Nini, A., Bailey, G., Guo, D., Grieve, J. (2020) The graphical representation of phonetic dialect features of the North of England on social media. In Honeybone, P. & Maguire, W. (eds), *Dialect Writing and the North of England*, 266-296, Edinburgh: Edinburgh University Press

The graphical representation of phonetic dialect features of the North of England on social media

Nini, A., Bailey, G., Guo, D., Grieve, J.

12.1. Introduction

This chapter offers a complementary perspective to the subject of the book by looking at social media and on whether and how the graphological reflection of dialect writing is affecting these new forms of communication. By foregrounding the importance of the ways in which speakers construct and project personae (Eckert, 2012), third wave theoretical approaches to the study of linguistic variation would predict that users can break orthographic conventions in order to encode their dialect and linguistic identity on social media. However, the extent to which users of social media use spelling resources to convey dialectal identities and to what purposes is not immediately obvious and one of the objectives of this chapter is to shed light on this phenomenon both quantitatively and qualitatively.

Recent research on dialect variation using social media data has so far provided evidence that spelling variants that reflect phonetic dialect s are found in social media posts, such as tweets. This is an important finding because it opens the possibility of analysing the dialect of a region using naturally occurring social media posts as opposed to using interviews or questionnaires.

In addition, if users do adopt dialectal spelling variants in their social media communications, it is not clear whether their geographical patterning would match their respective phonetic forms. The question of the extent to which social media like Twitter can be used to answer questions about language variation and change is still an open one, even though the evidence so far is particularly promising. The other objective of this study is therefore to attempt to derive dialect patterns from social media data on the basis of the frequency with which dialectal spelling is used and estimate the degree of match of these dialect patterns with established knowledge of their phonetic equivalents gathered using survey methods. For these reasons, the analysis will focus on a sub-set of features of the dialects of the north of England that can potentially be expressed using spelling variation.

In this study, using a corpus of 183 million geo-coded tweets totalling 1.8 billion words, we explore how phonetic features of the dialects of the North of England such as HAPPY-laxing (e.g. happy > happeh; funny > funneh) and the retention of UW in MOUTH are realised graphically by social media users. We present results that show that the geographical distribution of these features as found on Twitter is similar to the one attested from other studies carried out with other more conventional methods. Furthermore, our research reveals how and how often these dialect features are used in written online communication, adding to our understanding of the relationship between language and the projection of identity.

In the following sections, we firstly provide an overview of previous dialectological studies that have used social media data and, secondly, a brief description of Northern English dialectology with an outline of the specific features under consideration in this study.

12.2. Social media data for dialectology

Data for corpus and computational linguistics has been collected using the Internet for more than a decade and, despite theoretical issues concerning representativeness, has so far led to interesting and useful results for various corpus linguistics endeavours (Kilgarriff and Grefenstette, 2003). Among all the types of Internet data, though, the one type that started a revolution is social media data, which is responsible for the emergence of the new field of *computational sociolinguistics* (Nguyen *et al.* 2016). Social media data offers several advantages to a sociolinguist or dialectologist compared to survey data sets such as, above all, the substantially large size and the fact that it is 'observer paradox' free. Conversely, however, social media data comes with the problem of biased sampling and with its 'bad language' (Eisenstein, 2013) in terms of its adherence to standard variety, which is however not necessarily a drawback, as many studies on the emergence and spread of innovations based on social media data have demonstrated. The drawbacks of using social media data are, however, small compared to its benefits: this type of data allowed research of unprecedented scale that could finally tackle some research questions or tasks that would otherwise be impossible, such as identifying neologisms as they occur (Kerremans, Stegmayr and Schmid, 2012) or studying how words emerge and become popular (Grieve, Nini and Guo, 2017).

Dialectology is probably the branch of linguistics that can benefit the most from social media data such as Twitter data because of its availability with geocoded information. Users of social media can choose to add the GPS coordinates of the place they sent their message from and thus corpora of geo-coded tweets can be used to directly observe dialectal variation almost in real time. Dialectological studies using Twitter data have already been carried out and, because of the ease with which it is possible to extract lexical items, geo-coded Twitter data has been mostly explored in terms of lexical variation (Eisenstein et al., 2012; Gonçalves and Sánchez, 2014; Kulkarni, Perozzi and Skiena, 2016; Shoemark, Kirby and Goldwater, 2017).

