen guy boys
	photo watching #xfactor #cbb day #scotland loving posted #gbbo life #glasgow #bbuk #love #edinburgh love
	video #auspol liked game awesome watch time apple iphone play app games phone buy facebook
	oh bit news ha twitter story brilliant bbc read book called tv look dear wonder
	day time morning night car run food bit nice week train getting tea eat days
	tonight day week time tomorrow night glasgow morning looking edinburgh forward coming weeks hear live
Sports	cup win ireland #glasgow2014 irish time team final match scottish round top games race live
	game celtic team football season league fans mate goal win play players club player haha
Politics	people read agree question thanks issue debate political article course mean change indeed etc politics
	news police pm russia minister russian via eu report ukraine president ebola court uk #ukraine #russia
	#ferguson rt obama #ukraine police #cdnpoli ukraine video via mt people news american time america
	labour uk ukip cameron party tory ed tax vote tories english mps miliband boris david
	people lol look tell money time stop wrong please believe mean job care saying talking
	israel #gaza war via isis gaza #isis world people children israeli #israel police hamas support
Indyref	#indyref scotland #voteyes #yes vote scottish independence #scotdecides #indyrefpic #bettertogether salmond #bbcindyref #the45 campaign debate
	scotland vote uk labour scottish snp scots union oil party wm country westminster voters voting

Table 3: Topic themes and the top 15 keywords for each topic within each theme

	Audience \ Topic	Chatter	Lifestyle	Politics	All
(a)	Broadcast	598,673 (2.7)	334,143 (2.3)	295,981 (1.8)	1,228,797 (2.4)
	Initial Mention	352,981 (3.0)	164,909 (2.9)	188,191 (1.9)	706,081 (2.7)
	Internal Mention	92,682 (1.8)	63,242 (1.5)	56,727 (1.2)	212,651 (1.6)
	Hashtag	67,630 (1.8)	69,833 (1.4)	80,504 (1.2)	217,967 (1.4)
	All	1,111,966 (2.7)	632,127 (2.3)	621,403 (1.7)	2,365,496 (2.3)
(b)	Broadcast	308,797 (1.3)	341,592 (0.9)	658,520 (0.8)	1,308,909 (1.0)
	Initial Mention	644,459 (1.1)	394,036 (1.0)	1,026,634 (0.6)	2,065,129 (0.8)
	Internal Mention	76,403 (0.6)	96,123 (0.5)	203,275 (0.4)	375,801 (0.5)
	Hashtag	124,333 (0.7)	197,925 (0.5)	862,089 (0.5)	1,184,347 (0.5)
	All	1,153,992 (1.1)	1,029,676 (0.8)	2,750,518 (0.6)	4,934,186 (0.8)

Table 4: Counts of variable instances in the (a) SG Users and (b) IH Users datasets, broken down by Topic and Audience. In each cell, the percentage of variable instances that are Scottish variants is given in parentheses.

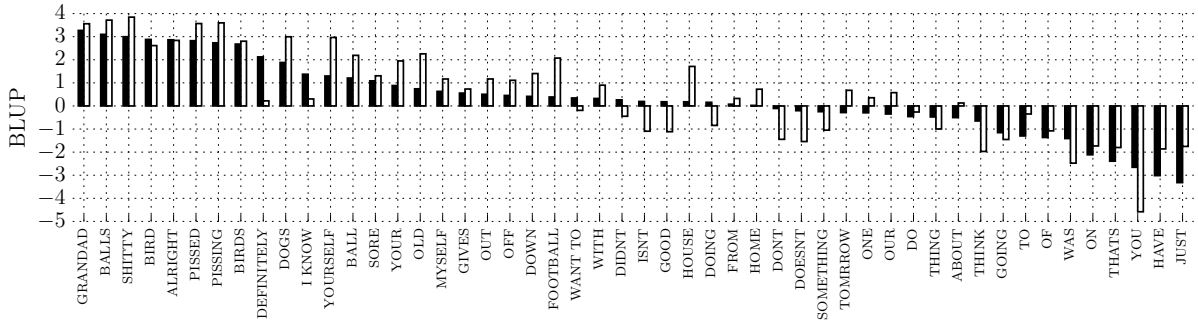


Figure 3: Barplots of by-variable BLUPs for SG Users (black bars) and for IH Users (white bars).

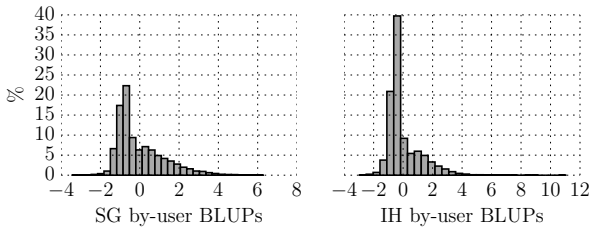


Figure 4: Histograms of by-user BLUPs.

6 Results and Discussion

6.1 Random Intercepts

We begin by constructing null models that predict the log odds of choosing a Scottish variant using only intercepts, which we allow to vary randomly by each user and by each lexical variable. The estimated variances of the by-user and by-variable adjustments to the intercept are given in Table 5a, for SG and IH Users, respectively.

The Best Linear Unbiased Predictors (BLUPs) of the by-variable random intercerpts (i.e. the posterior estimates, given the data and model parameters, of the adjustment to the intercept for each variable) are shown in Figure 3. In both datasets, open class variables (e.g. **GRANDAD**, **BALLS**, **DOGS**) tend to have higher rates of Scottish variant usage than closed class variables (e.g. **WAS**, **OF**, **YOU**).

Figure 4 shows the distributions of by-user BLUPs. Although the models assume a normal distribution over the by-user intercepts, the BLUPs are positively skewed. We suspect the BLUPs reflect the fact that our datasets contain a mixture of two populations: Scottish speakers, who use Scottish variants at a range of rates, and non-Scottish speakers, who rarely if ever use Scottish variants. The non-Scottish speakers are responsible for the large number of users with slightly negative intercepts. Unfortunately there is no straightforward way to separate these groups (especially for users with a relatively small number of observations). However, users with a constant near-zero rate of Scottish variant usage should, at worst, dilute any true effects of topic or audience on rates of usage, but should not change the direction of those effects.

6.2 Random Intercepts + Audience Effects

We now check whether Pavalanathan and Eisenstein's (2015a) reported effects of hashtags and mentions on the odds of using regional variants in US tweets, are replicated for distinctively Scottish variants in our two datasets.

We augment our null models with our dummy-coded audience factors as fixed effects. For each dataset, we assess the goodness-of-fit using chi-square tests on the log-likelihoods. Compared to the null models with only random effects, including fixed effects for audience significantly improves the fit on both datasets (SG: $\chi 2(3) = 643.05$, p = <2.2e-16; IH: $\chi 2(3) = 232.69$, p = <2.2e-16).

Parameters of the models with Audience effects are in Table 5b. Our results for SG Users largely accord with those of Pavalanathan and Eisenstein (2015a): Scottish variants are positively associated with tweet-initial mentions, and negatively associated with hashtags. Relative to broadcast tweets, the odds of seeing Scottish variants are about 28% higher in tweets with initial mentions, and about 17% lower in tweets with hashtags. However, while Pavalanathan and Eisenstein also found an association between the use of tweet-internal mentions and local/non-standard words in their data, our model does not show such a relationship in the SG dataset.

In the IH dataset, the audience effects in our model do not follow the pattern that Pavalanathan and Eisenstein observed in US tweets. Unlike for SG Users, there is no association between hashtags and Scottish variants, and the effects of mentions are in the opposite direction to those found by Pavalanathan and Eisenstein (2015a). Amongst

IH Users, initial mentions are *negatively* associated with Scottish variants, though the effect size is very small. Internal mentions are also negatively associated with Scottish variants, and in this case the effect is relatively large (in contrast with SG Users, for whom we found no effect). We discuss possible reasons for this result in Section 6.4.

6.3 Random Intercepts + Topic Effects

Next, we test for a relationship between the topic of a tweet and the odds of Scottish variant usage. For both datasets, models with fixed effects for topic significantly improve the fit over random-effects-only models (SG: $\chi2(2) = 570.48$, p = <2.2e-16; IH: $\chi2(2) = 1241$, p = <2.2e-16).

Parameters of the models are in Table 5c. The effects of tweet topic are qualitatively similar in each dataset: relative to 'chatter' tweets, tweeting about the 'lifestyle' topic reduces the odds of choosing Scottish variants by 11% for SG Users and 5% for IH Users, while tweeting about politics reduces them by 27% for SG Users, and 39% for IH Users.

6.4 Full Models

For each dataset, including fixed effects for audience and topic together significantly improves the model fit, both over the models with fixed effects for audience only (SG Users: $\chi2(2) = 508.67$, p = <2.2e-16; IH Users: $\chi2(2) = 1298.9$, p = <2.2e-16), and over the models with fixed effects for topic only (SG: $\chi2(3) = 581.25$, p = <2.2e-16; IH: $\chi2(3) = 290.6$, p = <2.2e-16).

Parameters of the full models are in Table 5d. When fixed effects for audience and topic are included together in the SG model, their effect sizes barely change. These results suggest that for SG Users, audience and topic have independent effects on Scottish usage, and that even after accounting for topic, the effects of audience size are as predicted by Pavalanathan and Eisenstein (2015a).

In the full IH model, while most of the fixed effect sizes are relatively unchanged, a positive association between the use of hashtags and Scottish variants emerges. Thus, the model reveals that the qualitative behavior of these users is very different from that of the SG Users. Although topic and audience are both significant factors in the models for each group, initial mentions and hashtags have the opposite effects for IH Users as for SG Users (and for Pavalanathan and Eisenstein's user sample).

Although they primarily interpret their findings in terms of audience size, Pavalanathan and Eisenstein acknowledge that mentions and hashtags can affect the composition of the audience in more nuanced ways than just its size. As an alternative explanation for the positive associations they found between mentions and local/non-standard words, they suggest that authors may use such words to express particular identities or stake claims to local authenticity, specifically when addressing users for whom such claims are meaningful.

In theory, this account could also apply to the positive association we find in the IH dataset between *hashtags* and local variants: while on the one hand, the indexing function of hashtags can be conceived of as broadening the audience of a tweet, on the other hand it could serve to narrow the tweet's intended audience, by explicitly targeting it at a circumscribed community. Hence, when using hashtags associated with communities for whom the notion of Scottish identity has strong currency (e.g. people with strong views on indyref, or supporters of a particular sports team), IH Users may use Scottish variants initiatively, in order to emphasise that part of their identity.

Suppose that authors tended to decrease their use of Scottish variants when discussing most political issues, but increase it when discussing Scottish independence—either to emphasise their own Scottish identity, or to accommodate towards an audience which is likely to contain many Scottish people. If this were the case, our models would be unable to account for this variation directly, since we have grouped indyref and other political issues together. However, since a large proportion (55%) of IH Users hashtag tweets are actually about indyref, one way the IH model could account for a difference between indyref and general politics is to increase the weight for hashtags. If this were the case, then including 'indyref' as a distinct topic should improve the model fit and alleviate the impact on the audience weights. To test this hypothesis, we performed a follow-up study where we split the topics into finer-grained categories.

6.5 Finer-grained topics

We added two topic categories, 'sport' and 'indyref', which we split off from the 'lifestyle' and 'politics' categories, respectively (see Table 3). Contrary to our hypothesis, re-defining the topic categories in this way made little difference to the model fit: the log-likelihoods for the new full model are -174169.4 for SG Users, and -121447.8

	Scottish Geotag Users				Indyref Hashtag Users			
(a)	*Log-likelihood:* -174758.0				*Log-likelihood:* -122240.2			
	σ^2 *users:* 2.769		σ^2 *variables:* 2.477		σ^2 *users:* 3.058		σ^2 *variables:* 3.444	
(b)	*Log-likelihood:* -174436.4				*Log-likelihood:* -122123.9			
	σ^2 *users:* 2.750		σ^2 *variables:* 2.503		σ^2 *users:* 3.039		σ^2 *variables:* 3.443	
Fixed Ef.	*OR*	*95% CI*	*z*	*Pr (>|z|)*	*OR*	*95% CI*	*z*	*Pr (>|z|)*
@init	1.28	(1.25, 1.31)	21.2	<2e-16	0.96	(0.93, 0.99)	-2.8	0.005
@intrnl	0.96	(0.92, 1.00)	-1.9	0.052	0.62	(0.59, 0.67)	-15.4	<2e-16
hashtag	0.83	(0.80, 0.86)	-8.9	<2e-16	0.97	(0.93, 1.01)	-1.6	0.111
(c)	*Log-likelihood:* -174472.7				*Log-likelihood:* -121619.7			
	σ^2 *users:* 2.758		σ^2 *variables:* 2.472		σ^2 *users:* 3.069		σ^2 *variables:* 3.427	
Fixed Ef.	*OR*	*95% CI*	*z*	*Pr (>|z|)*	*OR*	*95% CI*	*z*	*Pr (>|z|)*
lifestyle	0.89	(0.87, 0.91)	-9.9	<2e-16	0.95	(0.92, 0.98)	-3.2	0.001
politics	0.73	(0.71, 0.75)	-24.2	<2e-16	0.61	(0.59, 0.63)	-33.6	<2e-16
(d)	*Log-likelihood:* -174182.1				*Log-likelihood:* -121474.4			
	σ^2 *users:* 2.742		σ^2 *variables:* 2.496		σ^2 *users:* 3.063		σ^2 *variables:* 3.416	
Fixed Ef.	*OR*	*95% CI*	*z*	*Pr (>|z|)*	*OR*	*95% CI*	*z*	*Pr (>|z|)*
@init	1.27	(1.24, 1.29)	20.6	<2e-16	0.93	(0.90, 0.95)	-5.04	<5e-07
@intrnl	0.96	(0.92, 1.00)	-1.9	0.052	0.63	(0.60, 0.67)	-15.3	<2e-16
hashtag	0.85	(0.82, 0.89)	-7.6	<3e-14	1.08	(1.04, 1.12)	3.9	<1e-04
lifestyle	0.90	(0.88, 0.92)	-8.7	<2e-16	0.95	(0.91, 0.98)	-3.4	<0.001
politics	0.74	(0.72, 0.76)	-22.9	<2e-16	0.60	(0.58, 0.61)	-34.3	<2e-16

Table 5: Summary of model parameters for the two datasets: (a) random intercepts only, (b) random intercepts + audience effects, (c) random intercepts + topic effects, (d) full model. σ^2 *users* and σ^2 *variables* are variance estimates for the random intercepts. *Fixed Ef.* tables show odds ratios (*OR*) derived from logit coefficients, with roughly estimated confidence intervals (using approximate standard errors), and results of Wald's z-tests.

for IH Users (c.f. Table 5d).

In general, the effect sizes and directions of the newly defined subtopics are similar to those of the broad topics from which they were isolated, and more importantly, changing the topic definitions has no effect on the audience coefficients for either user group. This provides some evidence that our results are not highly sensitive to the precise choice of topics.

7 Conclusion

This study examined how Twitter users shift their use of Scottish variants depending on the topic and audience. We looked at two groups of users with different overall rates of Scottish usage and found that both topic and audience affect usage in both groups. The qualitative effects of topic were similar across the two groups, demonstrating a clear relationship between the topic or genre of discussion and the odds of choosing Scottish variants. However, the sizes and directions of the audience affects are inconsistent across the two groups: for Scottish Geotag Users we found (as in a previous study) that local variants are used more in tweets with initial mentions and less in tweets with hashtags, but for Indyref Hashtag Users we found the opposite. The demographics and usage patterns of these two groups are very different, and one interesting possibility is that they might be using the affordances of mentions and hashtags in different ways and focusing on different aspects of how these affect their potential audience. In any case, our results underscore the need for caution when drawing broad conclusions from studies of social media data, until the results of those studies are shown to hold across a variety of user samples.

Acknowledgments

This work was supported in part by the EPSRC Centre for Doctoral Training in Data Science, funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016427/1) and the University of Edinburgh.

References

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics* 18(2):135–160.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1):1–48.

Allan Bell. 1984. Language style as audience design. *Language in society* 13(02):145–204.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3(Jan):993–1022.

Jacob Eisenstein. 2015. Systematic patterning in phonologically-motivated orthographic variation. *Journal of Sociolinguistics* 19(2):161–188.

Jacob Eisenstein, Amr Ahmed, and Eric P. Xing. 2011. Sparse additive generative models of text. In *Proceedings of the International Conference on Machine Learning*. pages 1041–1048.

Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences* 101(suppl 1):5228–5235.

Susan C Herring and John C Paolillo. 2006. Gender and genre variation in weblogs. *Journal of Sociolinguistics* 10(4):439–459.

Liangjie Hong and Brian D Davison. 2010. Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics*. ACM, pages 80–88.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 752–762.

Fei Huang and Alexander Yates. 2014. Improving word alignment using linguistic code switching data. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Gothenburg, Sweden, pages 1–9.

Yuan Huang, Diansheng Guo, Alice Kasakoff, and Jack Grieve. 2016. Understanding U.S. regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems* 59:244 – 255.

Dong Nguyen, Dolf Trieschnigg, and Leonie Cornips. 2015. Audience and the use of minority languages on twitter. In *Proceedings of the Ninth International Conference on Web and Social Media*. pages 666–669.

Umashanthi Pavalanathan and Jacob Eisenstein. 2015a. Audience-modulated variation in online social media. *American Speech* 90(2):187–213.

Umashanthi Pavalanathan and Jacob Eisenstein. 2015b. Confounds and consequences in geotagged Twitter data. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 2138–2148.

R Core Team. 2013. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

John R Rickford and Faye McNair-Knox. 1994. Addressee-and topic-influenced style shift: A quantitative sociolinguistic study. *Sociolinguistic perspectives on register* pages 235–276.

Philippa Shoemark, Debnil Sur, Luke Shrimpton, Iain Murray, and Sharon Goldwater. 2017. Aye or naw, whit dae ye hink? Scottish independence and linguistic identity on social media. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics, pages 1239–1248.

Luke Sloan and Jeffrey Morgan. 2015. Who tweets with their location? Understanding the relationship between demographic characteristics and the use of geoservices and geotagging on Twitter. *PloS one* 10(11):e0142209.

Wessel Stoop and Antal van den Bosch. 2014. Using idiolects and sociolects to improve word prediction. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Gothenburg, Sweden, pages 318–327.

Yogarshi Vyas, Spandana Gella, Jatin Sharma, Kalika Bali, and Monojit Choudhury. 2014. POS tagging of English-Hindi code-mixed social media content. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. volume 14, pages 974–979.

5.4 Comments on the paper

5.4.1 Why use mixed effects models?

While mixed effects models are now well-established in various fields of linguistics (Baayen et al., 2008; Johnson, 2009; Linck and Cunnings, 2015), they are not so commonly used in computational linguistics and NLP. We chose to use them here because we have multiple observations per user and per lexical variable, and the observations associated with a particular user (or with a particular lexical variable) are not independent of one another.

A standard logistic regression model would not be able to account for the fact that our datapoints are grouped by user and by lexical variable, and that each user and each lexical variable has its own idiosyncratic baseline rate of Scottish variant usage. This problem is intensified by the fact that the number of observations in our dataset is highly imbalanced both across users (Figure 5.1) and across variables (Figure5.2), as are the rates of Scottish variant usage (Figures 5.3 & 5.4). Suppose that a considerable proportion of the political tweets in our dataset were from one particularly active Twitter user who tweets primarily, but not exclusively, about politics; and that this user happened to use distinctively Scottish variants at relatively low rates, compared to other users, *across all topics*. In this scenario, a model which treats all variable instances as independent, with no knowledge of their grouping by user, would over-estimate the inhibitory effect of the political topic on Scottish variant usage.