Despite the latest achievements of the studies that used geo-coded Twitter data, Eisenstein (2018) stresses that important limitations in terms of representativeness can be a danger, since arguably the population of users of Twitter is not necessarily a random sample of the general population of a country. Studies carried out in the UK indeed confirm that the population of Twitter users is not generally representative of the UK population as a whole. For example, Longley, Adnan and Lansley (2015) inferred demographic information for each user in a corpus of 4 million geo-tagged tweets sent in London using their usernames and concluded that the population of Twitter users is biased towards men and younger adults compared to 2011 Census data. Results about ethnicity indicate that all ethnic groups beside White are underrepresented, although the confidence over these results is unclear. Additional evidence comes from Mellon and Prosser (2017), who carried out a study using the 2015 British Election Study survey and concluded that, compared to a random and representative sample of eligible British voters, users of Twitter are younger (mean age of 34 compared to 48), more likely to be men, more likely to have A-levels or a degree, and more liberal in terms of political views than non-users.

In addition to these biases of Twitter users in general, there seem to exist an additional bias in the population of users who use geo-coded posts. Pavalanathan and Eisenstein (2015) studied the possible bias in geotagged US Twitter data and uncovered that geolocation is preferred by younger people and women. Interestingly,

they find that the users who choose to geo-tag their tweets are also the ones that are more likely to use non-standard regional words.

Although the evidence therefore points to a substantial disparity between the population of Twitter users and geocoded Twitter users compared to the population of a country as a whole, this by itself does not mean that dialectological work carried out on Twitter is necessarily invalid, and, in spite of the representativeness shortcomings, so far research has managed to successfully replicate dialect studies carried out with more conventional methods (Cook, Han and Baldwin, 2014; Doyle, 2014; Eisenstein, 2015; Jones, 2015; Huang et al., 2016; Rahimi, Cohn and Baldwin, 2017). Recently, Grieve *et al.* (2019) performed a comparison of 139 dialect maps from the BBC Voices survey to the equivalent maps generated using Twitter data and found a high degree of alignment which further validates the use of Twitter for dialectology.

An area that has so far showed promising advances is the study of graphological reflections of dialectal phonetic variation, with a few studies so far showing that graphological representations are consistent with the same variation present in speech. The act of encoding phonetic features in orthographic forms long pre-dates the advent of social media platforms and has been widely documented, for example, in nineteenth century poetry or in 'dialect literature' (see Asprey; Braber, this volume). As highlighted by the other chapters in this volume, one of the defining features of dialect writing is not simply the use of specific lexical items but also the creative way in which writers flout orthographic norms in order to reflect their spoken accent.

Parallels can be drawn between this traditional style of dialect writing and the contemporary stylistic practices of users on social media platforms such as Twitter. The open nature of Twitter data makes it a fruitful area of linguistic research, and there have been a number of existing studies that have explored the relationship between phonetic features in speech production and the way they are reflected orthographically on social media. Analysing a corpus of over 100 million geotagged Tweets, Eisenstein (2015) compares the variable use of *g-dropping* (e.g. <walkin> for *walking*) and *td-deletion* (e.g. <jus> for *just*) on Twitter with their pattern of variation in speech, finding similarities in the way they are conditioned between these two mediums. Specifically, the widely-established 'nominal-verbal continuum' that sees verbs favour [ɪn] and nouns [ɪŋ] (Labov, 1989) is reflected in the orthographic variation, as is the phonological conditioning of *td-deletion* in which deletion is favoured before consonant-initial words but inhibited pre-vocalically (e.g. Guy, 1991; Tagliamonte and Temple, 2005; Tanner, Sonderegger and Wagner, 2017). Although the grammatical conditioning of *g-dropping* was not replicated in UK Twitter data by Bailey (2016), non-standard <in> and <in'> spellings were found to occur at higher frequencies in northern England and Scotland, reflecting the same regional patterning that has been reported for the phonological alternation (Labov, 2001: 90).

In a separate line of work, Tatman (2016) investigated how sociophonetic features of New York City English – such as /ɹ/-deletion (e.g. <beah> for *beer*) and certain vocalic differences (e.g. <woyld> for *wild* and <nawt> for *not*) – are encoded on Twitter in impersonations of this dialect, arguing that the salience of these features modulates the extent to which they are employed in the projection of dialect on Twitter.

Despite these recent advances in our understanding of how phonetically-motivated spelling is utilised by users of social media, the extent to which varieties of Northern English are reflected in this way remains an unexplored avenue of research.

12.3. The north of England

Broadly speaking, the 'linguistic north' of England has been defined as the region where the vowels in FOOT and STRUT, and in TRAP and BATH, are pronounced the same; these are two well-known and highly-salient features that characterise the linguistic north/south divide in England (Wells, 1982). The isoglosses for these two vocalic features run approximately from the River Severn to the Wash, and as such they include the Midlands regions and the dialects spoken therein; in this chapter we focus primarily on those varieties spoken in the North West (e.g. Manchester, Merseyside, Lancashire etc.), the North East (e.g. Tyneside and Wearside), and Yorkshire, excluding Birmingham and Black Country varieties spoken in the Midlands.

There is a long history of dialectology in the north of England (e.g. Wakelin, 1977; Wells, 1982; Beal, 2004; Hughes, Trudgill and Watt, 2012). Additionally, contemporary variationist linguistics has often placed the lens of inquiry on northern dialects such as those spoken in Manchester (Drummond, 2012; Baranowski and Turton, 2015; Baranowski, 2017), Liverpool (Honeybone, 2007; Watson, 2007; Cardoso, 2015), Tyneside (Milroy et al., 1994; Watt, 2002; Beal, Burbano-Elizondo and LLamas, 2012), and Yorkshire (Petyt, 1985; Tagliamonte, 2004).

Dialectal variation in the North of England is a fruitful area of research given the extreme linguistic diversity we find here; it is said that differences between dialects are "sharper in the north than in any other part of England" (Wells, 1982: 351), and this statement has been echoed by Beal (2004: 120) who claims that more features differentiate northern dialects from each other than are common to all of them. For example, even within the North East alone there are significant differences between Newcastle and Sunderland varieties despite their close proximity to each other (Beal, 2000; Beal, Burbano-Elizondo and LLamas, 2012). Trudgill (1999) also notes how traditional dialect speakers can still be found in the north of England, particularly in rural areas and more endocentric, geographically-peripheral communities. The wealth of linguistic diversity in the North could in part stem from geography and its role in inhibiting contact-induced levelling (cf. southern areas such as the Fens that are conducive to dialect contact, as explored by Britain (2002)) and, on a related note, how its distance from London more generally inhibits the influence of the standard on the more traditional varieties spoken in the North. In the Old English period, regional norms in the written language had already begun to develop from different scribal centres across the country, laying the foundation for later regional diversity (see Beal (2004): §1.2 for a brief overview of the history and development of Northern English). In this chapter, we show that variation in written English is still present, at least in the case of phonetically-motivated orthography on social media.

12.3.1. Northern dialect features

In selecting variable phonetic features to be included in this study, there are two important pre-requisites to be considered in addition to regional stratification: (1) they must be perceptually and socially salient enough to be used orthographically as an index of local dialects, and (2) they must plausibly be encoded in orthographic representations. As such, we cannot investigate northern features such as the presence of dark /l/ variants (Carter, 2002; Turton, 2017) or post-nasal [g]-presence in words such as *sing* or *wrong* (Wells, 1982: 365); both features have relatively low social

profiles (see Bailey, 2019) and it is not clear how either could be reflected transparently in the orthographic representation.

Eleven features have been selected for analysis, covering both consonantal and vocalic variation; all of these features have been said to occur in the north of England, with a subset of these features occurring exclusively in the North and many others said to occur more frequently in these dialects relative to the other regions of England. The features are listed below, alongside descriptions and examples of their phonetic and orthographic realisation.