Pavalanathan and Eisenstein (2015a) addressed this issue by carefully balancing their dataset such that each user contributed the same number of tweets containing a non-standard variant as tweets without a non-standard variant; that is to say, such that each user had the same baseline rate of non-standard usage (50%). This careful downsampling procedure necesitates throwing away a lot of datapoints, and was not feasible for us given that we started out with much smaller corpora (after spam filtering we had 1.75M tweets by SG-Users and 4.7M tweets by IH-Users, vs. Pavalanathan and Eisenstein's 114M tweets.)

Mixed effects models enable us to to fit models using all of the datapoints at our disposal, while preserving information about which user (and which lexical variable) contributed which datapoint. By specifying by-user and by-variable random intercepts, we allow the model to account for idiosyncratic differences in baseline rates of Scottish variant usage across users and variables, and better describe how audience size and topic relate to deviations from these baseline rates.

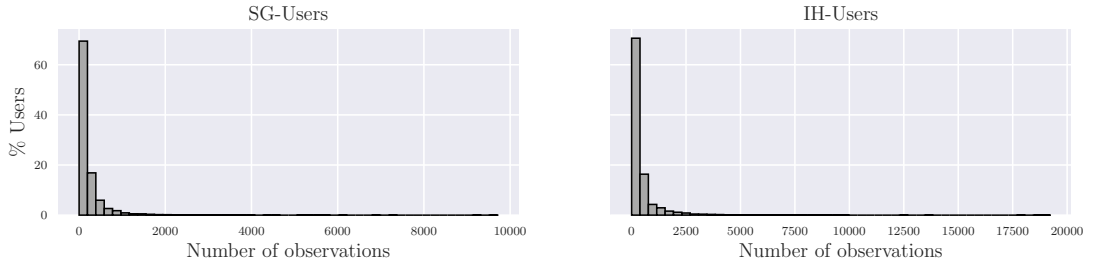


Figure 5.1: Histograms of number of observations per user in SG-Users and IH-Users datasets.

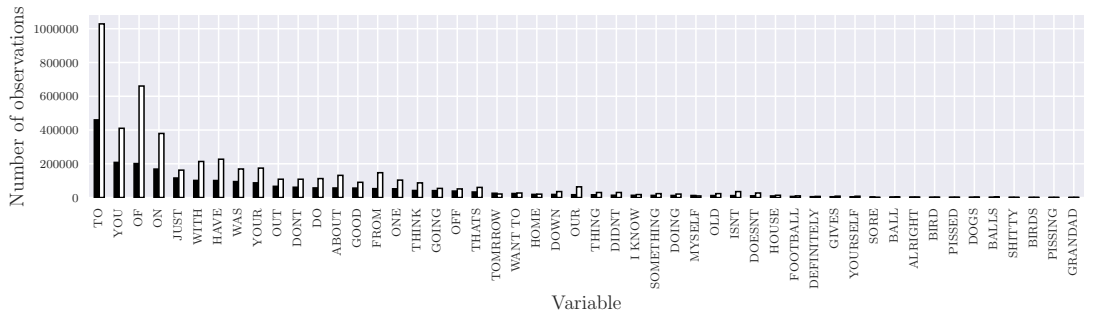


Figure 5.2: Bar chart of number of observations per variable in SG-Users (black bars) and IH-Users (white bars) datasets.

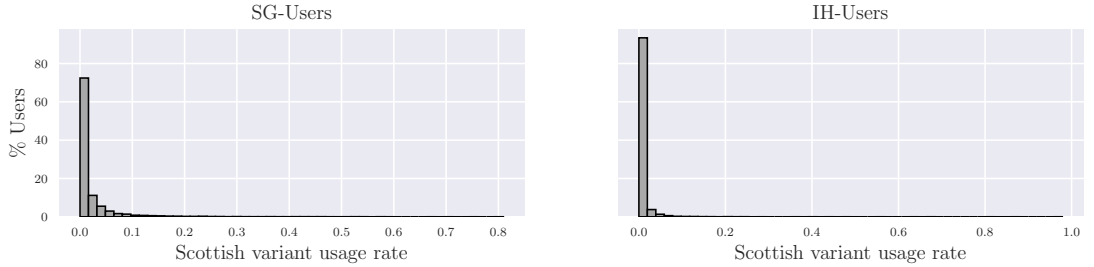


Figure 5.3: Histograms of overall rate of Scottish variant usage per user in SG-Users and IH-Users datasets.

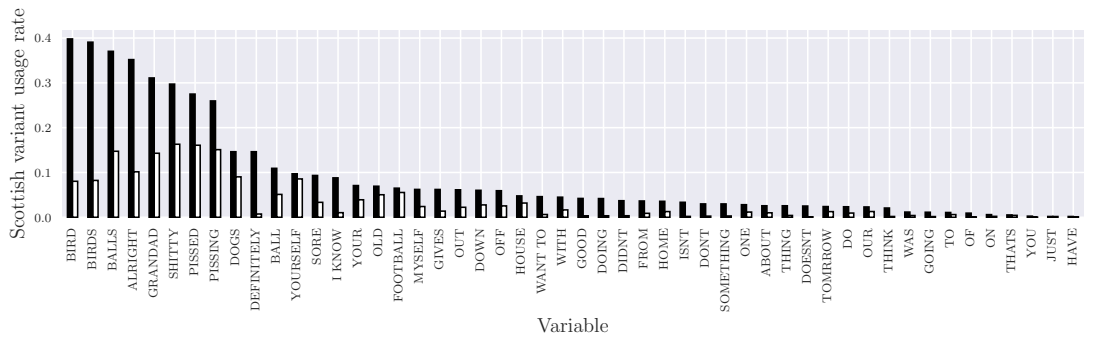


Figure 5.4: Bar chart of overall rate of Scottish variant usage per variable in SG-Users (black bars) and IH-Users (white bars) datasets.

5.4.2 Narrow topic themes

In the paper we considered a model with two additional topic categories, 'sport' and 'indyref', which we split off from the 'lifestyle' and 'politics' categories, respectively. Detailed results for these narrower topic themes were not presented in the paper for space reasons, so we include them here.

	Scottish Geotag Users				**Indyref Hashtag Users**			
	Log-likelihood: -174169.4				*Log-likelihood:* -121447.8			
	σ^2 *users:* 2.748 σ^2 *variables:* 2.496				σ^2 *users:* 3.043 σ^2 *variables:* 3.467			
Fixed Ef.	*OR*	*95% CI*	*z*	*Pr (>|z|)*	*OR*	*95% CI*	*z*	*Pr (>|z|)*
@init	1.27	(1.24, 1.30)	20.8	<2e-16	0.93	(0.91, 0.96)	-4.5	<7e-06
@intrnl	0.96	(0.92, 1.00)	-2.2	0.029	0.63	(0.59, 0.67)	-15.3	<2e-16
hashtag	0.85	(0.82, 0.89)	-7.6	<3e-14	1.06	(1.02, 1.10)	2.8	0.005
lifestyle	0.93	(0.91, 0.95)	-5.6	<2e-08	0.96	(0.92, 0.99)	-2.5	0.014
sport	0.83	(0.80, 0.87)	-9.1	<2e-16	0.90	(0.86, 0.96)	-3.5	<0.001
politics	0.74	(0.72, 0.76)	-19.2	<2e-16	0.54	(0.52, 0.57)	-31.0	<2e-16
indyref	0.75	(0.73, 0.78)	-16.1	<2e-16	0.63	(0.61, 0.65)	-27.4	<2e-16

Table 5.1: Summary of model parameters in full models using finer-grained topic distinctions. σ^2 *users* and σ^2 *variables* are variance estimates for the random intercepts. *Fixed Ef.* tables show odds ratios (*OR*) derived from logit coefficients, with roughly estimated confidence intervals (using approximate standard errors), and results of Wald's z-tests.

In general, the newly defined subtopics behave similarly to the previous topics: for Scottish Geotag Users, the 'indyref' and 'politics' topics have almost identical effect sizes, while for Indyref Hashtag Users, tweeting about the independence referendum has a slightly smaller effect than tweeting about other political themes, though this is still large compared with the effects for 'lifestyle' and 'sport'. For both user groups, the effect of the 'sport' topic is slightly greater than that of the'lifestyle' topic.

5.4.3 By-hashtag random intercepts

In the paper, we found hashtag use in general to be associated with higher odds of choosing distinctively Scottish variants for the group of users sampled on the basis of having used hashtags relating to the Scottish Independence referendum. Conversely, we found hashtags to be associated with lower odds of choosing distinctively Scottish variants for the user group which was sampled on the basis of Scottish geotag use.

To gain some insight into which sorts of hashtags are most associated with higher or lower odds of Scottish variant usage for each user group, we fit new models in which we included by-hashtag random intercepts, in addtion to the by-user and by-variable random intercepts, and then examined the Best Linear Unbiased Predictors of the adjustments to the intercept for individual hashtags. For tweets which contain multiple hashtags, we considered only the first hashtag in the tweet. We coded all tweets in which no hashtags occur with a 'NO_HASHTAG' label. Because the frequencies of individual hashtags have a long-tailed distribution, with the majority of hashtags occuring very infrequently, we replaced all hashtags which occur in fewer than 10 tweets with an 'INFREQUENT_HASHTAG' label.

The 20 hashtag labels with the largest positive adjustments to the intercept (i.e. those most associated with increased odds of Scots usage) and the 20 hashtag labels with the largest negative adjustments to the intercept (i.e. those most associated with decreased odds of Scots usage) are shown in Tables 5.2 and 5.3, for Scottish Geotag Users and Indyref Hashtag Users respectively. Tables 5.2 and 5.3 also give, for each listed hashtag, the number of tweets which are labeled with that hashtag and with 'Scottish' variable instances, and the number which are labeled with that hashtag and with 'Standard' variable instances.

Despite excluding hashtags which occur in fewer than 10 tweets overall, co-occurence counts of individual hashtags with distinctively Scottish variants are vary sparse; it is therefore difficult to read much into the results. Although some associations are clearly spurious (e.g. the three occurences of distinctively 'Scottish' variants with #makeamoviecanadian are instances of 'aboot', which also happens to be distinctively Canadian), it is perhaps worth noting that hashtags relating to the 2014 Commonwealth Games in Glasgow (#glasgow2014, #commonwealthgames, #closingceremony, #bbcglasgow2014, #kylie[1]) are among those with the largest positive BLUPs for both datasets. The Games were intended to promote national and civic pride, and many of

[1] Kylie Minogue performed at the Commonwealth Games closing ceremony

the tweets containing these hashtags do indeed evoke such sentiments, e.g.:

- *#commonwealthgames im reckoning im enjoying the commie games probably a wee bit more than the olympics. lindsay sharp, team tattie tae.*
- *auld lang syne, bagpipes and some fireworks.. canny help but hae a wee greet #closingceremony*
- *glesga... that wis nae bad. #glasgow2014*

Most Positive BLUPs			**Most Negative BLUPs**		
Hashtag	BLUP	'Scot'/'Std'	Hashtag	BLUP	'Scot'/'Std'
#glasgow2014	2.15	16 / 319	#nowplaying	-0.77	0 / 622
#commonwealthgames	1.72	7 / 119	#watp	-0.50	0 / 56
#yeswindaes	1.19	4 / 6	#fact	-0.49	0 / 94
#closingceremony	1.01	4 / 164	#soundhound	-0.47	0 / 88
#stillgame	0.90	3 / 31	#rydercup	-0.47	0 / 208
#gbbo	0.83	6 / 422	#celtic	-0.45	0 / 322
#eh1	0.81	2 / 15	#mwi	-0.43	0 / 19
#bbcqt	0.77	2 / 59	#buzzing	-0.43	0 / 133
#alsicebucketchallenge	0.69	2 / 31	#mufc	-0.40	0 / 235
#sonsofanarchy	0.68	2 / 24	#shazam	-0.38	0 / 51
#lad	0.67	2 / 28	#soundcloud	-0.35	0 / 180
#pt	0.67	2 / 59	#scenes	-0.32	0 / 41
#fuckoff	0.67	2 / 57	#bbuk	-0.32	0 / 115
#rip	0.64	3 / 109	#lfc	-0.32	1 / 262
#brutal	0.64	3 / 19	#yesbecause	-0.32	1 / 453
#fuckoffscotland	0.64	3 / 32	#hmfc	-0.30	0 / 139
#yolo	0.64	2 / 43	#fitfam	-0.29	0 / 11
#previoustweet	0.64	2 / 65	#bettertogether	-0.28	3 / 420
#bigbigdebate	0.61	8 / 404	#shocker	-0.28	0 / 15
NO_HASHTAG	0.60	18,525 / 977,180	#patronisingbtlady	-0.28	0 / 74

Table 5.2: The 20 hashtags with largest positive and negative adjustments to the intercept, for the Scottish Geotag Users dataset.

Most Positive Blups			**Most Negative BLUPs**		
Hashtag	BLUP	'Scot'/'Std'	Hashtag	BLUP	'Scot'/'Std'
#makeamoviecanadian	2.94	3 / 23	#gaza	-1.25	0 / 2888
#patronisingbtladypic	2.76	4 / 49	#hopeoverfear	-1.25	0 / 304
#closingceremony	2.54	34 / 1100	#ferguson	-1.25	0 / 3536
#commonwealthgames	2.46	22 / 692	#no	-1.10	2 / 1311
#goodish	2.34	2 / 10	#masscanvass	-1.02	1 / 97
#glasgow2014	2.27	39 / 1893	#celebritybigbrother	-0.90	0 / 54
#lastnightoftheproms	2.17	3 / 83	#england	-0.87	0 / 251
#celebritieswhopumpdugs	2.17	11 / 0	#lab14	-0.84	1 / 2118
#ladyalba	2.04	3 / 21	#isis	-0.82	0 / 2614
#philosophy	2.00	2 / 100	#mufc	-0.79	0 / 1491
#aye	1.96	5 / 40	#uk	-0.77	0 / 464
#polsco	1.89	3 / 36	#reddit	-0.75	0 / 27
#occupycentralpic	1.88	2 / 19	#skynews	-0.74	0 / 372
#murphy	1.86	2 / 8	#freescotland	-0.71	0 / 28
#gersco	1.80	6 / 148	#fail	-0.71	0 / 140
#bbcglasgow2014	1.72	4 / 123	#indyscot	-0.71	0 / 413
#kylie	1.70	2 / 42	#nowplaying	-0.70	0 / 766
#arrow	1.68	2 / 52	#explainafilmplotbadly	-0.70	1 / 1006
#lies	1.67	2 / 33	#c4news	-0.69	1 / 1865
#georgegalloway	1.58	2 / 56	#45andrising	-0.69	0 / 100

Table 5.3: The 20 hashtags with largest positive and negative adjustments to the intercept, for the Indyref Hashtag Users dataset.

Some of the other hashtags associated with an increase in the log-odds of Scottish variant usage are also related to notions of national or regional identity and pride. For example, #stillgame refers to a popular and critically acclaimed Scottish sitcom set in a fictional Glaswegian housing estate, while #gersco and #polsco refer to Scotland's European Championship Qualifying football matches against Germany and Poland. Moreover, for both datasets, the top-20 lists of hashtags with the largest positive BLUPs include hashtags relating to the Scottish independence referendum: #yeswindaes, #bbcqt, #fuckoffscotland, and #bigbigdebate for Scottish Geotag Users; and #patronisingbtladypic, #ladyalba, #aye, #murphy, #lies, and #georgegalloway for Indyref Hashtag Users.

On the other hand, for both datasets the top 20 hashtags associated with *decreased* odds of Scottish variant usage also include hashtags relating to Scottish independence (both pro and anti). So in the end, there is no clear distinction for either group in the kinds of hashtags which are most and least associated with distinctively Scottish variant use.

Somewhat disconcertingly, one of the top-10 hashtags most associated with decreased odds of Scottish variant usage in the Scottish Geotag Users dataset is actually an example of distinctively Scottish lexis itself (#mwi). We manually inspected a sample of tweets that contain those hashtags which are most negatively associated with Scots variant usage, and observed that many such tweets actually do contain distinctively Scottish vocabulary; just not any of the variants on our list. For example, of the fifteen tweets in the Scottish Geotag Users dataset which contain the negatively associated hashtag #shocker, about half of them do contain distinctively Scottish vocabulary, e.g:

- *i canna believe there were birds in the bayview last night #shocker*
- *think ma maw on 3 blues could of ran quicker than i did tonight #shocker onnit then*
- *just tried to get into some randoms car thinking it was mines and the alarms went off . #shocker #wheresmacar*

This does not invalidate the results in the paper, since our unit of observation is the lexical alternation variable instance, not the tweet. Failing to include some distinctively Scottish variants in our set of lexical alternation variables does not result in the mislabeling of any data-points; it merely results in the analysis being based on a smaller sample of data-points than it could have been. Moreover, the fixed effect of using a hashtag in the original models is estimated on the basis of many more data-points than the BLUPs for each of the individual hashtags are here. Of course, ideally we would like to base our analysis on as many relevant lexical alternations as possible, but as we discussed in §2.3, identifying these is not straightforward and is a laborious and time-consuming process. This motivated us to develop a method to facilitate the identification of lexical alternation variables, which we present in Chapter 6.

5.4.4 Potential explanations for inconsistent findings

Our mixed-effects models indicate that among SG-Users (i.e. those who were sampled on the basis of Scottish geotag use), distinctively Scottish variant usage is positively associated with tweet-initial mentions, while it is negatively associated with hashtags, and there is no significant distance between the odds of choosing distinctitively Scottish variants in broadcast tweets and in tweets with internal mentions. Conversely, among IH-Users (i.e. those who were sampled on the basis of having used hashtags related to the Scottish independence referendum), distinctively Scottish variants have a slight negative association with tweet-initial mentions, a slight positive association with hashtags, and a relatively strong negative association with tweet-internal mentions. In the Conclusion of the paper, we suggested that these two groups of users might be using the affordances of mentions and hashtags in different ways and focusing on different aspects of how these affect their potential audience. In this section I will expand a little upon these hypotheses and how they could explain our findings.

Firstly, we have already shown in Figures 1 and 2 within the paper that there are considerable differences across the datasets in the relative frequency distributions of audience markers and of topics. Regarding audience markers, IH-Users not only posted tweets containing hashtags relatively more frequently than SG-Users did, but also tweets containing mentions, since the proportions of tweets in which the only audience markers were initial or internal mentions are similar across the two user groups, but IH-Users posted relatively more tweets with multiple audience markers (i.e. tweets with both initial and internal mentions, or tweets with both hashtags and mentions). In fact, tweets with multiple audience markers were the most common kind of tweet posted by IH-Users in autumn 2014, while for SG-Users it was broadcasts, i.e. tweets without any hashtags or mentions. Thus we can characterise IH-Users as generally making greater use of the affordances Twitter provides to target posts towards particular groups or individuals.

Regarding topics, close to 45% of SG-Users' tweets were labeled as 'Chatter', which more or less correspond to 'phatic posts' (Radovanovic and Ragnedda, 2012), i.e. posts which have the primary purpose of fostering, maintaining, and reinforcing relationships, as opposed to sharing information or ideas. For IH-Users, the most frequent topics were the Scottish independence referendum followed by other political topics, with only around 20% of tweets being categorized as 'Chatter'.