1. **T to R**: In some northern varieties, a /t/ final in a monosyllabic word can be realised as [ɹ] if followed by a vowel-initial word, e.g. *get off* [gɛɹɒf]; this is typically associated with Liverpool English (Watson, 2007; Buchstaller et al., 2013; Honeybone, Watson and van Eyndhoven, 2017), but is also present in other northern dialects such as Tyneside English (Watt and Milroy, 1999). However, there have been reports suggesting that this feature is stigmatised and receding over time (Foulkes and Docherty, 2007), and that its occurrence is restricted to highly-frequent collocations (Clark and Watson, 2011). e.g. *get off, lot of* → *geroff, lorra*
2. **HAPPY-laxing**: The HAPPY lexical set refers to the word-final unstressed /i/ vowel, which in most varieties is realised as either tense [i] or lax [ɪ]; in Manchester English, however, a super-lax variant [ɛ] is also possible (Ramsammy and Turton, 2012; Hughes, Trudgill and Watt, 2012). This is highly salient and encoded in stereotypes of supporters of the football club Manchester City, e.g. *citeh*; Braber (this volume) shows evidence of a productive graphical representation of this feature in dialect writing literature of the East Midlands. Although it is also a productive process in speech, it remains to be seen whether or not the orthographic form is restricted to this lexical item as a sociocultural symbol on social media. e.g. *happy* → *happeh*
3. **LETTER-backing**: Like HAPPY-laxing, LETTER-backing is another vocalic feature exclusive to Manchester English, targeting word-final unstressed /ə/. Ramsammy and Turton (2012) find that the phonetic quality of this vowel is approaching [ʌ] rather than [ɒ], but this is nevertheless perceived and stereotyped as the latter, and therefore will likely be reflected orthographically by replacement of <er> with <oh>. This is likely to be enregistered most strongly in the word Manchester itself, as is typically the case with place names. e.g. *Manchester* → *Manchestoh*
4. **AW to UW**: The retention of [uː] in MOUTH is one of the defining features of Tyneside English spoken in the North East (Hughes, Trudgill and Watt, 2012: 155) as well as of Scots (Johnston, 1997; McColl Millar, 2007). It is claimed to be particularly frequent in specific lexical items that enact local identity such as *toon* (*town*) and *broon* (*brown*), referring to Newcastle United Football Club and Newcastle Brown Ale, respectively (Beal, Burbano-Elizondo and LLamas, 2012: 35). As with HAPPY-laxing and LETTER-backing, it is possible that the orthographic representation of this feature may be lexically-restricted. e.g. *down* → *doon*
5. **FOOT-STRUT**: The lack of distinction between the FOOT and STRUT lexical sets, in which both are realised as /ʊ/, is associated with all dialects in the North and, to a lesser extent, the Midlands. Alongside the BATH-TRAP split, it is described as the most important characteristic differentiating northern and southern dialects (Wells, 1982), although there is recent evidence to suggest that the /ʌ/ vowel is spreading northwards (MacKenzie, Bailey and Turton, 2014). In

addition to this, Trudgill (1986) notes an apparent discrepancy in the relative social salience of these two features, stating that although northerners are stereotyped by southerners as lacking both the bath- trap and foot-strut oppositions, northerners themselves only comment on the southern [ɑː] vowel in bath. This is an important point to consider given the afore-mentioned issue of salience and the likelihood of these features being represented orthographically. e.g. *love, London → luv/lav, Landan*