5.4.4.1 Explaining inter-group differences in effects of mentions

A potential explanation for the discrepancies in the effects of initial and internal mentions is that our two groups of users differ not only in the relative frequencies with which they use mentions, but also in the sorts of accounts they tend to mention, and in their motivations for doing so.

Honeycutt and Herring (2009) analysed functions of mentions in a randomly sampled corpus of English-language tweets, and while the most common function of mentions in their dataset was to address other users in order to engage in conversation with them, they also observed mentions being used to refer to other users (i.e. to talk about them rather than to them). boyd et al. (2010) point out that in addition to addressivity and reference, mentions can also have an attention-seeking function, in that they sometimes appear to be specifically intended to draw the mentioned user's attention to the tweet (since users receive notifications about tweets in which they are mentioned). We therefore propose the following hypotheses:

H1 SG-Users primarily use tweet-initital mentions to **engage in conversation with the mentionee**, and their mentionees tend to be individuals in their local peer group.

H2 SG-Users primarily use tweet-internal mentions to **refer to the mentionee**.

H3 IH-Users are more likely to use either kind of mention to **draw the mentionee's attention to the content of the post**, and their mentionees are more likely to be individuals or organisations outwith their own follower group (who would thus have been unlikely to see the tweet had they not been mentioned in it).

Should these hypotheses turn out to be true, then our results could be interpreted as entirely coherent with Pavalanathan and Eisenstein (2015a) findings (based on Twitter users who use geotags in the USA) that socially marked terms are more likely to be used in conversational messages aimed at narrow, local audiences. According to the abobe hypotheses, the only category corresponding to conversational messages aimed at narrow, local audiences would be SG-Users' tweets with initial mentions, which are indeed associated with increased odds of distinctively Scottish variant use.

Since tweet-internal mentions do not affect whose Home Timelines a tweet is pushed to (see §3.1), the audience invoked by a tweet with an internal mentions is equivalent to that of a broadcast tweet (i.e. the author's followers)—except that the

mentionee receives a notification. If SG-Users use tweet-internal mentions primarily with the intention of directing their followers' attention *towards* the mentioned account as opposed to receiving attention *from* the mentioned account, then it makes sense that they would conceive of the audience in the same way they do for a broadcast tweet, and thus make similar lexical choices to those they would make in broadcast tweets.

Finally, if IH-Users are less inclined to use mentions for conversation or reference, but moreso to spread information and ideas to influential accounts *beyond* their immediate social network, then it makes sense that they might want to maximise the accessibility of such posts, and thus be more likely to inhibit their use of socially marked or regionally specific variants.

5.4.4.2 Explaining inter-group differences in effects of hashtags

Although we were not able to discern any clear systematic differences across the two user groups in the kinds of hashtags that are most associated with increased or decreased odds of Scottish variant usage (see §5.4.3), it could still be the case that the two groups differ systematically with respect to the frequencies with which they use different kinds of hashtags, and with respect to their motivations for using hashtags.

In a series of six empirical studies, Rauschnabel et al. (2019) systematically assessed why and how people use hashtags on social media, and uncovered ten distinct motivations which appear to drive different patterns of hashtagging behavior. Among the ten identified motivations are one the authors call TRENDGAGING (a portmanteau of 'trends' and 'engaging'), which is defined as the desire to engage in and be associated with popular conversations and trendy topics, and is suggested to be driven by the positive effects on an individual's self-esteem that are brought about by the feeling of belonging to an attractive and popular social group. They distinguish trendgaging from BONDING, which they define as the motivation to show that one belongs to an in-group, stemming from the universal human need for a more intimate kind of belonging than can be derived from interactions with new acquaintances. Other identified motivations include ENDORSING: using hashtags for the prosocial purpose of promoting people, brands, events, or topics that one identifies with or finds interesting; and REACHING: using hashtags to share one's opinion about, raise awareness of, or engage in debates on important topics, by addressing specific communities of users who are interested in the relevant topic.

Rauschnabel et al. (2019) relate the ten motivations to five hashtagging 'styles' derived from a factor analysis of different kinds of hashtags people use. Their re-

sults suggest that the trendgaging and endorsing motivations are most associated with a 'modern' hashtagging style characterised by hashtags that are widely popular across social media; while the reaching and bonding motivations are both strongly associated with a 'related' hashtagging style, characterised by hashtags that are connected with specific interest groups or communities of practice. Based on their analyses, we propose the following hypotheses:

H4 SG-Users tend to use hashtags for **trendgaging and endorsing**, so they use more hashtags that are widely popular across social media.

H5 IH-Users tend to use hashtags for **reaching and bonding**, so they use more niche hashtags to engage with specific communities.

Should these hypotheses turn out to be true, the inhibitory effect of hashtags on SG-Users' rates of Scottish variant usage could be explained in terms of the mass audiences invoked by their relatively frequent use of widely popular hashtags; while the slight positive association between hashtags and Scottish variant use for IH-Users could be explained in terms of their bonding motivation, as well as the idea that while the audience of tweets with niche hashtags is not *restricted* since they are still pushed to the feeds of all of the author's followers, the use of niche hashtags can still serve as an indicator of the niche audience it is *intended* for, such that the author may feel licensed to use lexical choices that resonate specifically with that group, even if they may be unfamiliar to followers from outwith the group.

While it might seem counter-intuitive that IH-Users would use mentions to solicit the attention of entities beyond their in-group and hashtags to strengthen bonds within it, some additional grounding for these hypotheses comes from a systematic literature review of Twitter use during election campaigns by Jungherr (2016), in which he notes that Bruns and Highfield (2013) and Conover et al. (2012, 2011) have found that politically vocal Twitter users frequently reach out across party lines when using mentions, while findings by Bode et al. (2015), Hanna et al. (2011, 2013), and Lietz et al. (2014) indicate that politically partisan users tend to create within-group communication spaces through their use of hashtags.

5.4.4.3 Reconciling the findings from this chapter with those from Chapter 4

In Chapter 4, our analyses were based on the IT dataset (see §3.4), a smaller collection of tweets authored by a larger set of Indyref Hashtag Users. We found lower average Scottish usage rates in tweets containing Indyref hashtags than in the rest of the IT

dataset, and originally attributed this finding to a posited audience-broadening function of hashtags (as opposed to the topic being Indyref), since there was no significant difference in Scottish usage rates between tweets containing Indyref hashtags and tweets that contained other hashtags. However, the mixed-effects models we have presented here indicate that when we carefully account for effects of topic, for IH-Users hashtags are actually associated with *higher* odds of choosing distinctively Scottish variants.

Nevertheless, the results from this chapter and the previous one are not incoherent. As we have acknowledged throughout, the studies in Chaper 4 did not enable us to conclusively distinguish effects of audience and topic. Since our mixed-effects models have revealed that the negative effect of tweeting about politics is considerably larger than the positive effect of using hashtags, we can now re-interpret the lower average rate of distinctively Scottish variant use in Indyref hashtag tweets vs all other tweets in the IT dataset as being driven by the inhibitory effect of the Indyref topic after all. To explain the reduction in the difference in rates of Scottish variant use between Indyref hashtag tweets and control tweets when we restricted the control set to other tweets which contain hashtags, we can consider the different topic distributions across all tweets vs. tweets with hashtags. In the IH-Users dataset (in which we have inferred the topic of every tweet), 56% of all tweets are political, but this proportion rises to 72% in the subset of tweets which contain hashtags. Assuming these proportions are similar in the IT dataset, the diminished difference in Scottish variant usage rates can thus be explained by a diminished difference in the proportion of tweets which are political.

5.5 Future work

A limitation of our paper is its coarse-grained operationalisation of audience size on the basis of mentions and hashtags. While mentions and hashtags do enable users to manipulate the likely composition of the audience, the assumption that mentions invoke narrower audiences than hashtags is rather reductive: the size of the prospective audience will naturally depend on how widely followed the mentioned user or included hashtag is. Hence, to better understand the implications of the difference we observed in audience effects across our two user groups, future studies could incorporate more nuanced characterisations of audience.

One approach that could be taken is to estimate the popularity of relevant hashtags based on the frequency with which they occur in a sample of tweets, and of mentioned

users based on their follower counts, in order to make more fine-grained distinctions as a better proxy for audience size. Future studies could also operationalise other aspects of the composition of the prospective audience. For example, following Nguyen et al. (2015) they could consider the prospective audience's own Scottish vocabulary usage rates, estimated on the basis of Scottish vocabulary usage rates in other tweets which contain the relevant hashtag, or other users who follow (or have also mentioned) the relevant mentioned user. In case there is little variation in these rates, future studies could alternatively consider the geographic dispersion of hashtags, and/or followers or fellow-mentioners of mentioned users. Pavalanathan and Eisenstein (2015a) found that users in the USA were more likely to use non-standard terms in tweets containing mentions when the mentioned user was geographically close to the author. Because this finding held true not only for regionally-specific non-standard terms, but also for non-standard terms which were widely used in tweets across the USA, they speculated that the underlying factor behind this geographical proximity effect may in fact be social tie strength. Another interesting direction for future work would therefore be to consider a more direct measure of social tie strength between tweet authors and mentioned users, perhaps based on a graph of follower relationships or mentions (though these are particularly tricky to sample; see González-Bailón et al. 2014), or based on the Smallest Common Hashtag measure introduced by Romero et al. (2013). Romero et al. found that the frequency of the least popular hashtag that a given pair of users have both used is highly predictive of the probability of those users having followed or mentioned each other, thus we propose it could potentially function as a proxy for social tie strength.

Lastly (though by no means least importantly), our quantitative analysis could also be complemented by qualitative content analyses looking in detail at particular users and/or hashtags, as well as surveys or interviews in order to test the hypotheses we put forward in §5.4.4 about how differences in users' *motivations* for using hashtags and mentions could explain our present findings.

Chapter 6

Lexical variable discovery

6.1 Introduction

In Chapters 4 and 5, we analysed intra- and inter- speaker variation in usage rates of distinctively Scottish terms. We used a data-driven method to identify distinctively Scottish terms, but manually paired them with Standard English equivalents. This manual step was time consuming and required a high degree of familiarity with both Standard English and Scots/Scottish English. Hence it would be useful if we could also use data-driven methods to facilitate the identification of *pairs* of variants from two specified language varieties, which have the same referential meaning and can occur in the same syntactic contexts. As well as speeding up the process of curating lists of lexical variables, such a method could suggest variables that a researcher might not have otherwise considered.

The task we aim to solve is similar to bilingual lexicon induction. However, state of the art methods for bilingual lexicon induction (e.g. Conneau et al. 2017; Artetxe et al. 2018) use separate monolingual corpora for each language. We are interested in analysing variation in the use of minority language varieties, dialects, registers or sociolects, for which mono-lectal corpora are often difficult to construct. For closely related language variety pairs like English and Scots, which share a lot of vocabulary and are frequently code-mixed within single utterances, it is often impossible to definitively say whether a document belongs to one language variety or the other. Therefore, unlike in the typical setting for bilingual lexicon induction, we want to be able to induce lexical variables from a *single* corpus which contains a mixture of language varieties. Our task involves not only mapping terms in one variety onto terms in the other, but also working out which terms belong to which variety.

In this chapter we present a simple method for lexical variable induction from code-mixed text, which provides researchers with a ranked list of candidate variant pairs that they only have to accept or reject. We show that with as few as five manually curated seed pairs, our proposed method can efficiently identify large numbers of additional variables.

6.2 Author contributions

The paper is co-authored by me, James Kirby, and Sharon Goldwater. As the leading author, I conceived and developed the method, wrote the code, performed the experiments, and drafted the paper. James Kirby and Sharon Goldwater supervised the project, offered suggestions, and helped to revise the final manuscript.

6.3 The paper

The paper was accepted for publication at the 4th Workshop on Noisy User-generated Text at EMNLP 2019 in Copenhagen, where it was featured as an oral presentation and won a Best Paper Award. The publication reference is as follows:

> Shoemark, P., Kirby, J., & Goldwater, S. (2018, November). Inducing a lexicon of sociolinguistic variables from code-mixed text. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text* (pp. 1-6).

Inducing a lexicon of sociolinguistic variables from code-mixed text

Philippa Shoemark[*] p.j.shoemark@ed.ac.uk
James Kirby[†] j.kirby@ed.ac.uk
Sharon Goldwater[*] sgwater@inf.ed.ac.uk

[*]School of Informatics
University of Edinburgh

[†]Dept. of Linguistics and English Language
University of Edinburgh

Abstract

Sociolinguistics is often concerned with how variants of a linguistic item (e.g., *nothing* vs. *nothin'*) are used by different groups or in different situations. We introduce the task of inducing lexical variables from code-mixed text: that is, identifying equivalence pairs such as (*football*, *fitba*) along with their linguistic code (*football*→British, *fitba*→Scottish). We adapt a framework for identifying gender-biased word pairs to this new task, and present results on three different pairs of English dialects, using tweets as the code-mixed text. Our system achieves precision of over 70% for two of these three datasets, and produces useful results even without extensive parameter tuning. Our success in adapting this framework from gender to language variety suggests that it could be used to discover other types of analogous pairs as well.

1 Introduction

Large social media corpora are increasingly used to study variation in informal written language (Schnoebelen, 2012; Bamman et al., 2014; Nguyen et al., 2015; Huang et al., 2016). An outstanding methodological challenge in this area is the bottom-up discovery of sociolinguistic *variables*: linguistic items with identifiable variants that are correlated with social or contextual traits such as class, register, or dialect. For example, the choice of the term *rabbit* versus *bunny* might correlate with audience or style, while *fitba* is a characteristically Scottish variant of the more general British *football*.

To date, most large-scale social media studies have studied the usage of individual variant forms (Eisenstein, 2015; Pavalanathan and Eisenstein, 2015). Studying how a variable *alternates* between its variants controls better for 'Topic Bias' (Jørgensen et al., 2015), but identifying the relevant variables/variants may not be straightforward. For example, Shoemark et al. (2017b) used a data-driven method to identify distinctively Scottish terms, and then manually paired them with Standard English equivalents, a labour intensive process that requires good familiarity with both language varieties. Our aim is to facilitate the process of curating sociolinguistic variables by providing researchers with a ranked list of candidate variant pairs, which they only have to accept or reject.

This task, which we term *lexical variable induction*, can be viewed as a type of bilingual lexicon induction (Haghighi et al., 2008; Zhang et al., 2017). However, while most work in that area assumes that monolingual corpora are available and labeled according to which language they belong to, in our setting there is a single corpus containing code-mixed text, and we must identify both translation equivalents (*football, fitba*) as well as linguistic code (*football*→British, *fitba*→Scottish). To illustrate, here are some excerpts of tweets from the Scottish dataset analysed by Shoemark et al., with Standard English glosses in italics:[1]

1. need to come hame fae the football
   need to come home from the football
2. miss the fitba
   miss the football
3. awwww man a wanty go tae the fitbaw
   awwww man I want to go to the football

The lexical variable induction task is challenging: we cannot simply classify documents containing *fitba* as Scottish, since the *football* variant may also occur in otherwise distinctively Scottish texts, as in (1). Moreover, if we start by knowing only a few variables, we would like a way to learn what other likely variables might be. Had we not known

[1]Note that it is hard to definitively say whether tweets such as these are mixing English and Scots codes, or whether they are composed entirely in a single Scottish code, which happens to share a lot of vocabulary with Standard English.

the (*football, fitba*) variable, we might not detect that (2) was distinctively Scottish. Our proposed system can make identifying variants quicker and also suggest variant pairs a researcher might not have otherwise considered, such as (*football, fitbaw*) which could be learned from tweets like (3).

Our task can also be viewed as the converse of the one addressed by Donoso and Sanchez (2017), who proposed a method to identify geographical regions associated with different linguistic codes, using pre-defined lexical variables. Also complementary is the work of Kulkarni et al. (2016), who identified words which have the same form but different semantics across different linguistic codes; here, we seek to identify words which have the same semantics but different forms.

We frame our task as a ranking problem, aiming to generate a list where the best-ranked pairs consist of words that belong to different linguistic codes, but are otherwise semantically and syntactically equivalent. Our approach is inspired by the work of Schmidt (2015) and Bolukbasi et al. (2016), who sought to identify pairs of words that exhibit gender bias in their distributional statistics, but are otherwise semantically equivalent. Their methods differ in the details but use a similar framework: they start with one or more seed pairs such as {(*he, she*), (*man, woman*)} and use these to extract a 'gender' component of the embedding space, which is then used to find and rank additional pairs.

Here, we replace the gendered seed pairs with pairs of sociolinguistic variants corresponding to the same variable, such as {(*from, fae*), (*football, fitba*)}. In experiments on three different datasets of mixed English dialects, we demonstrate useful results over a range of hyperparameter settings, with precision@100 of over 70% in some cases using as few as five seed pairs. These results indicate that the embedding space contains structured information not only about gendered usage, but also about other social aspects of language, and that this information can potentially be used as part of a sociolinguistic researcher's toolbox.

2 Methods

Our method consists of the following steps.[2]

Train word embeddings We used the Skip-gram algorithm with negative sampling (Mikolov et al., 2013) on a large corpus of code-mixed text to obtain a unit-length embedding $\boldsymbol{w}$ for each word in the input vocabulary V.[3]

Extract 'linguistic code' component Using seed pairs $S = \{(\boldsymbol{x_i}, \boldsymbol{y_i}), i = 1 \ldots n\}$, we compute a vector $\boldsymbol{c}$ representing the component of the embedding space that aligns with the linguistic code dimension. Both Schmidt and Bolukbasi et al. were able to identify gender-biased word pairs using only a single seed pair, defining the 'gender' component as $\boldsymbol{c} = \boldsymbol{w_{she}} - \boldsymbol{w_{he}}$. However, there is no clear prototypical pair for dialect relationships, so we average our pairs, defining $\boldsymbol{c} = \frac{1}{n}\sum_i \boldsymbol{x_i} - \frac{1}{n}\sum_i \boldsymbol{y_i}$.[4] We experiment with the number of required seed pairs in §5.

Threshold candidate pairs From the set of all word pairs in $V \times V$, we generate a set of candidate output pairs. We follow Bolukbasi et al. (2016) and consider only pairs whose embeddings meet a minimum cosine similarity threshold δ. We set δ automatically using our seed pairs: for each seed pair $(\boldsymbol{x_i}, \boldsymbol{y_i})$ we compute $\cos(\boldsymbol{x_i}, \boldsymbol{y_i})$ and set δ equal to the lower quartile of the resulting set of cosine similarities.

Rank candidate pairs Next we use $\boldsymbol{c}$ to rank the remaining candidate pairs such that the top-ranked pairs are the most indicative of distinct linguistic codes, but are otherwise semantically equivalent. We follow Bolukbasi et al. (2016),[5] setting $score(\boldsymbol{w_i}, \boldsymbol{w_j}) = \cos(\boldsymbol{c}, \boldsymbol{w_i} - \boldsymbol{w_j})$.

Filter top-ranked pairs High dimensional embedding spaces often contain 'hub' vectors, which are the nearest neighbours of a disproportionate number of other vectors (Radovanović et al., 2010). In preliminary experiments we found that many of our top-ranked candidate pairs included such 'hubs', whose high cosine similarity with the word vectors they were paired with did not reflect genuine semantic similarity. We therefore discard all pairs containing words that appear in more than m of the top-n ranked pairs.[6]

[2] Code is available at `github.com/pjshoemark/lexvarinduction`.

[3] In preliminary experiments we also tried CBOW and FastText, but obtained better output with Skip-gram.

[4] Bolukbasi et al. (2016) introduced another method to combine multiple seed pairs, using Principal Component Analysis. We compared it and a variant to our very simple difference of means method, and found little difference in their efficacy. Details can be found in the Supplement. All results reported in the main paper use the method defined above.

[5] See Supplement for comparison with an alternative scoring method devised by Schmidt (2015).

[6] The choice of $m \in \{5, 10, 20\}$ and $n \in \{5k, 10k, 20k\}$ made little difference, although we did choose the best pa-

3 Datasets

We test our methods on three pairs of language varieties: British English vs Scots/Scottish English; British English vs General American English; and General American English vs African American Vernacular English (AAVE). Within each data set, individual tweets may contain words from one or both codes of interest, and the *only* words with a known linguistic code (or which are known to have a corresponding word in the other code) are those in the seed pairs.

BrEng/Scottish For our first test case, we combined the two datasets collected by Shoemark et al. (2017a), consisting of complete tweet histories from Aug-Oct 2014 by users who had posted at least one tweet in the preceding year geotagged to a location in Scotland, or that contained a hashtag relating to the 2014 Scottish Independence referendum. The corpus contains 9.4M tweets.

For seeds, we used the 64 pairs curated by Shoemark et al. (2017b). Half are discourse markers or open-class words (*dogs, dugs*), (*gives, gees*) and half are closed-class words (*have, hae*), (*one, yin*). The full list is included in the Supplement.

BrEng/GenAm For our next test case we recreated the entire process of collecting data and seed variables from scratch. We extracted 8.3M tweets geotagged to locations in the USA from a three-year archive of the public 1% sample of Twitter (1 Jul 2013–30 Jun 2016). All tweets were classified as English by langid.py (Lui and Baldwin, 2012), none are retweets, none contain URLs or embedded media, and none are by users with more than 1000 friends or followers. We combined this data with a similarly constructed corpus of 1.7M tweets geotagged to the UK and posted between 1 Sep 2013 and 30 Sep 2014.

To create the seed pairs, we followed Shoemark et al. (2017b) and used the Sparse Additive Generative Model of Text (SAGE) (Eisenstein et al., 2011) to identify the terms that were most distinctive to UK or US tweets. However, most of these terms turned out to represent specific dialects *within* each country, rather than the standard BrEng or GenAm dialects (we discuss this issue further below). We therefore manually searched through the UK terms to identify those that are standard BrEng and differ from GenAm by spelling only, and paired each one with its GenAm spelling variant, e.g. (*color, colour*), (*apologize, apologise*), (*pajamas, pyjamas*). This process involved looking through thousands of words to identify only 27 pairs (listed in the Supplement), which is a strong motivator for our proposed method to more efficiently increase the number of pairs.

GenAm/AAVE While creating the previous dataset, we noticed that many of the terms identified by SAGE as distinctively American were actually from AAVE. To create our GenAm/AAVE seed pairs, we manually cross-referenced the most distinctively 'American' terms with the AAVE phonological processes described by Rickford (1999). We then selected terms that reflected these processes, paired with their GenAm equivalents, e.g. (*about, bou*), (*brother, brudda*). The full list of 19 open-class and 20 closed-class seed pairs is included in the Supplement.

4 Evaluation Procedure

We evaluate our systems using Precision@K, the percentage of the top K ranked word pairs judged to be valid sociolinguistic variables. We discard any seed pairs from the output before computing precision. Since we have no gold standard translation dictionaries for our domains of interest, each of the top-K pairs was manually classified as either valid or invalid by the first author.

For a pair to be judged as valid, (a) each member must be strongly associated with one or the other language variety, (b) they must be referentially, functionally, and syntactically equivalent, and (c) the two words must be ordered correctly according to their language varieties, e.g. if the seeds were (BrEng, GenAm) pairs, then the BrEng words should also come first in the top-K output pairs.

Evaluation judgements were based on the author's knowledge of the language varieties in question; for unfamiliar terms, tweets containing the terms were sampled and manually inspected, and cross-referenced with urbandictionary.com and/or existing sociolinguistic literature.

Our strict criteria exclude pairs like *(dogs, dug)* which differ in their inflection, or *(quid, dollar)* whose referents are distinct but are equivalent across cultures. In many cases it was difficult to judge whether or not a pair should be accepted, such as when not all senses of the words were interchangable, e.g. BrEng/GenAm *(folk, folks)*

rameters for each language pair: $m = 20$, $n = 20k$ for BrEng/Scottish; $m = 5$, $n = 5k$ for GenAm/AAVE; and $m = 10$, $n = 5k$ for BrEng/GenAm.

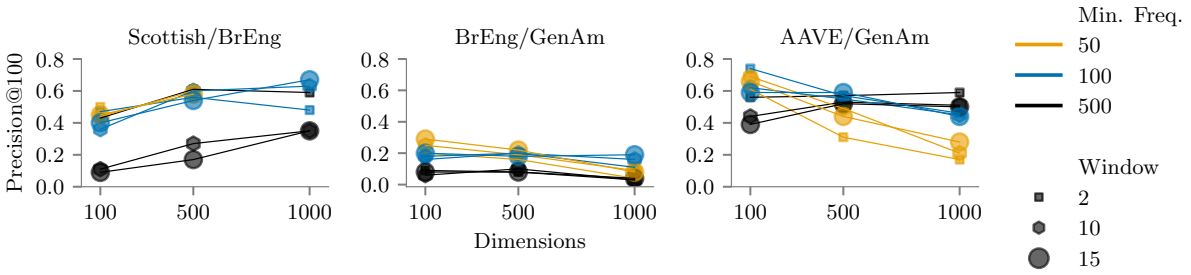


Figure 1: Precision@100 for various Skip-gram hyperparameter settings.

works for the 'people' sense of *folk*, but not the adjectival sense: *(folk music, *folks music)*. The BrEng/GenAm dataset also yielded many pairs of words that exhibit different frequencies of usage in the two countries, but where both words are part of both dialects, such as *(massive, huge)*, *(vile, disgusting)*, and *(horrendous, awful)*. We generally marked these as incorrect, although the line between these pairs and clear-cut lexical alternations is fuzzy. For some applications, it may be desirable to retrieve pairs like these, in which case the precision scores we report here are very conservative.

5 Results and Discussion

We started by exploring how the output precision is affected by the hyperparameters of the word embedding model: the number of embedding dimensions, size of the context window, and minimum frequency below which words are discarded. Results (Figure 1) show that the context window size does not make much difference and that the best scores for each language use a minimum frequency threshold of 50-100. The main variability seems to be in the optimal number of dimensions, which is much higher for the BrEng/Scottish data set than for GenAm/AAVE. Although the precision varies considerably, it is over 40% for most settings, which means a researcher would need to manually check only a bit over twice as many pairs as needed for a study, rather than sifting through a much larger list of individual words and trying to come up with the correct pairs by hand. Results for BrEng/GenAm are worse than for the other two datasets, for reasons which become clear when we look at the output.

Table 1 shows the top 10 generated pairs for each pair of language varieties, using the best hyperparameters for each of the datasets. The top 100 are given in the Supplement. According to our strict evaluation criteria, many of the output pairs for the BrEng/GenAm dataset were scored as incorrect. However, most of them are actually sensible, and examples of the kinds of grey areas and cultural analogies (e.g., (*amsterdam, vegas*), (*bbc, cnn*)) that we discussed in §4. These types of pairs likely predominate because BrEng and GenAm are both standardized dialects with very little difference at the lexical level, so cultural analogies and frequency effects are the most salient differences.

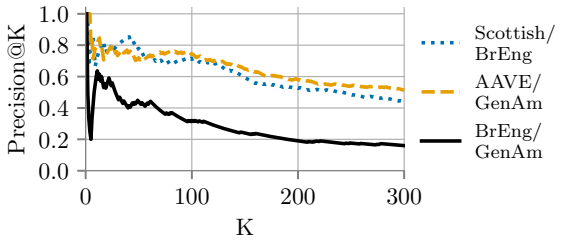


Figure 2: Precision@K from K=1 to 300 for each language variety pair.

BrEng / Scottish	BrEng / GenAm	GenAm / AAVE
now / noo	mums / moms	the / tha
what / whit	*dunno / idk*	with / wit
wasnt / wis	*yeh / yea*	getting / gettin
cant / canny	*shouting / yelling*	just / jus
would / wid	*quid / dollars*	*and / nd*
doesnt / disny	learnt / learned	making / makin
cant / cannae	favour / favor	*when / wen*
going / gonny	sofa / couch	looking / lookin
want / wanty	advert / commercial	something / somethin
anyone / embdy	adverts / commercials	going / goin

Table 1: Top 10 ranked variables for each language pair (invalid variables in italics).

To show how many pairs can be identified effectively, Figure 2 plots Precision@K as a function of $K \in \{1..300\}$. For BrEng/Scottish and GenAm/AAVE, more than 70% of the top-100 ranked word pairs are valid. Precision drops off fairly slowly, and is still at roughly 50% for these two datasets even when returning 300 pairs.

To assess the contribution of the 'linguistic code' component, we compared the performance of our system with a naïve baseline which does not use the 'linguistic code' vector $\boldsymbol{c}$ at all. Since translation equivalents such as *fitba* and *football* are likely

	Baseline	Our Method
BrEng / Scottish	0.00	0.71
BrEng / GenAm	0.04	0.32
GenAm / AAVE	0.08	0.74

Table 2: Precision@100 for our method and the baseline for each language pair.

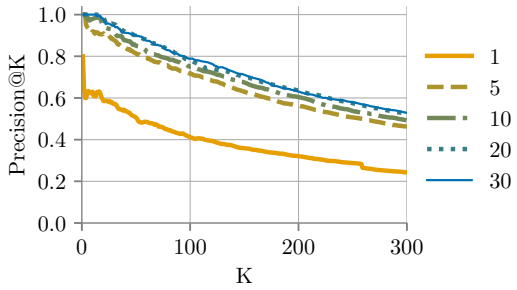
to be very close to one another in the embedding space, it is worth checking whether they can be identified on that basis alone. The baseline ranks all unordered pairs of words in the vocabulary just by their cosine similarity, $cos(\boldsymbol{w_i}, \boldsymbol{w_j})$. Since this baseline gives us no indication of which word belongs to which language variety, we evaluated it only on its ability to correctly identify translation equivalents (i.e. using criteria (a) and (b), see §4), and gave it a free pass on assigning the variants to the correct linguistic codes (criterion (c)). Results are in Table 2. Despite its more lenient evaluation criteria, the baseline performs very poorly. Perhaps if we were starting with a pre-defined set of words from one language variety which were known to have equivalents in the other, then simply returning their nearest neighbours might be sufficient. However, in this more difficult setting where we lack prior knowledge about which words belong to our codes of interest, an additional signal clearly is needed.

Finally, we examined how performance depends on the particular seed pairs we used and the number of seed pairs. Using the BrEng/Scottish and GenAm/AAVE datasets, we subsampled between 1 and 30 seed pairs from our original sets. Over 10 random samples of each size, we found similar average performance using just 5 seed pairs as when using the full original sets (see Figure 3). Performance increased slightly when using only open-class seed pairs: P@100 rose to 0.77 for Scottish/BrEng and 0.75 for GenAm/AAVE (cf. 0.71 and 0.74 using all the original seed pairs). These results indicate our method is robust to the number and quality of seed pairs.

6 Conclusion

Overall, our results demonstrate that sociolinguistic information is systematically encoded in the word embedding space of code-mixed text, and that this implicit structure can be exploited to identify sociolinguistic variables along with their linguistic code. Starting from just a few seed variables, a simple heuristic method is sufficient to identify a large number of additional candidate pairs with precision of 70% or more. Results are somewhat sensitive to different hyperparameter settings but even non-optimal settings produce results that are likely to save time for sociolinguistic researchers. Although we have so far tested our system only on varieties of English[7], we expect it to perform well with other pairs of language varieties which have a lot of vocabulary overlap or are frequently code-mixed *within* sentences or short documents, including code-mixed languages as well as dialects. This framework may also be useful for identifying variation across genres or registers.

(a) Scottish/BrEng



(b) AAVE/GenAm

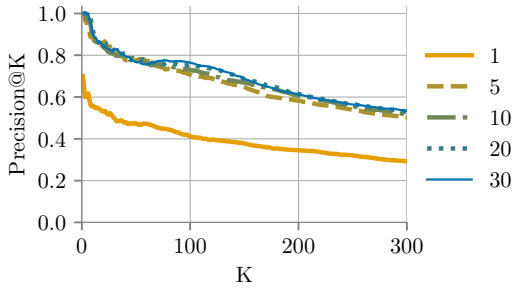


Figure 3: Mean Precision@K curves for different sized samples from the original seed pair lists. Each curve is averaged across 10 random samples of n seed pairs, for $n \in \{1, 5, 10, 20, 30\}$.

Acknowledgments

This work was supported in part by the EPSRC Centre for Doctoral Training in Data Science, funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016427/1) and the University of Edinburgh.

[7]The Scots language, while not a variety of Modern English, developed from a dialect of Old English and in practise is often inextricably mixed with Scottish English.

References

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 4356–4364. Curran Associates Inc.

Gonzalo Donoso and David Sanchez. 2017. Dialectometric analysis of language variation in Twitter. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects*, VarDial'17, pages 16–25. Association for Computational Linguistics.

Jacob Eisenstein. 2015. Systematic patterning in phonologically-motivated orthographic variation. *Journal of Sociolinguistics*, 19(2):161–188.

Jacob Eisenstein, Amr Ahmed, and Eric P. Xing. 2011. Sparse additive generative models of text. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 1041–1048. Omnipress.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies*, ACL'08: HLT, pages 771–779. Association for Computational Linguistics.

Yuan Huang, Diansheng Guo, Alice Kasakoff, and Jack Grieve. 2016. Understanding U.S. regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems*, 59:244 – 255.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2015. Challenges of studying and processing dialects in social media. In *Proceedings of the Workshop on Noisy User-generated Text*, W-NUT'15, pages 9–18. Association for Computational Linguistics.

Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2016. Freshman or fresher? quantifying the geographic variation of language in online social media. In *Proceedings of the Tenth International Conference on Web and Social Media*, ICWSM'16, pages 615–618. AAAI Press.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, ACL'12, pages 25–30. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119. Curran Associates Inc.

Dong Nguyen, Dolf Trieschnigg, and Leonie Cornips. 2015. Audience and the use of minority languages on Twitter. In *Proceedings of the Ninth International Conference on Web and Social Media*, ICWSM'15, pages 666–669. AAAI Press.

Umashanthi Pavalanathan and Jacob Eisenstein. 2015. Audience-modulated variation in online social media. *American Speech*, 90(2):187–213.

Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. 2010. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(Sep):2487–2531.

John R Rickford. 1999. *African American vernacular English: Features, evolution, educational implications*, chapter 1. Blackwell Malden, MA.

Ben Schmidt. 2015. Rejecting the gender binary: a vector-space operation. http://bookworm.benschmidt.org/posts/2015-10-30-rejecting-the-gender-binary.html.

Tyler Schnoebelen. 2012. Do you smile with your nose? stylistic variation in Twitter emoticons. *University of Pennsylvania Working Papers in Linguistics*, 18(2):14.

Philippa Shoemark, James Kirby, and Sharon Goldwater. 2017a. Topic and audience effects on distinctively scottish vocabulary usage in Twitter data. In *Proceedings of the Workshop on Stylistic Variation*, pages 59–68. Association for Computational Linguistics.

Philippa Shoemark, Debnil Sur, Luke Shrimpton, Iain Murray, and Sharon Goldwater. 2017b. Aye or naw, whit dae ye hink? Scottish independence and linguistic identity on social media. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, EACL'17, pages 1239–1248. Association for Computational Linguistics.

Meng Zhang, Haoruo Peng, Yang Liu, Huan-Bo Luan, and Maosong Sun. 2017. Bilingual lexicon induction from non-parallel data with minimal supervision. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, pages 3379–3385. AAAI Press.

Supplementary Information

A Full lists of seed variables used

Tables 1 to 3 list the full sets of seed variables that were used in our experiments.

BrEng	Scottish	BrEng	Scottish
alright	awright	about	aboot
alright	awrite	didnt	didny
alright	awryt	do	dae
ball	baw	doesnt	doesny
balls	baws	dont	deh
bird	burd	dont	dini
birds	burds	dont	dinny
definitely	defos	from	fae
dogs	dugs	have	hae
doing	dain	isnt	isny
down	doon	myself	masel
football	fitba	myself	maself
gives	gees	of	ae
gimme	gees	off	aff
going	gawn	on	oan
good	gid	one	wan
grandad	granda	one	yin
grandfather	granda	our	oor
grandpa	granda	out	oot
home	hame	thats	ats
house	hoose	to	tae
just	jist	too	tae
old	auld	wanna	wanty
pissed	pished	was	wis
pissing	pishing	with	wi
shit	shite	with	wae
shitty	shitey	you	yi
something	suhin	your	yer
sore	sare	your	yir
thing	hing	youre	yer
think	hink	youre	yir
tomorrow	morra	yourself	yersel

Table 1: BrEng/Scottish seed pairs. Left: open-class; Right: closed-class.

BrEng	GenAm	BrEng	GenAm
color	colour	apologize	apologise
coloring	colouring	apologized	apologised
colors	colours	apologizing	apologising
colored	coloured	realize	realise
favorite	favourite	realized	realised
favorites	favourites	realizes	realises
favorited	favourited	realizing	realising
behavior	behaviour	gray	grey
neighbor	neighbour	theater	theatre
neighbors	neighbours	theaters	theatres
humor	humour	tire	tyre
rumor	rumour	tires	tyres
rumors	rumours	math	maths
		pajamas	pyjamas

Table 2: BrEng/GenAm seed pairs (all open-class).

GenAm	AAVE	GenAm	AAVE
brother	brova	about	bou
brother	brudda	about	bout
everybody	erbody	after	afta
everybody	errybody	before	befo
everyday	errday	dont	ont
everyday	erryday	every	erry
everyone	erryone	for	fa
everything	errthang	for	fah
everything	everythang	for	fo
hundred	hunna	that	dat
hundred	hunnid	the	da
hundred	hunnit	there	der
later	lata	there	dere
nothing	nuffin	these	deez
nothing	nuttin	these	dese
partner	partna	they	dey
partner	patna	this	dis
round	roun	to	tah
thing	thang	with	wid
		with	witt

Table 3: GenAm/AAVE seed pairs. Left: open-class; Right: closed-class.

B Comparison of methods

B.1 Alternative methods for extracting 'linguistic code' component

In addition to the very simple method presented in the main paper (which we will refer to here as MEANSDIFF), we tested two additional methods for combining multiple seed pairs to identify a single 'linguistic code' component.

The first is a version of the method Bolukbasi et al. used in their full debiasing algorithm, which we call OFFSETSPCA.[1] We compute the mean $\boldsymbol{m_i}$ of each seed pair $(\boldsymbol{x_i}, \boldsymbol{y_i})$ and the offset vectors $\boldsymbol{m_i} - \boldsymbol{x_i}$ and $\boldsymbol{m_i} - \boldsymbol{y_i}$. We then apply PCA to the resulting collection of offsets, and set $\boldsymbol{c}$ equal to the first principal component. We also consider a simplere variant, INDIVPCA, wherein we define $\boldsymbol{c}$ to be the first principal component of the set of all our individual seed vectors; no pairwise information is used.