6. **G-dropping**: The variable realisation of unstressed *-ing* clusters as either alveolar [ɪn] or velar [ɪŋ] is widely studied in sociolinguistics and has been attested throughout the British Isles. Although this feature is not restricted to northern dialects in the same way as the features discussed thus far, it has nevertheless been argued that [ɪn] occurs more frequently in the north of England and Scotland (Moore, Meech and Whitehall, 1935; Houston, 1985; Watts, 2005), so much so that Levon and Fox (2014: 201) describe it as a regional (rather than social) variable in the British Isles. A third possible variant, [ɪŋg], *is* unique to the North West and West Midlands, but is unlikely to be reflected orthographically given the existing presence of <g> in the standard spelling. e.g. *walking → walkin, walkin'*
7. **TH-stopping**: Word-initial dental fricatives /θ, ð/ can sometimes be realised as alveolar stops [t, d], e.g. *think* [tɪŋk]. Although this is more strongly associated with the performance of ethnic rather than regional identity and has strong ties to Multicultural London English (see Drummond, 2018), it has at least been attested in northern varieties such as Liverpool English (Watson, 2007) and Manchester English (Drummond, 2018). e.g. *think, this → tink, dis*
8. **TH-fronting**: Not to be confused with the afore-mentioned process of stopping, the dental fricatives /θ, ð/ can also undergo fronting to labiodental position, i.e. [f, v]. This process has been described as one of the fastest spreading sound changes in British English (Trudgill, 1999), but is possibly less frequent in the north given that it originated and subsequently diffused from London and the South East (Williams and Kerswill, 1999). It has been attested in Manchester English, where it is described as an "urban youth norm" (Baranowski and Turton, 2015: 303). e.g. *think, with → fink, wiv*
9. **H-dropping**: Deletion of word-initial /h/, e.g. *house* [aʊs], has been reported in dialects of British English for hundreds of years and is described as "the single most powerful pronunciation shibboleth in England", especially for function words (Wells, 1982: 254); as such, it is not necessarily a northern feature but it has been attested in Manchester English (Baranowski and Turton, 2015) where it has been claimed to be frequent in conversational speech (Hughes, Trudgill and Watt, 2012: 116). e.g. *happens, have → appens, av(e)*
10. **Consonant reduction**: Rather than being a single phonetic process, we use the term consonant reduction to refer to a group of pronunciation variants that are characteristic of conversational speech; there is no reason to believe these are geographically restricted, but they are often found in traditional dialect poetry from regions such as Lancashire. e.g. *doesn't, didn't, isn't, with → dunt, dint, int, wi*
11. **Vowel reduction**: As above, but for general process of vocalic reduction instead. Again, these are characteristic of conversational speech where unstressed vowels are often reduced and centralised to a more [ə]-like quality. e.g. *your, you, I've, our, my → yer, ye/ya, av, ar(e), mi/ma*

12.4. Data

The corpus used in this study was collected at the University of South Carolina using the Twitter API and consists of 183 million geo-coded tweets. The corpus contains a total of 1.8 billion words written by almost 2 million users in the United Kingdom for the year 2014[1]. Using the longitude and latitude of each geo-coded tweet the corpus was divided in sub-corpora, grouping together all tweets from the same postcode area. There are in total 124 postcode areas in the United Kingdom and the data for these areas varied widely, from 54,000 tweets in Outer Hebrides to 5.5 million in Manchester, which is the largest area because London is sub-divided in smaller areas. We decided not to alter the corpus in any way, as for example by filtering it for retweets or for tweets sent by bots as done in some previous studies. This decision was taken to guarantee that the sample analysed is a representative data set that replicates what a typical user would encounter in the real world, while the size of the data set allows for the geographical signal to be captured even despite the noise that might be given by these factors.

12.5. Methodology

To find the eleven features considered for this study we compiled a list of words belonging to each category ordered by frequency, with measures taken from the SUBTLEX-UK corpus of television subtitles and operationalised using the Zipf scale (van Heuven et al., 2014). For each of these words we replaced the standard spelling with the predicted spelling for its dialectal variant. We then filtered these lists in two ways, firstly by frequency and secondly by word search feasibility. We kept all words for each category with Zipf ≥ 5, since the most frequent types account for the majority of the tokens in the corpus, as per Zipf's law[2]. After this frequency filtering, we manually scrutinised the word lists and eliminated words that could be problematic when searched automatically. For example, the 'AW to UW' word *now* would be spelled in its dialectal form as *noo* but searching for this form would be problematic as there would be considerable noise coming from standard *no* with an elongated vowel representation. The descriptive statistics of the features in terms of their problematic words are reported in Table 1.

[1] No data was collected in two days in June and four days October for technical reasons

[2] The Zipf scale ranges from 1 to 7 and a value ≥ 4 indicates that the word is a high-frequency word. A value of 5 roughly corresponds to 100 occurrences per million words

	Number of words (≥ 5 zipf)	% of problematic words	Example of problematic words
T to R	6	0	-
HAPPY-laxing	60	0	-
LETTER-backing	2	0	-
AW to UW	25	20	*now → noo*
FOOT-STRUT	8	12.5	*done → dan*
G-dropping	40	2.5	*being → bein*
TH-stopping	54	22	*than → dan*
TH-fronting	54	74	*than → van*
H-dropping	44	41	*his → is*
Consonant reduction	6	33	*wasn't → want*
Vowel reduction	12	25	*I → a*
Average	28.27	20.91	

Table 1 – List of features including the number of words per feature and the percentage of these words that were judged to be problematic with examples.

Despite problems with certain categories, such as TH-fronting or H-dropping, the majority of the spelling replacements were not highly problematic in terms of being confused with other existing standard or non-standard words of English.

Each word that was judged not to be problematic was then searched in the corpus and the relative frequency of each word was calculated by normalising the number of occurrences of the nonstandard variant by the sum of the occurrences of the nonstandard variant plus the standard variant. The formula to calculate the relative frequency was therefore as follows, where $rf_{w,a}$ is the relative frequency of the word *w* in the area *a*, $f_{n,a}$ is the number of times the non-standard variant of the word occurs in the area *a* and $f_{s,a}$ is the number of times the corresponding standard variant occurs in the area *a.*

$$rf_{w,a} = \frac{f_{n,a}}{f_{s,a} + f_{n,a}}$$

The overall relative frequency of the feature across all words for each area, $rf_{f,a}$, was calculated as follows

$$rf_{f,a} = \frac{\sum f_{n,a}}{\sum (f_{s,a} + f_{n,a})}$$

Each array of frequencies was then mapped to represent the geographical distribution of both the single words and the features. Using this method, all the words belonging to a feature contribute equally to the calculation of the overall relative frequency of the feature. A full list of the words that were searched in the corpus can be found in the appendix.

12.6. Results

The analysis resulted in various findings, some of which confirm previously established patterns of phonetic variations while others are harder to interpret. This section outlines the results for each feature leaving the interpretation of the general

patterns to the discussion section. A general trend that can be noticed across the maps is that the relative frequencies of the nonstandard variants and of the features are relatively small. Despite this limitation, for most of them clear geographical patterns can be detected and this suggests that the geographical signal contained in these frequencies is also relatively strong.

The results for the 'T to R' feature are consistent with the expectations given by what is known from dialectology studies carried out using conventional methods.