B.2 Alternative methods for ranking candidate word pairs

As well as the ranking method that we used throughout the main paper (which we adopted from Bolukbasi et al. (2016) and will refer to here as DIFFSIM), we also tested an alternative ranking method devised by Schmidt (2015). Schmidt's

[1] Bolukbasi et al.'s method is more general, and can in principle be used to extract multi-dimensional subspaces using equivalency seed sets of any cardinality, not just pairs. This was not necessary for our application.

method, which we will call REJECT, first 'rejects' $\boldsymbol{c}$ from each word $\boldsymbol{w_i}$ in a candidate pair, projecting $\boldsymbol{w_i}$ onto the plane orthogonal to $\boldsymbol{c}$, as $\boldsymbol{w_i^{(r)}} = \boldsymbol{w_i} - (\boldsymbol{w_i} \cdot \boldsymbol{c})\boldsymbol{c}$. It then defines $score(\boldsymbol{w_i}, \boldsymbol{w_j})$ as the ratio between the similarity of the rejected vectors and the originals: $\frac{cos(\boldsymbol{w_i^{(r)}}, \boldsymbol{w_j^{(r)}})}{cos(\boldsymbol{w_i}, \boldsymbol{w_j})}$

B.3 Results of comparison

Table 4 compares the different methods outlined above for extracting the 'linguistic code' component and ranking the candidate word pairs. There is little difference in P@100 between OFFSETSPCA, INDIVPCA, and MEANSDIFF except on the most difficult dataset (BrEng/GenAm), where INDIVPCA (the only method that doesn't explicitly use pairwise information) fails entirely. DIFFSIM and REJECT perform similarly except on the AAVE/GenAm dataset, providing some evidence that DIFFSIM is more robust.[2]

		DIFFSIM	REJECT
a)	OFFSETSPCA	0.69	0.64
	INDIVPCA	0.70	0.65
	MEANSDIFF	**0.71**	0.68
b)	OFFSETSPCA	**0.32**	0.30
	INDIVPCA	0.000	0.000
	MEANSDIFF	**0.32**	0.30
c)	OFFSETSPCA	0.72	0.47
	INDIVPCA	**0.74**	0.45
	MEANSDIFF	**0.74**	0.47

Table 4: Precision@100 for three methods of extracting the 'linguistic code' component crossed with two methods for ranking candidate word pairs. a) BrEng/Scottish, b) BrEng/GenAm, c) GenAm/AAVE

C Output from our system

Tables 8 to 10 display the top-100 ranked variant pairs generated by our system for each language pair. Although there are admittedly some inconsistencies in the kinds of pairs that were accepted/rejected, our system clearly returned many more clear-cut lexical alternations for BrEng/Scottish and GenAm/AAVE than for BrEng/GenAm. That being said, a lot of the BrEng/GenAm pairs we rejected do accurately reflect cultural differences between the UK and USA.

[2]However, recall that we tuned our embedding hyperparameters using DIFFSIM; another setting might yield better results for REJECT.

D Output from baseline

Tables 5 to 7 show the top-10 ranked variant pairs generated by the simple cosine-similarity baseline for each language pair. None of these were judged to be correct.

Rank	Variant 1	Variant 2	Cosine similarity
1	*umwandlung*	*auftauchen*	*0.99*
2	*verificadas*	*contribuyeron*	*0.99*
3	*umwandlung*	*angeschaut*	*0.98*
4	*benutzer*	*drehte*	*0.98*
5	*donnerstag*	*sonntag*	*0.98*
6	*auftauchen*	*angeschaut*	*0.98*
7	*harryyy*	*ilysssm*	*0.98*
8	*irby*	*pensby*	*0.98*
9	*erscheint*	*ihrer*	*0.98*
10	*dienstag*	*donnerstag*	*0.98*

Table 5: Top 10 generated variables by the baseline for BrEng/Scottish.

Rank	Variant 1	Variant 2	Cosine similarity
1	*tmin*	*tmax*	*1.00*
2	*clr*	*ksjc*	*0.99*
3	*rbngate*	*wpm*	*0.99*
4	*pocus*	*hocus*	*0.98*
5	*imma*	*ima*	*0.98*
6	*klax*	*rmk*	*0.98*
7	*soooo*	*sooo*	*0.98*
8	*aiko*	*jhene*	*0.98*
9	*til*	*till*	*0.97*
10	*clr*	*rmk*	*0.97*

Table 6: Top 10 generated variables by the baseline for BrEng/GenAm.

Rank	Variant 1	Variant 2	Cosine similarity
1	*clr*	*ksjc*	*0.99*
2	*til*	*till*	*0.99*
3	*imma*	*ima*	*0.99*
4	*trynna*	*tryna*	*0.98*
5	*clr*	*rmk*	*0.97*
6	*canceled*	*cancelled*	*0.97*
7	*plz*	*pls*	*0.97*
8	*cus*	*cuz*	*0.97*
9	*horrible*	*terrible*	*0.97*
10	*ksjc*	*rmk*	*0.97*

Table 7: Top 10 generated variables by the baseline for GenAm/AAVE.

Rank	BrEng Variant	Scottish Variant	Score	Rank	BrEng Variant	Scottish Variant	Score
1	**now**	**noo**	**0.54**	*51*	*does*	*disnae*	*0.40*
2	**what**	**whit**	**0.54**	*52*	*getting*	*gettin*	*0.40*
3	*wasnt*	*wis*	*0.51*	**53**	**cant**	**cah**	**0.40**
4	**cant**	**canny**	**0.50**	*54*	*cant*	*couldny*	*0.40*
5	**would**	**wid**	**0.49**	*55*	*being*	*bein*	*0.40*
6	**doesnt**	**disny**	**0.47**	*56*	*was*	*wasny*	*0.40*
7	**cant**	**cannae**	**0.47**	**57**	**yep**	**aye**	**0.40**
8	*going*	*gonny*	*0.47*	**58**	**wasnt**	**wasny**	**0.40**
9	*want*	*wanty*	*0.46*	**59**	**doesnt**	**doesnae**	**0.39**
10	**anyone**	**embdy**	**0.46**	**60**	**couldnt**	**couldnae**	**0.39**
11	**wasnt**	**wisny**	**0.46**	*61*	*werent*	*wisny*	*0.39*
12	**wrong**	**wrang**	**0.46**	*62*	*would*	*wouldny*	*0.39*
13	**yeah**	**aye**	**0.46**	*63*	*need*	*needty*	*0.39*
14	**into**	**inty**	**0.46**	**64**	**abt**	**aboot**	**0.39**
15	**didnt**	**didnae**	**0.46**	**65**	**youd**	**yed**	**0.39**
16	**into**	**intae**	**0.45**	*66*	*bloody*	*fuckin*	*0.39*
17	**im**	**ahm**	**0.45**	**67**	**tomorrow**	**mora**	**0.39**
18	*does*	*disny*	*0.45*	*68*	*going*	*goin*	*0.38*
19	**wouldnt**	**widnae**	**0.45**	**69**	**what**	**wit**	**0.38**
20	*was*	*wisny*	*0.45*	**70**	**couldnt**	**couldny**	**0.38**
21	**nothing**	**nuhin**	**0.45**	**71**	**dont**	**dinna**	**0.38**
22	**isnt**	**isnae**	**0.45**	*72*	*did*	*didnae*	*0.38*
23	**your**	**yur**	**0.44**	*73*	*coming*	*comin*	*0.38*
24	**wouldnt**	**widny**	**0.44**	**74**	**wouldnt**	**wouldnae**	**0.38**
25	**ive**	**ahve**	**0.44**	*75*	*birthdayx*	*brer*	*0.38*
26	**doesnt**	**disnae**	**0.44**	**76**	**gonna**	**gonnae**	**0.38**
27	*going*	*gonnae*	*0.43*	**77**	**cannot**	**cannae**	**0.38**
28	**giving**	**geein**	**0.43**	**78**	**cant**	**canni**	**0.38**
29	**dont**	**dinnae**	**0.43**	**79**	**round**	**roon**	**0.38**
30	**give**	**gie**	**0.43**	**80**	**cant**	**canna**	**0.38**
31	**dont**	**diny**	**0.42**	**81**	**wouldnt**	**wouldny**	**0.37**
32	**cant**	**cany**	**0.42**	*82*	*taking*	*takin*	*0.37*
33	**cant**	**cani**	**0.42**	*83*	*does*	*doesny*	*0.37*
34	**good**	**guid**	**0.42**	**84**	**ok**	**awright**	**0.37**
35	**before**	**afore**	**0.42**	**85**	**just**	**jus**	**0.37**
36	**dont**	**dinni**	**0.42**	**86**	**well**	**weel**	**0.37**
37	**cannot**	**canny**	**0.41**	**87**	**million**	**hunner**	**0.37**
38	**whats**	**whits**	**0.41**	*88*	*anything*	*anythin*	*0.37*
39	**gonna**	**gonny**	**0.41**	**89**	**someone**	**somecunt**	**0.37**
40	**wasnt**	**wisnae**	**0.41**	**90**	**lots**	**hunners**	**0.37**
41	**everyone**	**everycunt**	**0.41**	**91**	**good**	**smashin**	**0.37**
42	*going*	*goni*	*0.41*	**92**	**cannot**	**cany**	**0.37**
43	**outside**	**ootside**	**0.41**	**93**	**kids**	**weans**	**0.37**
44	**going**	**gon**	**0.40**	**94**	**football**	**fitbaw**	**0.37**
45	*cant*	*couldnae*	*0.40*	**95**	**stupid**	**stupit**	**0.37**
46	*fucking*	*fuckin*	*0.40*	**96**	**nobody**	**naebody**	**0.36**
47	*was*	*wisnae*	*0.40*	**97**	**photos**	**photies**	**0.36**
48	*did*	*didny*	*0.40*	*98*	*morning*	*mornin*	*0.36*
49	**anyone**	**anycunt**	**0.40**	**99**	**cannot**	**cani**	**0.36**
50	**round**	**roond**	**0.40**	*100*	*does*	*doesnae*	*0.36*

Table 8: Top 100 generated variant pairs for British English vs Scots/Scottish English. Pairs we accepted are in bold, and those we rejected are in italics.

] Rank	BrEng Variant	GenAm Variant	Score	Rank	BrEng Variant	GenAm Variant	Score
1	**mums**	**moms**	**0.65**	*51*	*lecturer*	*teacher*	*0.39*
2	*dunno*	*idk*	*0.65*	*52*	*stuart*	*scott*	*0.39*
3	*yeh*	*yea*	*0.64*	**53**	**pavement**	**sidewalk**	**0.38**
4	*shouting*	*yelling*	*0.62*	*54*	*horrendous*	*awful*	*0.38*
5	*quid*	*dollars*	*0.61*	*55*	*loosing*	*losing*	*0.38*
6	**learnt**	**learned**	**0.60**	*56*	*cosy*	*comfy*	*0.38*
7	**favour**	**favor**	**0.57**	**57**	**revision**	**studying**	**0.38**
8	**sofa**	**couch**	**0.56**	**58**	**toilets**	**bathrooms**	**0.37**
9	**advert**	**commercial**	**0.56**	*59*	*bbe*	*bby*	*0.37*
10	**adverts**	**commercials**	**0.55**	**60**	**shittest**	**shittiest**	**0.37**
11	**petrol**	**gas**	**0.53**	**61**	**moustache**	**mustache**	**0.37**
12	*vile*	*disgusting*	*0.52*	*62*	*guna*	*gonna*	*0.37*
13	**grandad**	**grandpa**	**0.52**	*63*	*wid*	*wit*	*0.37*
14	*ure*	*ur*	*0.52*	*64*	*sympathetic*	*insensitive*	*0.37*
15	**cos**	**cuz**	**0.52**	*65*	*morn*	*mornin*	*0.37*
16	*yeh*	*yeahh*	*0.51*	*66*	*bbc*	*cnn*	*0.37*
17	*shouting*	*screaming*	*0.49*	*67*	*shittest*	*worst*	*0.36*
18	**cos**	**cus**	**0.49**	*68*	*ion*	*ionn*	*0.36*
19	*favourite*	*fav*	*0.48*	*69*	*defiantly*	*definitely*	*0.36*
20	**honour**	**honor**	**0.48**	*70*	*loads*	*tons*	*0.36*
21	**mummy**	**mommy**	**0.47**	*71*	*shouted*	*yelled*	*0.36*
22	**windscreen**	**windshield**	**0.46**	*72*	*granddaughter*	*daughter*	*0.36*
23	*emirates*	*stadium*	*0.46*	*73*	*favourite*	*fave*	*0.36*
24	*grandad*	*uncle*	*0.46*	*74*	*retard*	*dumbass*	*0.36*
25	**spoilt**	**spoiled**	**0.45**	*75*	*iont*	*ionn*	*0.36*
26	*il*	*ill*	*0.45*	**76**	**spliff**	**bleezy**	**0.36**
27	*nandos*	*sushi*	*0.45*	*77*	*fkn*	*fucken*	*0.36*
28	*photos*	*pictures*	*0.44*	*78*	*arkham*	*batman*	*0.36*
29	*tidy*	*clean*	*0.43*	**79**	**paddys**	**pattys**	**0.36**
30	*gran*	*grandpa*	*0.43*	*80*	*munching*	*eating*	*0.35*
31	*favourites*	*favs*	*0.43*	**81**	**hammered**	**drunk**	**0.35**
32	**slag**	**slut**	**0.42**	*82*	*gatwick*	*airport*	*0.35*
33	*massive*	*huge*	*0.42*	*83*	*cocktails*	*margaritas*	*0.35*
34	*gona*	*gonna*	*0.42*	*84*	*prem*	*league*	*0.35*
35	*netball*	*lacrosse*	*0.42*	*85*	*infront*	*front*	*0.35*
36	**spelt**	**spelled**	**0.42**	*86*	*outreach*	*community*	*0.35*
37	*folk*	*folks*	*0.42*	*87*	*wonna*	*wanna*	*0.35*
38	*nearly*	*almost*	*0.41*	*88*	*tenerife*	*cancun*	*0.35*
39	*dunno*	*idek*	*0.41*	*89*	*wkend*	*wknd*	*0.35*
40	*terrific*	*great*	*0.41*	*90*	*bedroom*	*room*	*0.35*
41	**lecturer**	**professor**	**0.41**	*91*	*pundits*	*analysts*	*0.35*
42	*mep*	*senator*	*0.40*	*92*	*council*	*community*	*0.35*
43	**revising**	**studying**	**0.40**	*93*	*gota*	*gotta*	*0.34*
44	**nans**	**grandmas**	**0.40**	*94*	*amsterdam*	*vegas*	*0.34*
45	*lucozade*	*powerade*	*0.39*	*95*	*truely*	*truly*	*0.34*
46	**cosy**	**cozy**	**0.39**	*96*	*tidying*	*cleaning*	*0.34*
47	*portuguese*	*spanish*	*0.39*	**97**	**flavour**	**flavor**	**0.34**
48	**films**	**movies**	**0.39**	*98*	*unbeaten*	*undefeated*	*0.34*
49	**criticise**	**criticize**	**0.39**	*99*	*een*	*eem*	*0.34*
50	*shops*	*shop*	*0.39*	**100**	**flavours**	**flavors**	**0.34**

Table 9: Top 100 generated variant pairs British English vs General American English. Pairs we accepted are in bold, and those we rejected are in italics.

Rank	GenAm Variant	AAVE Variant	Score	Rank	GenAm Variant	AAVE Variant	Score
1	**the**	**tha**	**0.85**	**51**	**waiting**	**waitin**	**0.65**
2	**with**	**wit**	**0.82**	**52**	**another**	**anotha**	**0.64**
3	**getting**	**gettin**	**0.79**	*53*	*what*	*wat*	*0.64*
4	**just**	**jus**	**0.76**	**54**	**putting**	**puttin**	**0.64**
5	*and*	*nd*	*0.74*	**55**	**something**	**sumthin**	**0.64**
6	**making**	**makin**	**0.74**	*56*	*until*	*til*	*0.64*
7	*when*	*wen*	*0.74*	**57**	**keeping**	**keepin**	**0.64**
8	**looking**	**lookin**	**0.73**	**58**	**throwing**	**throwin**	**0.64**
9	**something**	**somethin**	**0.72**	**59**	**helping**	**helpin**	**0.64**
10	**going**	**goin**	**0.72**	**60**	**laying**	**layin**	**0.64**
11	**being**	**bein**	**0.72**	**61**	**knowing**	**knowin**	**0.63**
12	**doing**	**doin**	**0.72**	**62**	**listening**	**listenin**	**0.63**
13	**taking**	**takin**	**0.71**	**63**	**nothing**	**nuthin**	**0.63**
14	**working**	**workin**	**0.71**	**64**	**thats**	**thas**	**0.63**
15	**something**	**sumn**	**0.71**	**65**	**staying**	**stayin**	**0.63**
16	*someone*	*somebody*	*0.71*	**66**	**shopping**	**shoppin**	**0.63**
17	**watching**	**watchin**	**0.70**	**67**	**telling**	**tellin**	**0.63**
18	**having**	**havin**	**0.70**	**68**	**hoping**	**hopin**	**0.62**
19	**looking**	**lookn**	**0.70**	**69**	**playing**	**playin**	**0.62**
20	**just**	**juss**	**0.70**	*70*	*dont*	*dnt*	*0.62*
21	**everyone**	**errbody**	**0.70**	**71**	**drinking**	**drinkin**	**0.62**
22	*that*	*tht*	*0.69*	**72**	**eating**	**eatin**	**0.62**
23	**thinking**	**thinkin**	**0.69**	*73*	*tonight*	*tonite*	*0.62*
24	*everyone*	*everybody*	*0.69*	**74**	**things**	**thangs**	**0.62**
25	*bc*	*cus*	*0.69*	**75**	**thw**	**tha**	**0.62**
26	**coming**	**comin**	**0.69**	**76**	**wouldve**	**woulda**	**0.62**
27	**over**	**ova**	**0.69**	**77**	**running**	**runnin**	**0.62**
28	**thats**	**dats**	**0.69**	*78*	*this*	*thiss*	*0.62*
29	**pushing**	**pushin**	**0.69**	*79*	*work*	*wrk*	*0.62*
30	*someone*	*sumbody*	*0.68*	**80**	**anymore**	**nomore**	**0.61**
31	*know*	*kno*	*0.68*	**81**	**showing**	**showin**	**0.61**
32	**and**	**n**	**0.68**	*82*	*cannot*	*kant*	*0.61*
33	*anyone*	*anybody*	*0.68*	*83*	*whats*	*wats*	*0.61*
34	**never**	**neva**	**0.68**	**84**	**asking**	**askin**	**0.61**
35	**your**	**yo**	**0.68**	**85**	**my**	**ma**	**0.61**
36	**getting**	**gettn**	**0.68**	**86**	**better**	**betta**	**0.61**
37	**other**	**otha**	**0.67**	*87*	*noting*	*nuthin*	*0.61*
38	**yourself**	**yaself**	**0.67**	**88**	**my**	**mah**	**0.61**
39	**even**	**een**	**0.67**	**89**	**driving**	**drivin**	**0.61**
40	*school*	*skool*	*0.67*	**90**	**sleeping**	**sleepin**	**0.61**
41	**little**	**lil**	**0.67**	**91**	**cannot**	**caint**	**0.60**
42	**with**	**widd**	**0.67**	*92*	*back*	*bacc*	*0.60*
43	*from*	*frm*	*0.66*	**93**	**to**	**ta**	**0.60**
44	**nothing**	**nothin**	**0.66**	*94*	*but*	*bt*	*0.60*
45	*bc*	*cuz*	*0.65*	*95*	*releasing*	*droppin*	*0.60*
46	*about*	*abt*	*0.65*	**96**	**dropping**	**droppin**	**0.60**
47	**morning**	**mornin**	**0.65**	**97**	**giving**	**givin**	**0.60**
48	**seeing**	**seein**	**0.65**	**98**	**nothing**	**nun**	**0.59**
49	**wearing**	**wearin**	**0.65**	**99**	**starting**	**startin**	**0.59**
50	**wanting**	**wantin**	**0.65**	*100*	*thay*	*dat*	*0.59*

Table 10: Top 100 generated variant pairs for General American English vs African American Vernacular English. Pairs we accepted are in bold, and those we rejected are in italics.