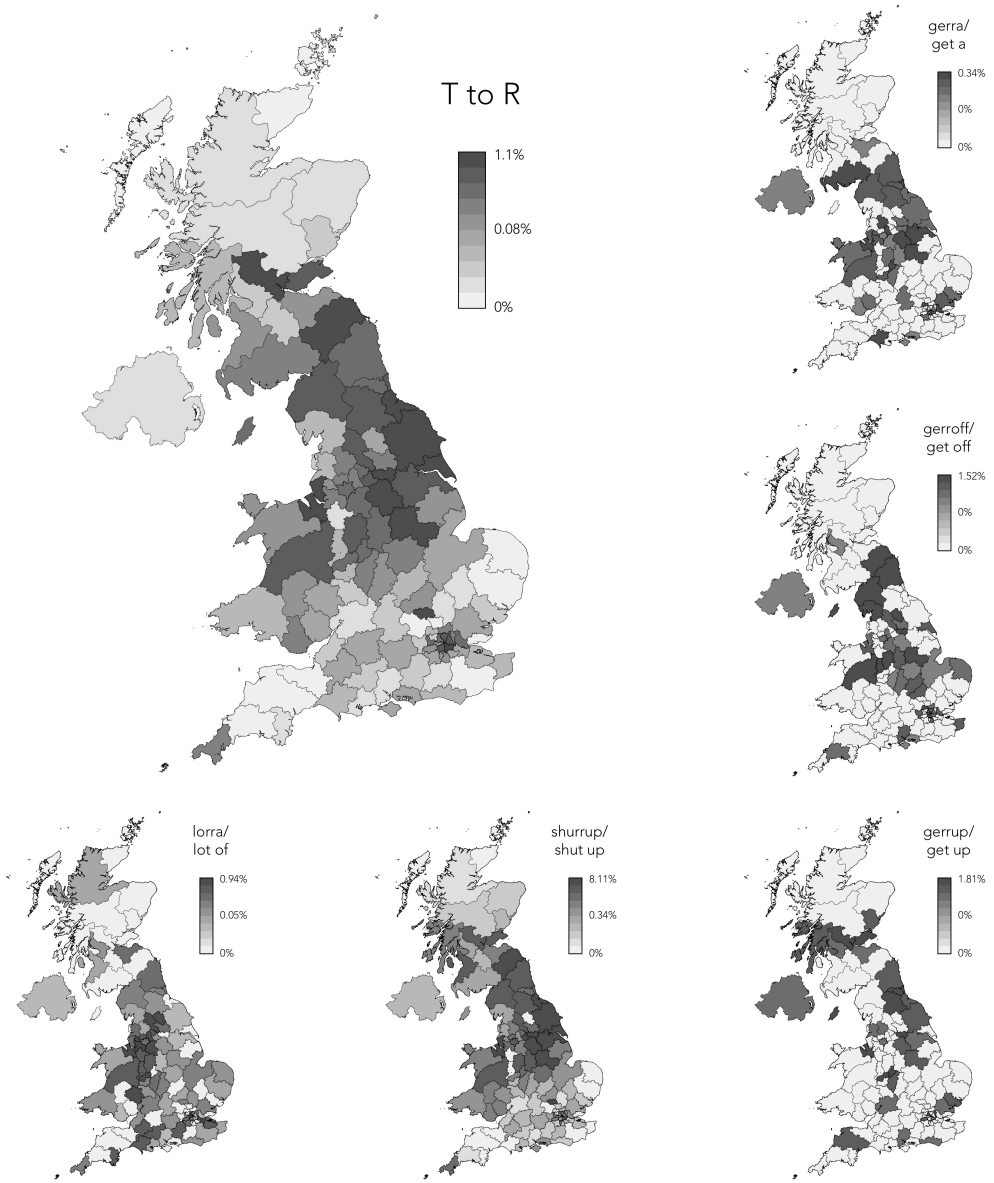


Figure 1 shows that overall the non-standard spellings of these 'T to R' words are mostly in the north of England (top left map), with some words more common in Tyneside (*gerra, gerrup*), as in the examples below[3]:

(1) Some people walk so stupid that it actually annoys me. **Gerra** grip ya idiots
(2) Time to **gerrup** and work out before the derby.

[3] The Twitter examples in this book chapter were slightly altered to protect the anonymity of the users. A few words in the tweets were replaced with synonymous ones so that the tweets cannot be easily traced back to their users.

alternatively, other words are more common in the North West (*gerroff, lorra*):

(3) pub quiz is tomorrow night. Come & have a **lorra lorra** laughs and get your thinking caps on
(4) so much for revising ya liar **gerroff** twitter