6.4 Comments on the paper

6.4.1 Principal component analysis

In the paper's Supplementary Information (§B), we presented a comparison of three methods for defining a 'linguistic code' component. These were MEANSDIFF: the vector difference of the mean of the L1 seed word embedding and the mean of the L2 seed word embeddings; INDIVPCA: the first principal component of the whole set of seed word embeddings; and OFFSETSPCA: the first principle component of the set of vector differences between L1 seed word embeddings and their corresponding L2 seed word embeddings. These performed more or less equally well, except that the INDIVPCA method was completely ineffective for the British English / General American English language variety pair.

Figures 6.1, 6.2, and 6.3 show the seed word embeddings projected onto their first two principal components, for BrEng/Scottish, GenAm/AAVE, and BrEng/GenAm respectively. In Figures 6.1 and 6.2, the first principal component quite cleanly delineates the two linguistic codes, but this is not the case for BrEng/GenAm (Figure 6.3). It appears, then, that the distinction between British and General American English is less strongly encoded in the embedding space than is the case for our other language variety pairs.

One potential explanation for this has to do with the composition of the corpora on which we trained our embeddings:

- The embeddings we used to identify **BrEng/Scottish** variables were trained on the concatenation of our SG-Users and IH-Users datasets (see §3.4), which consist of tweets posted in autumn 2014 by users sampled on the basis of having used either Scottish geotags or hashtags relating to Scottish independence. Distinctively Scottish terms are reasonably evenly (though sparsely) distributed throughout this corpus.

- The embeddings we used to identify **GenAm/AAVE** variables were trained on our G-USA dataset, which consists of tweets with USA geotags from the 'Spritzer' sample between June 30th 2013 and July 1st 2016. Terms distinctive to African American English are reasonably evenly distributed throughout this corpus.

- The embeddings we used to identify **BrEng/GenAm** variables were trained on the concatenation of our G-UK dataset—which consists of tweets with UK geo-

> tags from the 'Spritzer' sample between September 1st 2013 and September 30th 2014—and our G-USA dataset. Terms distinctive to British English are *not* evenly distributed throughout this corpus; instead they are predominantly concentrated within the first sixth of it (i.e. within the G-UK dataset).

Perhaps it is significant that all of the UK tweets (and therefore most instances of British English variants) are at the beginning of the corpus, whereas the other two language variety pairs are interleaved throughout their respective corpora. Antoniak and Mimno (2018) investigated factors affecting the stability of geometric relationships among word embeddings, and found that *randomly* shuffling the order of documents in a training corpus did not substantially change the geometry of the embeddings. However, they did report that

> "anecdotally, we had observed cases where the embeddings were affected by groups of documents (e.g. in a different language) at the beginning of training." (p. 117)

On the other hand, the apparently weaker encoding of the variability between British and American English may be due to an inherent difference in the nature of that variability. Both the African American and distinctively Scottish language varieties include numerous distinctive variants of so-called closed-class words, whereas all of our BrEng/GenAm seed pairs, and most of the additional BrEng/GenAm pairs we identified using our system, are open-class. Closed-classed words (a.k.a. function words) tend to be frequent and contextually diverse, and the more a word-form co-occurs with other word-forms that are distinctive to the same language variety, the more likely its vector should be to diverge from that of the equivalent word-form in the other language variety. It may therefore be interesting for future work to explore the hypothesis that distinctive function words are particularly helpful in providing a signal during embedding training which helps to distinguish lexical variants associated with different language varieties. Note that this would not necesarily mean that the seed pairs we input to our system would need to include function words in order for it to be effective (indeed, as noted in the paper, we actually observed slight increases in performance for both BrEng/Scottish and GenAm/AAVE when we did *not* include closed-class seeds). Rather, if distinctive function words occur in the training corpus then these may help to shape the embedding space such that a strong 'linguistic code' component emerges, and can then be identified using any reasonable seed-pair set. If this were the case then our system would generally be less effective for pairs of language varieties across which there is little variation in the forms of function words.

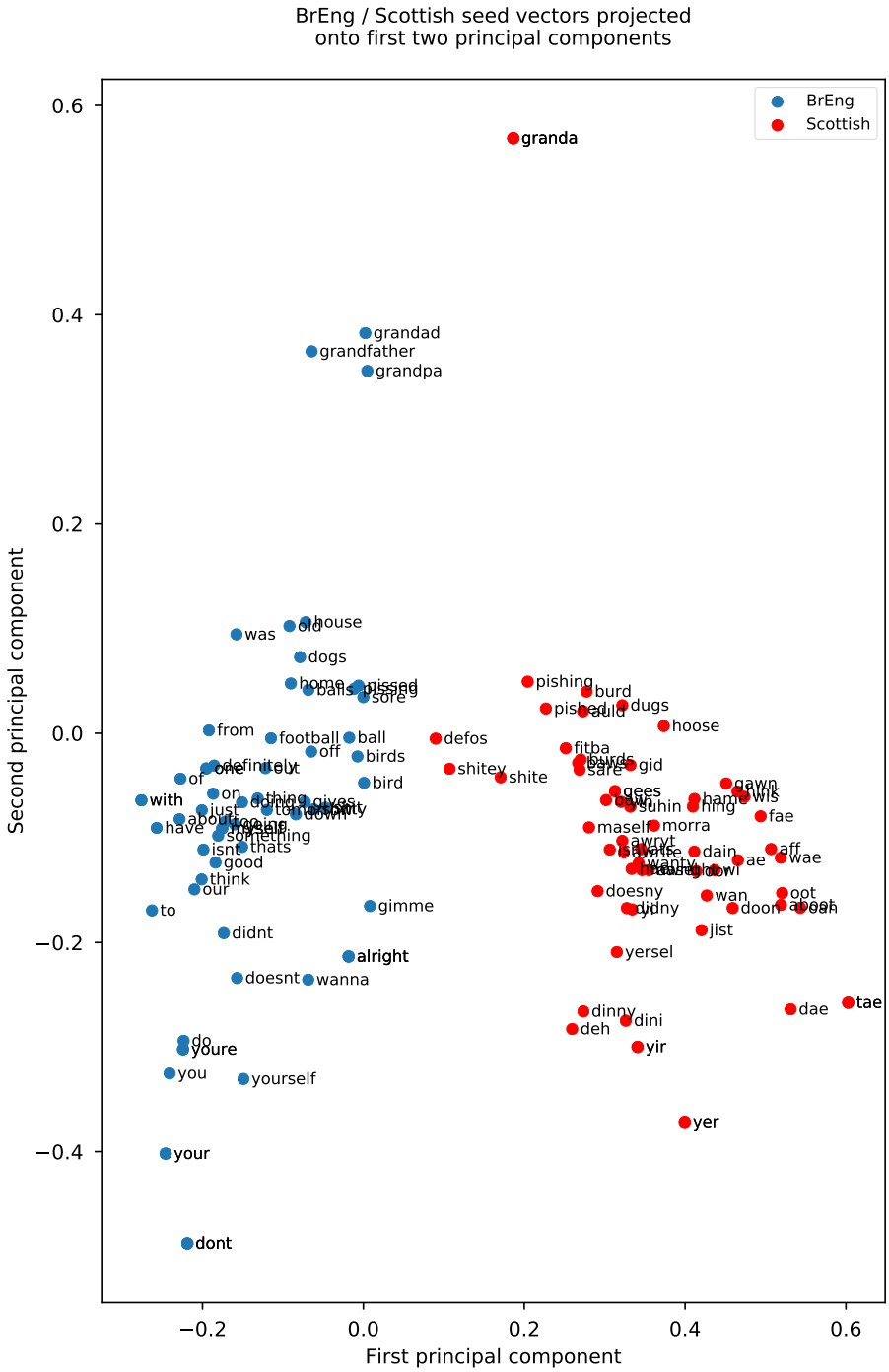


Figure 6.1: British English / Scottish seed word embeddings projected onto their first two principal components.

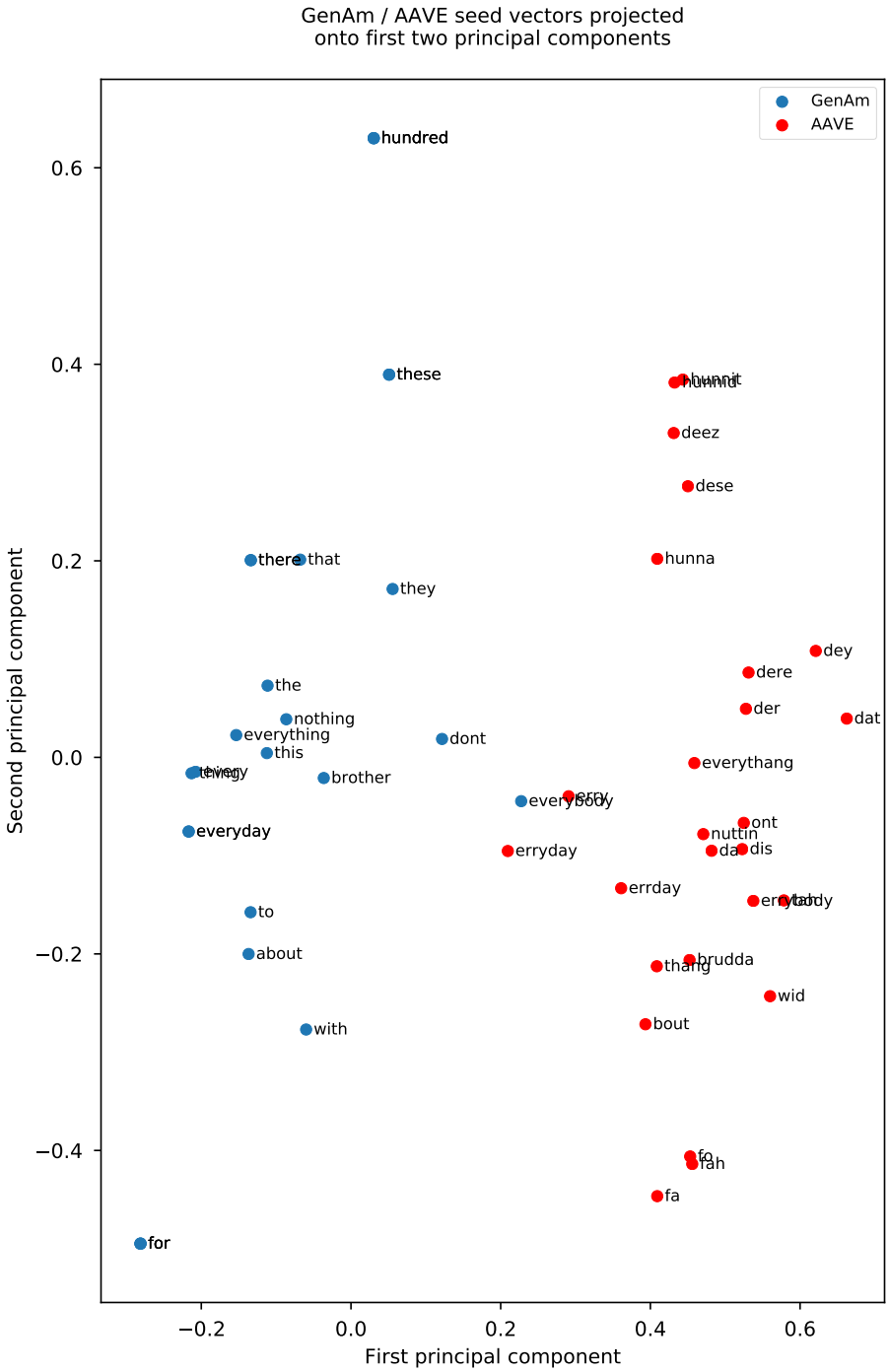


Figure 6.2: General American English / African American English seed word embeddings projected onto their first two principal components.

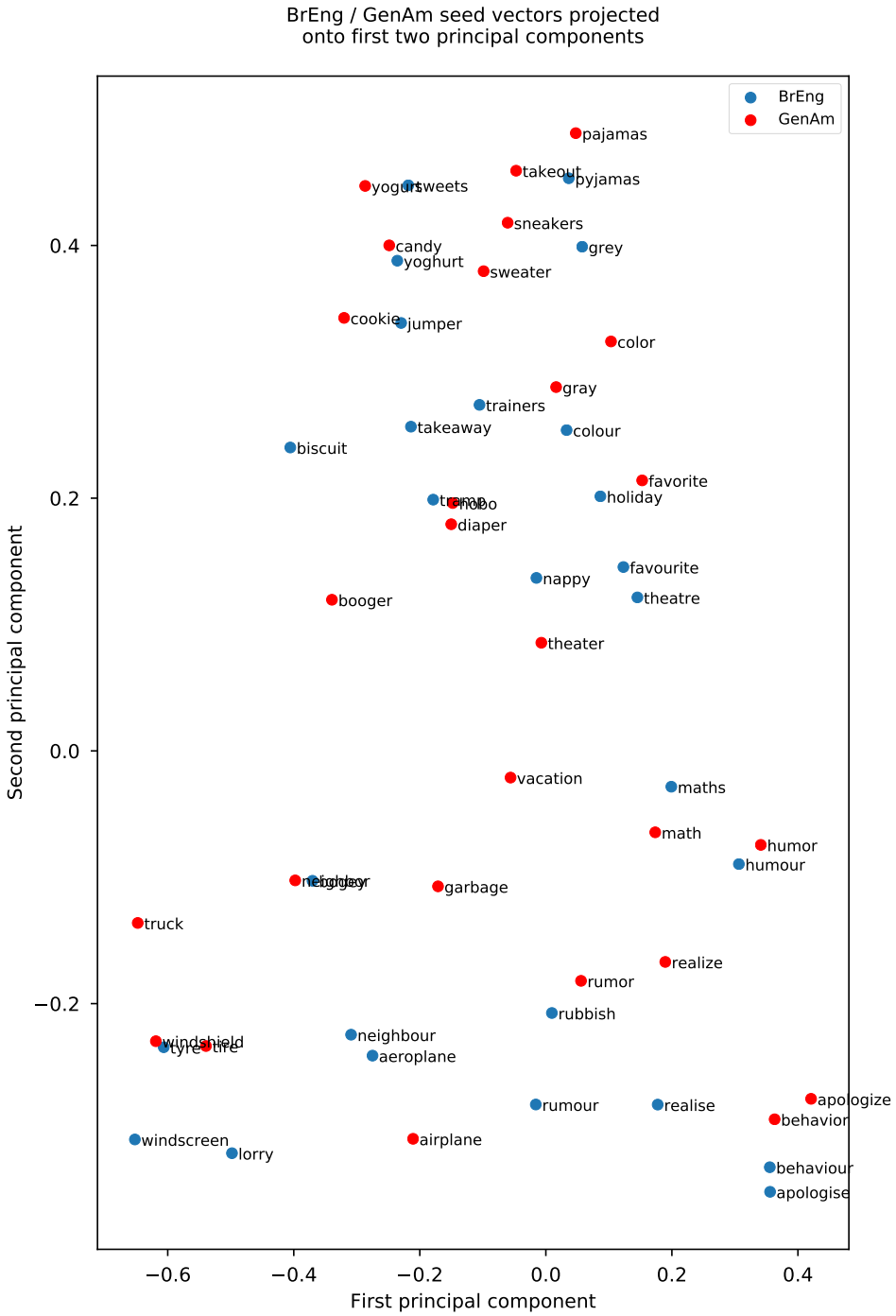


Figure 6.3: General British English / General American English seed word embeddings projected onto their first two principal components.

6.4.2 Other related work

In addition to the related works mentioned in the paper, a few others have also addressed the problem of lexical variable discovery. Before the advent of Word2Vec (Mikolov et al., 2013) and other neural network based distributional semantic models, Peirsman et al. (2010) used distributional models based on co-occurence matrices weighted by Pointwise Mutual Information (PMI). They addressed the same objectives as we do—automatically detecting words which are distinctive to one linguistic variety, and pairing them with synonyms from another—but whereas we address both of these objectives simultaneously within a single module, Piersman et al addressed them separately. Their cross-lectal synomy detection approach requires as input a list of words which are known to be distinctive to one of the language varieties in question. Because the accuracy of their lectal marker system is only around 25%, this means they still must go through the labour-intensive process of manually filtering its output, or else obtain a gold standard list of lectal markers from elsewhere. The main advantage of our all-in-one approach is that it obviates this step. Another key difference is that Piersman et al's approach is designed to be applied to two separate monolingual corpora, whereas ours is designed for a single code-mixed corpus.

Gouws et al. (2011) also used PMI-based distributional models to identify semantically similar pairs of terms across linguistic varieties, focusing on variation across domains, rather than geographical regions. They train two separate distributional models, one on a 'common English' corpus and one on a domain-specific corpus. They then identify pairs of terms which are semantically similar according to the domain-specific model but not the 'common' English model. Finally, they re-order the remaining pairs according to their orthographic similarity. An advantage of their method is that it doesn't require any seed pairs, but on the other hand it does require two separate corpora. This method could potentially be applied to our own use-case if we were able to obtain a relatively 'pure' corpus for one language variety, and use a code-mixed corpus for the 'domain-specific' model; we'd then be looking for pairs of words which are similar in the code-mixed corpus but not in the single-code corpus. The two corpora would both need to be from the same domain, or else this method would also return word pairs consisting of variants from either domain, as opposed to either regional language variety. Moreover, if the code-mixed corpus happened to contain other language varieties beyond the two of interest to the study, pairs including words belonging to these other varieties would likely also be returned. An advantage of the method we

propose here is that using a small set of representative seed pairs we can specifiy the dimension of variation we are interested in. Another advantage of our method is that while Gouws et al. (2011) focus on pairs of variants which are both semantically and *orthographically* similar, our method is also able to identify pairs of words which are semantically equivalent but whose orthographic forms are completely unrelated (e.g. *kids*/*weans*; *pavement*/*sidewalk*).

More generally, while we introduced our task as a special case of unsupervised bilingual lexicon induction, it also bares close relation to lexical substitution tasks (McCarthy and Navigli, 2007; Mihalcea et al., 2010; Melamud et al., 2015), where the aim is to identify alternative words that can be substituted for a target word within a specific sentential context, while preserving the referential meaning and grammaticality of the sentence. Special kinds of lexical substitution tasks include text normalisation, wherein the aim is to identify words which are out-of-vocabulary for a downstream natural language processing model, and replace them with equivalent in-vocabulary words (Baldwin et al., 2015; Dirkson et al., 2019; Muller et al., 2019), and lexical simplification, wherein the aim is to identify complex words and replace them with simpler equivalents (Specia et al., 2012; Paetzold and Specia, 2017; Kriz et al., 2018). Lexical simplification is perhaps even closer to our task than major-world-language bilexicon induction is, since deciding which terms should be labeled as belonging to 'complex' or 'simple' varieties can be far from trivial (Paetzold and Specia, 2016; Yimam et al., 2018; Lee and Yeung, 2018)[1], just as it is far from trivial to decide which terms are unequivocally 'Standard' English and which are unequivocally Scots or Scottish English (or African American English, or British as opposed to American, etc.). A key way in which our lexical variable discovery task differs from the typical lexical substitution formulation is that we have tasked ourselves with finding substition pairs which work generally across a variety of contexts, rather than in one specific sentential context.

[1]That said, in some lexical simplification settings there is an *a priori* list of complex terms to be replaced, such that the task is only to identify and rank simplification candidates for those specific terms; and not to identify which terms *can* be simplified in the first place.

6.5 Future work

6.5.1 Better ranking

The performance of our system could potentially be improved by using negative examples as well as positive ones, and/or combining the ranking of candidate pairs based on embeddings with rankings based on other features (e.g. string similarity).

6.5.2 Context-sensitive substitutions

Our original motivation for developing a system for curating large sets of lexical sociolinguistic variables was our need to automatically identify, with high coverage in a large corpus, instances either of distinctively Scottish lexical variants, or of Standard English lexical variants for which a distinctively Scottish variant could be substituted. Rather than relying on a set of context-indepedent lexical substitution pairs (however large), this particular use-case could arguably be better served with a context-specific approach, which would enable us to consider pairs of word-forms which are interchangable in some contexts but not in others, and thereby obtain higher coverage of all the occasions on which users had the opportunity to use a distinctively Scottish variant.

Lexical simplification systems often follow a pipeline in which the first step (after the identification of complex words) is to generate context-*insensitive* candidate substitutes, and a subsequent step is to decide which of these can replace the target complex word in its specific sentential context without altering the meaning or grammaticality (Paetzold and Specia, 2017). A similar approach could perhaps be applied for our own use-case: first, the system we have presented here could be used to generate lexical variant pairs in a context-insenstive manner; we would then need to identify all instances in our dataset of either member of any of the resulting lexical variant pairs; and then for each instance we would need to decide whether the other variant in the relevant pair could really replace the observed variant in this particular sentential context without altering its referential meaning or grammaticality. Thus one potential avenue for future work is to explore whether similar approaches to those developed for Substitute Selection in lexical simplification tasks might be effective for this use-case.

6.5.3 Semasiological variables

Note that this thesis has been concerned with identifying ONOMASIOLOGICAL variables, i.e. pairs of words which have the same meaning, but differ in their form.

Another interesting direction for future work might be to identify SEMASIOLOGICAL variables, i.e. pairs of terms which are identical in form, but differ in their meanings (e.g. the term *suspenders* in British English denotes a garment used to hold up stockings and worn around the waist, while in US English this term instead denotes a garment used to hold up trousers and worn over the shoulders). Where separate corpora are available for the language varieties of interest, wordforms which exhibit semasiological variation across these language varieties could be identified by aligning and comparing word embeddings trained on each corpus, using similar techniques to those which have been developed for identifying words whose meanings change over time (e.g. Shoemark et al., 2019).

6.5.4 Leveraging sub-corpus level metadata

Developing an approach for identifying semasiological variables which can be applied to a single code-mixed corpus (without language variety labels at the word, sentence, or document level) would be more challenging, of course. However, sub-corpus level metadata regarding geographic location could perhaps be leveraged as a weak signal for the *likely* distribution over language varieties.

A model for incorporating sub-corpus level metadata to learn semantic representations of words which capture semasiological variation across different contexts has already been introduced by Bamman et al. (2014a), who used it to produce geographically-informed word embeddings from a dataset of U.S. tweets. They jointly learned a 'global' embedding matrix for the United States as a whole, and additional matrices for each individual state. The embedding for a particular instance of a word was computed as the sum of its global and state-specific vectors, such that the state-specific vectors indicate how the global representations should be shifted to reflect state-specific semantics. Kulkarni et al. (2016) used a similar model to identify statistically significant semantic differences in British vs. U.S. English.

Geographical metadata could also potentially be leveraged in the onomasiological variable induction task we have introduced here. We did not think the approaches of Bamman et al. (2014a) and Kulkarni et al. (2016) would add much value for the onomasiological task, since they are designed to reveal geographical differences in the meanings of individual terms. Suppose we were to use such an approach to train a 'global' embedding matrix using all our UK tweets, along with a Scottish-specific matrix using just the tweets from Scotland. The global embedding for the term *greet-*

ing might have nearest neighbours like *card* and *meeting*, while its Scotland-specific deviation vector might shift it closer to *crying* and *weeping*, reflecting the different meanings of *greeting* in English and Scots. However, the Scotland-specific deviation vector for a term like *hoose*—which does not have a different meaning in the rest of the UK than in Scotland, but is simply less frequently used in the rest of the UK (where the Standard English equivalent *house* is relatively more frequent)—would presumably not represent much of a shift, since the instances used to learn its Scotland-specific vector would make up most of those used to learn its 'global' vector, and any additional instances from the rest of the UK would be likely to occur in similar linguistic contexts.

A different approach to leveraging sub-corpus level metadata which perhaps could be useful for identifying onomasiological variables is to use Doc2Vec, a.k.a. Paragraph Vectors (Le and Mikolov, 2014), an extenstion of Word2Vec which embeds document labels in the same space as words. Word and document vectors can be trained simultaneously such that vectors representing semantically similar words end up close together (as with basic Word2Vec), but additionaly vectors representing linguistically similar documents also end up close together, and vectors representing words which are strongly associated with particular documents end up close to the vectors representing those documents (Lau and Baldwin, 2016). Hovy and Purschke (2018) used Doc2Vec to learn embeddings for cities from location-based threads on the social media platform Jodel. They were able to reproduce regional linguistic distinctions by clustering the resulting city embeddings, since these were indirectly based on the words used in the respective cities. Future work could explore whether this method of simultaneously learning embeddings for words and geographical regions (or indeed other kinds of metadata labels) could be used to enhance the extent to which sociolinguistic variation is encoded in the geometric relationships among word vectors.

If it is the case that our current approach is less effective for pairs of language varieties which lack distinctive function words (as I tentatively suggested in §6.4.1), then an approach which leverages additional metadata might be particularly beneficial for such language variety pairs. An exciting (though entirely speculative) prospect is that rather than identifying new variant pairs by comparing them with an average difference vector of representative seed pairs as in the system we have presented here, using something like Doc2Vec we might instead be able to compare them directly with the difference vector of the relevant geographical labels—thus obviating the need for seed pairs entirely.

Chapter 7

Conclusions

In this thesis we sought to answer questions about the use of Scots and Scottish English on Twitter, and to advance methodologies for analysing patterns of sociolinguistic variation in social media text more generally. In Chapters 4 and 5 we have presented two large-scale quantative analyses of extra-linguistic factors which condition usage rates of distinctively Scottish vocabulary on Twitter, and in Chapter 6 we have proposed a framework for using word embeddings as a tool to facilitate the curation of lexical alternation variables, which can both save time for researchers and reveal variables they may not have otherwise considered.

As we discussed in Chapter 2, social media text differs from other written domains in that it is typically both informal and public-facing, and it differs from everyday speech in that it is inherently persistent and replicable, often with the potential to be viewed by anyone in the world at any time after it has been produced. Relaxed orthographic norms license users to experiment with ways to authentically represent their speech in written form, and the indeterminacy of the audience may strenghten the impetus to index aspects of one's identity, stance, or affect through linguistic means. For some speakers of minority language varieties, social media provides an unprecedented opportunity to write in their native language without being corrected or chastised. Social media has also made minority language varieties such as Scots more visible to people outside of their offline communities of practise, as audiences on social media can be much larger and more geographically and demographically diverse than an individual's offline social network.

While sociolinguistic studies have traditionally used surveys or interviews to elicit particular variables of interest in carefully controlled contexts, an alternative approach is to use large, naturalistic corpora. Corpus-based methods enable the researcher to

avoid the Observer's Paradox, to collect data from many more 'informants' than would be practical using traditional elictation methods, and, given sufficient volumes of linguistic data and meta-information, a naturalistic corpus can be used to identify social categories and sociolinguistic variables in a transparent, data-driven way, rather than relying on intuition to determine which linguistic variables and extra-linguistic covariates are worthy of detailed analysis. The abundance of spontaneously produced informal language on social media, and its public and permanent nature, make it easier than ever before to collect large datasets for quantitative analyses of subtle sociolinguistic phenomena. That being said, various methodological challenges are involved in quantatively analysing minority language usage on social media.

First, it can be difficult to quantify usage rates of minority language varieties like Scots which have a high degree of overlap in vocabulary and grammar with another language variety—especially when these are often code-mixed within a single utterance. For this reason, we chose to focus on the lexical level, measuring usage rates of lexical variants which are distinctive to Twitter posts from Scotland. Defining the envelope of variation for a lexical analysis brings its own challenges, as existing dictionaries and word-lists are unlikely to provide good coverage of the innovative words and spelling variants in use on social media. Furthermore, we have argued for the use of alternation variables in order to control for variation in denotational meaning as opposed to form, which requires matching words from one language variety with denotationally equivalent words from another. Many won't have single-word equivalents, some will have single-word equivalents that only work in some contexts, and some will have multiple single-word equivalents that could easily be missed, making this a particularly arduous task.

There are also challenges involved in collecting social media data, which we discussed primarily with regards to Twitter in Chapter 3. While it can be difficult to reach certain populations using traditional sociolinguistic data collection methods such as interviews and surveys, when collecting large-scale social media datasets it can be difficult to even determine whether or not the target population has been reached. The biographical information associated with social media posts in the form of meta-data or user profiles is limited, and cannot always be taken at face value. Relying on public self-disclosure to determine demographic information risks systematically biasing the resultant dataset toward certain subgroups of the target population, and inadvertantly excluding others. Furthermore, aside from the difficulty of appropriately delineating the target population in the absence of reliable criteria, the sampling methodologies

used by Twitter's APIs are not transparent, and can introduce additional unintended and unquantifiable biases. Since prior work has indicated that Twitter's Streaming API *does* appear to provide a simple random sample of all Twitter traffic, we opted to post-filter that in order to collect an unbiased sample of tweets and users which met our criteria. However, since the Streaming API yields only 1% of Twitter traffic, this did limit the sizes of the datasets we were able to collect. Due to the ethical concerns we discussed in §3.3.1, we cannot release our datasets in full, but we have made them publically available at https://doi.org/10.5281/zenodo.3517244 in the form of tweet and user IDs, such that they can be re-collected by other researchers for use in their own studies.

7.1 Contributions

We conducted the first large-scale sociolinguistic study of British tweets, which was also the first to examine the relationship between sociolinguistic variation and political views using social media data. We collected tens of thousands of tweets posted during the 2014 Scottish independence referendum campaign, and investigated how people's usage rates of distinctively Scottish words varied in relation to their views on Scottish independence, the topic of their tweets, and the size of their target or imagined audience. We established that use of the Scots language and regionally-specific terms and spellings is prevalent on Twitter and corresponds to features known in the linguistic literature about Scots and Scottish English, though we also identified some new distinctively Scottish terms which are specific to social media text (the acronyms *MWI* and *GTF*). We found that people who used hashtags indicating support for Scottish independence tended to use more distinctively Scottish vocabulary than those who opposed it, but that both groups modulate their usage in relation to the topic and the likely composition of their audience.

While previous works have suggested that users modulate their usage of non-standard and geographically-specific language with respect to the size of their audience, we have taken greater care to control for the potential confound of topic, and to evaluate the extent to which our findings generalise across different subpopulations. Our results indicate that audience and topic have independent effects on the rate of distinctively Scottish usage in two groups of users sampled using different criteria. We observed a clear relationship between the topic or genre of discussion and the odds of choosing Scottish variants in both groups, but the sizes and directions of the au-

dience effects were not consistent across the two groups. Since topic emerged as a clear conditioning factor, we recommend including it in future studies of factors which condition variation on social media. While we leave the testing of our hypothesised explanations for the differences in audience effects across the two user groups to future work, in highlighting these differences we hope to stimulate more engagement with questions about the representativeness of user samples and the generalisability of findings in quantitative sociolinguistic studies using social media data.

In order to ensure that the effects we measured were truly effects on what language variety people were choosing to use to refer to things—and not on which things they were choosing to refer to—it was necesary to analyse peoples' relative usage frequencies of different words which mean approximately the same thing (e.g. *bairns* vs *children*). If done manually, the process of identifying terms which are distinctive to one language variety and then pairing these with semantically equivalent words from another variety can be extremely labour intensive. Furthermore, if researchers rely on their own experience and intuition to select the word pairs on which their analyses will be based, this can lead them to systematically miss datapoints from particular settings or segments of the community. To facilitate this process we devised a data-driven, computational method which can make identifying relevant word pairs much quicker, and can also suggest variant pairs a researcher might not have otherwise considered.

7.2 Future directions

Further to the suggestions we made in Chapters 4 to 6 for future work that could build on each of the papers individually, we will now propose some more directions for future work which pertain to the body of work we have presented here as a whole.

7.2.1 Representativeness of variable sets

The lexical variable induction method we introduced in Chapter 6 facilitates the curation of large sets of lexical alternations. The ability to efficiently gather large sets of lexical variables for use in quantitative sociolinguistic studies opens the door for future work to repeat studies of this kind using several different subsets of a larger variable set, and assess how sensitive the results are to the particular set of variables used.

7.2.2 Other language varieties

Though the studies we presented in Chapters 4 and 5 were concerned with the Anglic language varieties of Scotland, the same methodologies could be applied to other language varieties. Indeed, as dicussed in §4.5, our study on the relationship between regionally specific language use and support for independence has been replicated with respect to Catalan (Stewart et al., 2018). In that replication study, the analysis was conducted at the tweet level rather than the lexical level, since compared with Scots and English, Spanish and Catalan are less commonly code-mixed within a single tweet; and unlike Scots, Catalan is supported by well-established off-the-shelf language classification tools. The same approach could potentially be used for Scottish Gaelic, which belongs to a distinct branch of the Indo-European family from Scots and English, and is supported by the Compact Language Detector toolkit[1].

On the other hand, there are many minority language varieties which are not supported by existing NLP tools or which are frequently code-mixed with other varieties, for which a lexical-level analysis may be necesary. For example, the relationship between Swiss German and Standard German resembles that of Scots and English in several ways: Swiss German is a dialect continuum, widely used in everyday speech throughout the German-speaking regions of Switzerland; but while Standard German is an offical language of Switzerland, Swiss German is not. Like Scots, Swiss German lacks a standard written orthography and is rarely used in formal writing (where Standard German is used instead), but is increasingly being used in informal writing, such as on social media. Like Scots and English, Swiss German and Standard German have some vocabulary in common, such that it may not always be possible to unequivocally asign a short post to one variety or the other. Moreover 'Swiss German', like 'Scots', does not refer to a single homogenous variety, but a multitude of dialects. Swiss German and Standard German would therefore be an interesting language pair on which to test our lexical variable induction system. It would also be interesting to investigate whether or not usage rates of Swiss German lexis on Twitter are conditioned by topic in the same way as we found for distinctively Scottish lexis. We would not necesarily expect this to be the case, since Swiss German is widely spoken in most social contexts, and is not associated with inferior education or social status in the same way as Scots sometimes still is.

[1] https://github.com/google/cld3

7.2.3 Diachronic analyses

The analyses in Chapters 4 and 5 are based on data from 2014, the year in which the Scottish Independence Referendum was held. To be sure, this was an interesting time period in which to study distinctively Scottish vocabulary usage on Twitter, but since then 'Scottish Twitter' has taken off as something of a cultural phenomenon, in many ways echoing the emergence of the construct of 'Black Twitter' (see Florini, 2014). Tweets employing distinctively Scottish language and humour became increasingly popular with Twitter users around the world, and then were introduced to audiences beyond Twitter through their dissementation in listicles (e.g. Bailey, 2015; Robinson, 2019), and eventually in newspaper and magazine comment pieces (e.g. Livingston, 2019; Russell, 2019). By Autumn 2019, the construct of 'Scottish Twitter' had become so well established (and so widely adored) that Twitter paid homage to it with a physical pop-up museum, the 'Scottish Twitter Visitor Centre', at the Edinburgh Fringe Festival[2].

It would be interesting, therefore, for future studies to investigate empirically whether usage rates of Scots and Scottish English have increased since 2014, in tandem with the rise of 'Scottish Twitter' as a cultural phenomenon. If there has been an increase in usage, it would be interesting to establish whether this is primarily confined to humurous tweets, of the sort that are widely shared and celebrated in 'Scottish Twitter' listicles, or whether usage rates have also risen across a range of topics. Future studies might also investigate whether there are particular Scottish variants whose usage has changed over time, or whether patterns of moderation with respect to topic and/or audience have changed since 2014, regardless of whether or not overall usage has increased.

7.2.4 Applications in NLP

In addition to being used to curate variables for use in sociolinguistic studies, other potential applications of our lexical variable induction system (Chapter 6) include text normalisation and data augmentation, to improve performance of NLP systems on dialects and minority language varieties for which large training corpora are lacking. Although raw social media text is freely available in abundance, many NLP systems require labeled training data, and the requisite labels are typically very expensive to obtain. Even when large social media datasets with the requisite labels do exist, minority language varieties are typically underrepresented within them. Due to the rich

[2] https://Twitter.com/TwitterUK/status/1163816121926991873

diversity of styles and varieties that are used in social media text, training separate systems individually for each of them is infeasible (not least due to the high prevalance of code-mixing; AlGhamdi et al. 2016). Improving the performance of NLP systems on minority language varieties is particularly imporant from the standpoint of fairness, given that communities who use these varieties are often disempowered relative to speakers of the dominant language variety; and developing NLP tools which work well only for the dominant variety can perpertuate or even exacerbate this imbalance (Blodgett and O'Connor, 2017). With respect to quantitative sociolinguistic analyses, improving performance of NLP tools such as Part-Of-Speech taggers and dependency parsers could enable us to move beyond beyond the lexical level and also identify minority language usage on the basis of more complex constructions.

One way to improve the performance of NLP systems on texts which are not well represented by the training data is to normalise such texts as a pre-processing step before they are input to the system. Lexical substitution is a common approach to text normalisation (e.g. Gouws et al., 2011), and thus our system—designed specifically for identifying lexical substitution pairs from a single, unlabelled, code-mixed corpus—could be useful here.

A drawback of normalisation approaches is that they can obscure the social meaning and pragmatic information encoded in the original lexical choices (Eisenstein, 2013), which for some downstream tasks may be important to preserve. An alternative solution is to boost the representation of minority varieties in the training dataset by augmenting it with synthetic examples (e.g. Fadaee et al., 2017). Our system could potentially be used to produce synthetic examples by taking sentences from the dominant language variety and making lexical substitutions from other dialects or closely related / frequently code-mixed languages.

Of course, there are inherent limitations to text normalisation and data augmentation approaches based on lexical substitutions. To quote the Scots grammarian Purves (2002, p.7): "A passage in English cannot be transformed into Scots simply by substituting Scots words for English words without reference to structure and idiom." That being said, while lexical substitutions may not be sufficient to produce what Purves considers 'good Scots', their use in NLP pipelines may nevertheless help to improve performance on Scottish social media texts. After all, as Purves (2002, p.9) himself concedes: "bannoks is better nor nei breid."[3]

[3] An English equivalent to this Scots proverb is 'half a loaf is better than none'.

Bibliography

Aitken, A. J. (1982,2015). Bad Scots: some superstitions about Scots speech. In Macafee, C., editor, *Collected Writings on the Scots Language*. (2015), [online] Scots Language Centre. http://medio.scotslanguage.com/library/document/aitken/Bad_Scots_some_superstitions_about_Scots_speech_(1982) (Accessed 2019-02-05). Originally published Scottish Language 1 (1982), 30–44.

AlGhamdi, F., Molina, G., Diab, M., Solorio, T., Hawwari, A., Soto, V., and Hirschberg, J. (2016). Part of speech tagging for code switched data. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 98–107, Austin, Texas. Association for Computational Linguistics.

Androutsopoulos, J. (2014). Languaging when contexts collapse: Audience design in social networking. *Discourse, Context & Media*, 4:62–73.

Antoniak, M. and Mimno, D. (2018). Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics*, 6:107–119.

Artetxe, M., Labaka, G., and Agirre, E. (2018). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, 59(4):390–412.

Bailey, L. (2015). 35 reasons Scottish Twitter is the wildest place on the internet. *BuzzFeed*. Retrieved from: https://www.buzzfeed.com/lukebailey/oot-the-nite-aye/. Accessed: 2019-10-24.

Baldwin, T., de Marneffe, M. C., Han, B., Kim, Y.-B., Ritter, A., and Xu, W. (2015). Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 126–135, Beijing, China. Association for Computational Linguistics.

Bamman, D., Dyer, C., and Smith, N. A. (2014a). Distributed representations of geographically situated language. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 828–834, Baltimore, Maryland. Association for Computational Linguistics.

Bamman, D., Eisenstein, J., and Schnoebelen, T. (2014b). Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

Blodgett, L. S., Green, L., and O'Connor, B. (2016). Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130. Association for Computational Linguistics.

Blodgett, L. S. and O'Connor, B. (2017). Racial disparity in Natural Language Processing: A case study of social media African-American English. In *Fairness, Accountability, and Transparency in Machine Learning (FAT/ML) Workshop, KDD*.

Bode, L., Hanna, A., Yang, J., and Shah, D. V. (2015). Candidate networks, citizen clusters, and political expression: Strategic hashtag use in the 2010 midterms. *The Annals of the American Academy of Political and Social Science*, 659:149.

boyd, d. m. (2008). *Taken out of context: American teen sociality in networked publics*. PhD thesis, University of California, Berkeley.

boyd, d. m., Golder, S., and Lotan, G. (2010). Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter. In *43rd Hawaii International Conference on System Sciences*, pages 1–10. IEEE.

Bruns, A. and Highfield, T. (2013). Political networks on twitter: Tweeting the queensland state election. *Information, Communication & Society*, 16:667.

Bruns, A. and Stieglitz, S. (2014). Twitter data: What do they represent? *it - Information Technology*, 56(5):1–7.

Clark, T. (2018). Nobody actually talks like that – Why the fear of Scots? *The National*. Retrieved from: https://www.thenational.scot/news/17262527.nobody-actually-talks-like-that-why-is-everyone-terrified-of-scots/. Accessed: 2019-10-24.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2017). Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Conover, M. D., Davis, C., Ferrara, E., McKelvey, K., Menczer, F., and Flammini, A. (2013). The geospatial characteristics of a social movement communication network. *PloS one*, 8(3):e55957.

Conover, M. D., Gonçalves, B., Flammini, A., and Menczer, F. (2012). Partisan asymmetries in online political activity. *EPJ Data Science*, 1:1.

Conover, M. D., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., and Flammini, A. (2011). Political polarization on twitter. In *Fifth international AAAI conference on weblogs and social media*.

Dirkson, A., Verberne, S., Sarker, A., and Kraaij, W. (2019). Data-driven lexical normalization for medical social media. *Multimodal Technologies and Interaction*, 3(3):60.

Doyle, G. (2014). Mapping dialectal variation by querying social media. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 98–106. Association for Computational Linguistics.

Duggan, M. (2015). Mobile messaging and social media 2015. http://www.pewinternet.org/2015/08/19/mobile-messaging-and-social-media-2015/.

Duggan, M. and Brenner, J. (2013). The demographics of social media users, 2012. http://www.pewinternet.org/2013/02/14/the-demographics-of-social-media-users-2012/.

Duggan, M., Ellison, N. B., Lampe, C., Lenhart, A., and Madden, M. (2015). Social media update 2014. http://www.pewinternet.org/2015/01/09/social-media-update-2014/.

Duggan, M. and Smith, A. (2013). Social media update 2013. http://www.pewinternet.org/2013/12/30/social-media-update-2013/.

Durkin, P. (2012). Variation in the lexicon: the 'Cinderella'of sociolinguistics?: Why does variation in word forms and word meanings present such challenges for empirical research? *English Today*, 28(4):3–9.

Eisenstein, J. (2013). What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369, Atlanta, Georgia. Association for Computational Linguistics.

Eisenstein, J. (2015). Systematic patterning in phonologically-motivated orthographic variation. *Journal of Sociolinguistics*, 19(2):161–188.

Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E. P. (2014). Diffusion of lexical change in social media. *PloS one*, 9(11):e113114.

Eisenstein, J., Smith, N. A., and Xing, E. P. (2011). Discovering sociolinguistic associations with structured sparsity. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1365–1374. Association for Computational Linguistics.

Fadaee, M., Bisazza, A., and Monz, C. (2017). Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573.

Fiesler, C. and Proferes, N. (2018). Participant perceptions of Twitter research ethics. *Social Media and Society*, 4(1).

Florini, S. (2014). Tweets, Tweeps, and Signifyin': Communication and cultural performance on "Black Twitter". *Television & New Media*, 15(3):223–237.

Gerlitz, C. and Rieder, B. (2013). Mining one percent of Twitter: Collections, baselines, sampling. *M/C Journal*, 16(2). Retrieved from: http://journal.media-culture.org.au/index.php/mcjournal/article/view/620. Accessed: 2019-04-27.

Gonçalves, B. and Sánchez, D. (2014). Crowdsourcing dialect characterization through Twitter. *PloS one*, 9(11):e112074.

González-Bailón, S., Borge-Holthoefer, J., Rivero, A., and Moreno, Y. (2011). The dynamics of protest recruitment through an online network. *Scientific reports*, 1:197.

González-Bailón, S., Wang, N., Rivero, A., Borge-Holthoefer, J., and Moreno, Y. (2014). Assessing the bias in samples of large online networks. *Social Networks*, 38:16–27.

Görlach, M. (1985). Introduction. In Görlach, M., editor, *Focus on: Scotland (Varieties of English around the world, V.5)*, pages 3–5. Amsterdam/Philadelphia, John Benjamins Publishing Company.

Gouws, S., Hovy, D., and Metzler, D. (2011). Unsupervised mining of lexical variants from noisy text. In *Proceedings of the First workshop on Unsupervised Learning in NLP*, pages 82–90. Association for Computational Linguistics.

Grant, W. (1931). Phonetic description of Scottish language and dialects. In *The Scottish National Dictionary*, volume 1, pages 9–41. Online: http://www.dsl.ac.uk/about-scots/history-of-scots/.

Grant, W. and Murison, D. D. (1931). *The Scottish National Dictionary*. Scottish National Dictionary Association.

Greenwood, S., Perrin, A., and Duggan, M. (2016). Social media update 2016. http://www.pewinternet.org/2016/11/11/social-media-update-2016/.

Grieve, J. (2015). Dialect variation. In Biber, D. and Reppen, R., editors, *The Cambridge Handbook of English Corpus Linguistics*, pages 362–380. Cambridge University Press.

Grieve, J., Nini, A., and Guo, D. (2018). Mapping lexical innovation on american social media. *Journal of English Linguistics*, 46(4):293–319.

Hanna, A., Sayre, B., Bode, L., Yang, J., and Shah, D. (2011). Mapping the political twitterverse: Candidates and their followers in the midterms. In *Fifth International AAAI Conference on Weblogs and Social Media*.

Hanna, A., Wells, C., Maurer, P., Friedland, L., Shah, D., and Matthes, J. (2013). Partisan alignments and political polarization online: A computational approach to understanding the french and us presidential elections. In *Proceedings of the 2nd Workshop on Politics, Elections and Data*, pages 15–22.

Honeycutt, C. and Herring, S. C. (2009). Beyond microblogging: Conversation and collaboration via twitter. *42nd Hawaii International Conference on System Sciences*, pages 1–10.

Hovy, D. (2015). Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 752–762. Association for Computational Linguistics.

Hovy, D., Johannsen, A., and Søgaard, A. (2015). User review sites as a resource for large-scale sociolinguistic studies. In *Proceedings of the 24th International Conference on World Wide Web - WWW '15*, pages 452–461.

Hovy, D. and Purschke, C. (2018). Capturing regional variation with distributed place representations and geographic retrofitting. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4383–4394, Brussels, Belgium. Association for Computational Linguistics.

Huang, Y., Guo, D., Kasakoff, A., and Grieve, J. (2016). Understanding U.S. regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems*, 59:244 – 255.

Huberty, M. (2015). Can we vote with our tweet? On the perennial difficulty of election forecasting with social media. *International Journal of Forecasting*, 31(3):992–1007.

Irvine, J. T. (2002). *"Style" as distinctiveness: the culture and ideology of linguistic differentiation*, page 21–43. Cambridge University Press.

Johnson, D. E. (2009). Getting off the goldvarb standard: Introducing rbrul for mixed-effects variable rule analysis. *Language and Linguistics Compass*, 3(1):359–383.

Jones, T. (2015). Toward a description of African American Vernacular English dialect regions using "Black Twitter". *American Speech*, 90(4):403–440.

Jørgensen, A., Hovy, D., and Søgaard, A. (2015). Challenges of studying and processing dialects in social media. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 9–18. Association for Computational Linguistics.

Jungherr, A. (2016). Twitter use in election campaigns: A systematic literature review. *Journal of Information Technology & Politics*, 13(1):72–91.

Kay, B. (1988). *Scots: The Mither Tongue*. Grafton.

Kriz, R., Miltsakaki, E., Apidianaki, M., and Callison-Burch, C. (2018). Simplification using paraphrases and context-based lexical substitution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 207–217, New Orleans, Louisiana. Association for Computational Linguistics.

Kulkarni, V., Perozzi, B., and Skiena, S. (2016). Freshman or fresher? quantifying the geographic variation of language in online social media. In *ICWSM*, pages 615–618. AAAI Press.

Labov, W. (1978a). *The Social Motivation of a Sound Change*, chapter 1, pages 1–41. Basil Blackwell.

Labov, W. (1978b). *The Social Stratification of (r) in New York City Department Stores*, chapter 2, pages 43–69. Basil Blackwell.

Lau, J. H. and Baldwin, T. (2016). An empirical evaluation of doc2vec with practical insights into document embedding generation. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 78–86, Berlin, Germany. Association for Computational Linguistics.

Le, Q. V. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *ICML*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 1188–1196. JMLR.org.

Lee, J. and Yeung, C. Y. (2018). Personalizing lexical simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 224–232, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Lietz, H., Wagner, C., Bleier, A., and Strohmaier, M. (2014). When politicians talk: Assessing online conversational practices of political parties on Twitter. In *Eighth International AAAI Conference on Weblogs and Social Media*.

Linck, J. A. and Cunnings, I. (2015). The utility and application of mixed-effects models in second language research. *Language Learning*, 65(S1):185–207.

Livingston, E. (2019). Yer da sells Avon: how the Scots language found a new home on twitter. *The Face*. Retrieved from: https://theface.com/society/

yer-da-sells-avon-how-the-scots-language-found-a-new-home-on-twitter. Accessed: 2019-10-24.

Macafee, C. (2003). Studying Scots vocabulary. In Corbett, J., McClure, J. D., and Stuart-Smith, J., editors, *The Edinburgh Companion to Scots*, pages 50–71. Edinburgh, Edinburgh University Press.

Marwick, A. E. and boyd, d. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New media & society*, 13(1):114–133.

McAnulla, S. and Crines, A. (2017). The rhetoric of Alex Salmond and the 2014 Scottish independence referendum. *British politics*, 12(4):473–491.

McCarthy, D. and Navigli, R. (2007). Semeval-2007 task 10: English lexical substitution task. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 48–53. Association for Computational Linguistics.

Melamud, O., Levy, O., and Dagan, I. (2015). A simple word embedding model for lexical substitution. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 1–7, Denver, Colorado. Association for Computational Linguistics.

Mihalcea, R., Sinha, R., and McCarthy, D. (2010). Semeval-2010 task 2: Cross-lingual lexical substitution. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 9–14. Association for Computational Linguistics.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mishra, P., Del Tredici, M., Yannakoudakis, H., and Shutova, E. (2019). Author profiling for hate speech detection. *arXiv preprint arXiv:1902.06734*.

Morstatter, F., Pfeffer, J., and Liu, H. (2014). When is it biased?: assessing the representativeness of Twitter's streaming API. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 555–556. ACM.

Morstatter, F., Pfeffer, J., Liu, H., and Carley, K. M. (2013). Is the sample good enough? Comparing data from Twitter's Streaming API with Twitter's Firehose. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, pages 400–408.

Muller, B., Sagot, B., and Seddah, D. (2019). Enhancing BERT for lexical normalization. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 297–306, Hong Kong, China. Association for Computational Linguistics.

Mycock, A. (2012). SNP, identity and citizenship: Re-imagining state and nation. *National Identities*, 14(1):53–69.

Nguyen, D., Trieschnigg, D., and Cornips, L. (2015). Audience and the use of minority languages on Twitter. In *Proceedings of the Ninth International Conference on Web and Social Media*, pages 666–669.

Paetzold, G. and Specia, L. (2016). SemEval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569, San Diego, California. Association for Computational Linguistics.

Paetzold, G. H. and Specia, L. (2017). A survey on lexical simplification. *Journal of Artificial Intelligence Research*, 60:549–593.

Pavalanathan, U. and Eisenstein, J. (2015a). Audience-modulated variation in online social media. *American Speech*, 90(2):187–213.

Pavalanathan, U. and Eisenstein, J. (2015b). Confounds and consequences in geotagged Twitter data. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2138–2148. Association for Computational Linguistics.

Peirsman, Y., Geeraerts, D., and Speelman, D. (2010). The automatic identification of lexical variation between language varieties. *Natural Language Engineering*, 16(4):469–491.

Purves, D. (2002). *A Scots grammar: Scots grammar and usage*. The Saltire Society.

Radovanovic, D. and Ragnedda, M. (2012). Small talk in the digital age: Making sense of phatic posts. In *2nd Workshop on Making Sense of Microposts*, pages 59–68. Association for Computational Linguistics.

Rauschnabel, P. A., Sheldon, P., and Herzfeldt, E. (2019). What motivates users to hashtag on social media? *Psychology & Marketing*, 36(5):473–488.

Robinson, S. (2019). 19 times Scottish Twitter was the best Twitter. *BuzzFeed*. Retrieved from: https://www.buzzfeed.com/sydrobinson1/funny-scottish-tweets/. Accessed: 2019-10-24.

Romero, D. M., Tan, C., and Ugander, J. (2013). On the interplay between social and topical structure. In *Seventh International AAAI Conference on Weblogs and Social Media*.

Ruette, T., Speelman, D., and Geeraerts, D. (2014). Lexical variation in aggregate perspective. *Pluricentricity: Language variation and sociocognitive dimensions*, pages 103–126.

Russell, J. (2019). Scottish Twitter hilariously reacts to exam results. *The Daily Record*. Retrieved from: https://www.dailyrecord.co.uk/news/scottish-news/scottish-twitter-hilariously-reacts-exam-18839116. Accessed: 2019-10-24.

Sánchez-Rada, J. F. and Iglesias, C. A. (2019). Social context in sentiment analysis: Formal definition, overview of current trends and framework for comparison. *Information Fusion*, 52:344–356.

Schrading, N., Ovesdotter Alm, C., Ptucha, R., and Homan, C. (2015). #WhyIStayed, #WhyILeft: Microblogging to make sense of domestic abuse. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1281–1286, Denver, Colorado. Association for Computational Linguistics.

ScotCen (2013). Should Scotland be an independent country? (combined responses of those who have and those who haven't decided yet) broken down by 'Moreno' national identity. Retrieved from: http://whatscotlandthinks.org/. Accessed: 2016-09-30.

Scots Language Centre (2013). Brief analysis of the 2011 census results. Retrieved from: http://media.scotslanguage.com/library/document/SLC%20Analysis%20of%20Census%202011%20for%20Scots.pdf. Accessed: 2019-02-05.

Scottish Government (2010). Public attitudes towards the Scots language.

Scottish Language Dictionaries (2004). Dictionary of the Scots language. http://www.dsl.ac.uk/. Accessed: 2016-12-20.

Shoemark, P., Ferdousi Liza, F., Nguyen, D., Hale, S. A., and McGillivray, B. (2019). Room to glo: A systematic comparison of semantic change detection approaches with word embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. to appear.

Sloan, L. (2017). Who tweets in the united kingdom? Profiling the Twitter population using the British Social Attitudes Survey 2015. *Social Media + Society*, 3(1).

Sloan, L. and Morgan, J. (2015). Who tweets with their location? Understanding the relationship between demographic characteristics and the use of geoservices and geotagging on Twitter. *PloS one*, 10(11):e0142209.

Sobhani, P., Inkpen, D., and Zhu, X. (2017). A dataset for multi-target stance detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 551–557, Valencia, Spain. Association for Computational Linguistics.

Soule, D. P., Leith, M. S., and Steven, M. (2012). Scottish devolution and national identity. *National Identities*, 14(1):1–10.

Specia, L., Jauhar, S. K., and Mihalcea, R. (2012). Semeval-2012 task 1: English lexical simplification. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 347–355. Association for Computational Linguistics.

Stewart, I., Pinter, Y., and Eisenstein, J. (2018). Si o no, que penses? Catalonian independence and linguistic identity on social media. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 136–141. Association for Computational Linguistics.

Stuart-Smith, J. (2003). The phonology of modern urban Scots. In Corbett, J., McClure, J. D., and Stuart-Smith, J., editors, *The Edinburgh Companion to Scots*, pages 110–137. Edinburgh, Edinburgh University Press.

Stuart-Smith, J., Timmins, C., and Tweedie, F. (2007). 'Talkin' Jockney'? Variation and change in Glaswegian accent1. *Journal of Sociolinguistics*, 11(2):221–260.

Tagliamonte, S. A. (2006). *Analysing sociolinguistic variation*. Cambridge University Press.

Tatman, R. (2015). #go awn: Sociophonetic variation in variant spellings on Twitter. *Working Papers of the Linguistics Circle of the University of Victoria*, 25(2):97–108.

Tromble, R., Storz, A., and Stockmann, D. (2017). We don't know what we don't know: When and how the use of Twitter's public APIs biases scientific inference. In *Working Papers on SSRN*, pages 1–26.

Uri, A. (2018). The Scots leid is for aa, nae jist for nationalists. *The National*. Retrieved from: https://www.thenational.scot/news/17239284.the-scots-leid-is-for-aa-nae-jist-for-nationalists/. Accessed: 2019-10-24.

Weiner, E. J. and Labov, W. (1983). Constraints on the agentless passive. *Journal of linguistics*, 19(1):29–58.

Williams, M. L., Burnap, P., and Sloan, L. (2017). Towards an ethical framework for publishing Twitter data in social research: Taking into account users' views, online context and algorithmic estimation. *Sociology*, 51(6):1149–1168.

Wood-Doughty, Z., Smith, M., Broniatowski, D., and Dredze, M. (2017). How does Twitter user behavior vary across demographic groups? In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 83–89. Association for Computational Linguistics.

Yimam, S. M., Biemann, C., Malmasi, S., Paetzold, G., Specia, L., Štajner, S., Tack, A., and Zampieri, M. (2018). A report on the complex word identification shared task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78, New Orleans, Louisiana. Association for Computational Linguistics.