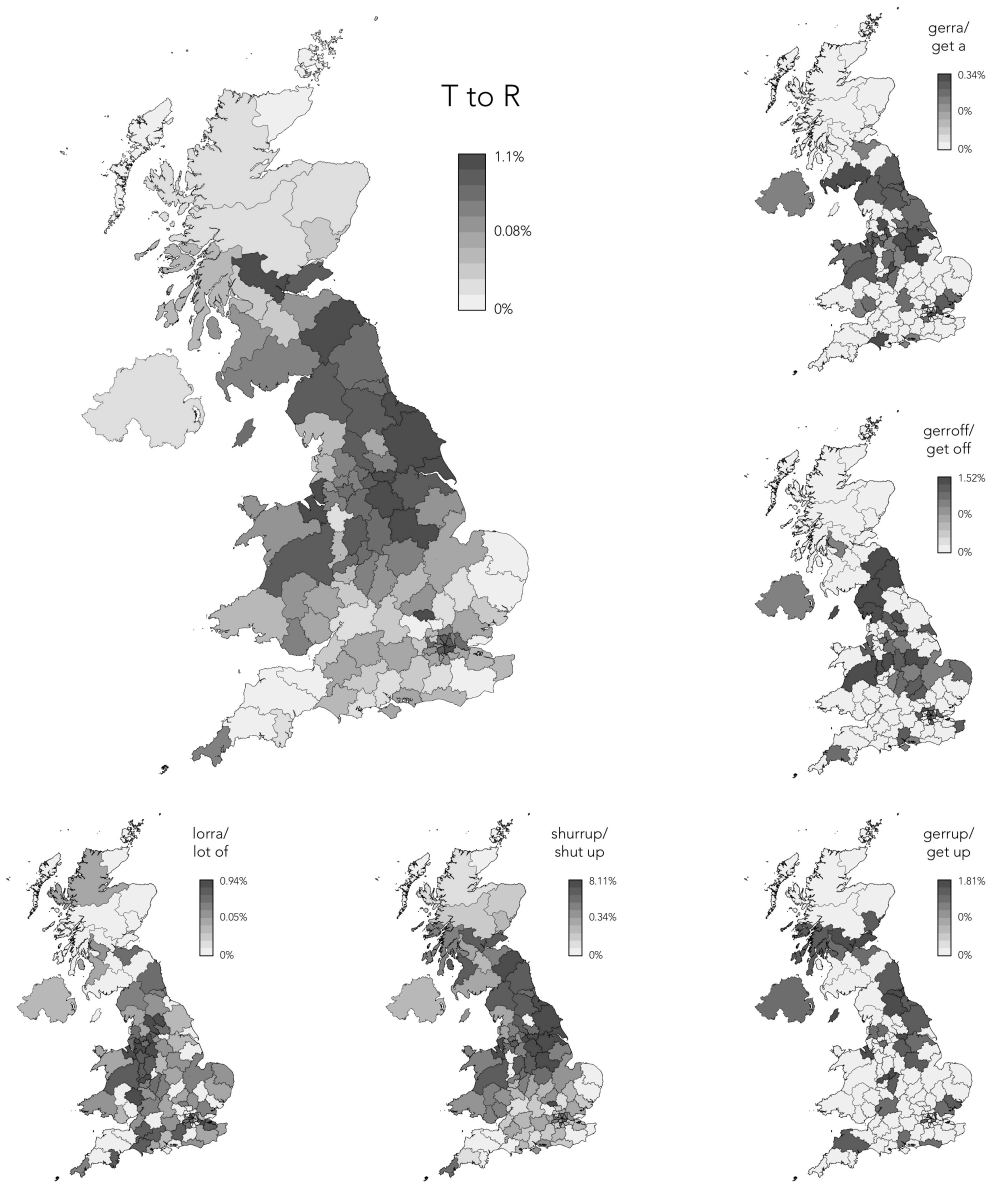


Figure 1 - Maps for the 'T to R' feature and for a sample of its words (*gerra, gerroff, gerrup, lorra, shurrup*)

Similarly, the results of the analysis of HAPPY-laxing confirm that this phenomenon is reflected in spelling in areas consistent with their phonetic equivalent. In addition, both the quantitative results and the manual exploration of the tweets, such as the examples below, suggest that the graphological variation is somewhat productive. Figure 2 shows that not only *citeh* but also *babeh*, *funneh*, or *happeh* are relatively more frequent in the north and even more in the North West.

(5) Don't care if it's pre-season, I'd like to see the red men smash **Citeh** tonight!

(6) oh yes so **happeh** to be eating chicken nugs
(7) Sorry we won't be having one of these tomorrow happy valentines **babeh**

Figure 2 - Maps for the HAPPY-laxing feature and for a sample of its words (*babeh, citeh, funneh, happeh, sorreh*)

Figure 3 - Maps for the ER lowering feature and for all of its words (*manchestah, manchestoh*)

Although the two words considered for ER-backing, *manchestah* and *manchestoh*, are clearly geographically marked, the maps for this feature in Figure 3 reveal that there is no uniform or clear spatial pattern in their use, a result probably due to the very low frequency of this feature. A qualitative exploration of a random sample of the tweets however suggests that there was no error in the analysis and that Twitter users do adopt these spellings to refer to Manchester

(8) that's why I'm on the vip list from Miami too **Manchestoh**
(9) he's the biggest FOOL in **Manchestah**
(10) tour tickets booked for **MANCHESTAH**. I'm a happy man!

The sparseness of the frequency of this feature could be attributed to the fact that it is a stereotypical form that can be used by all speakers, regardless of dialect, in order to imitate a Mancunian accent.

For the retention of UW in MOUTH, the maps resulting from the Twitter corpus analysis represented in Figure 4 indicate that its equivalent graphological phenomenon is distributed similarly to the phonetic variation, if not even more extensively. In addition to the expected focus in Tyneside, the words considered are also used very much in Scotland and, to a lesser extent, other parts of the United Kingdom.

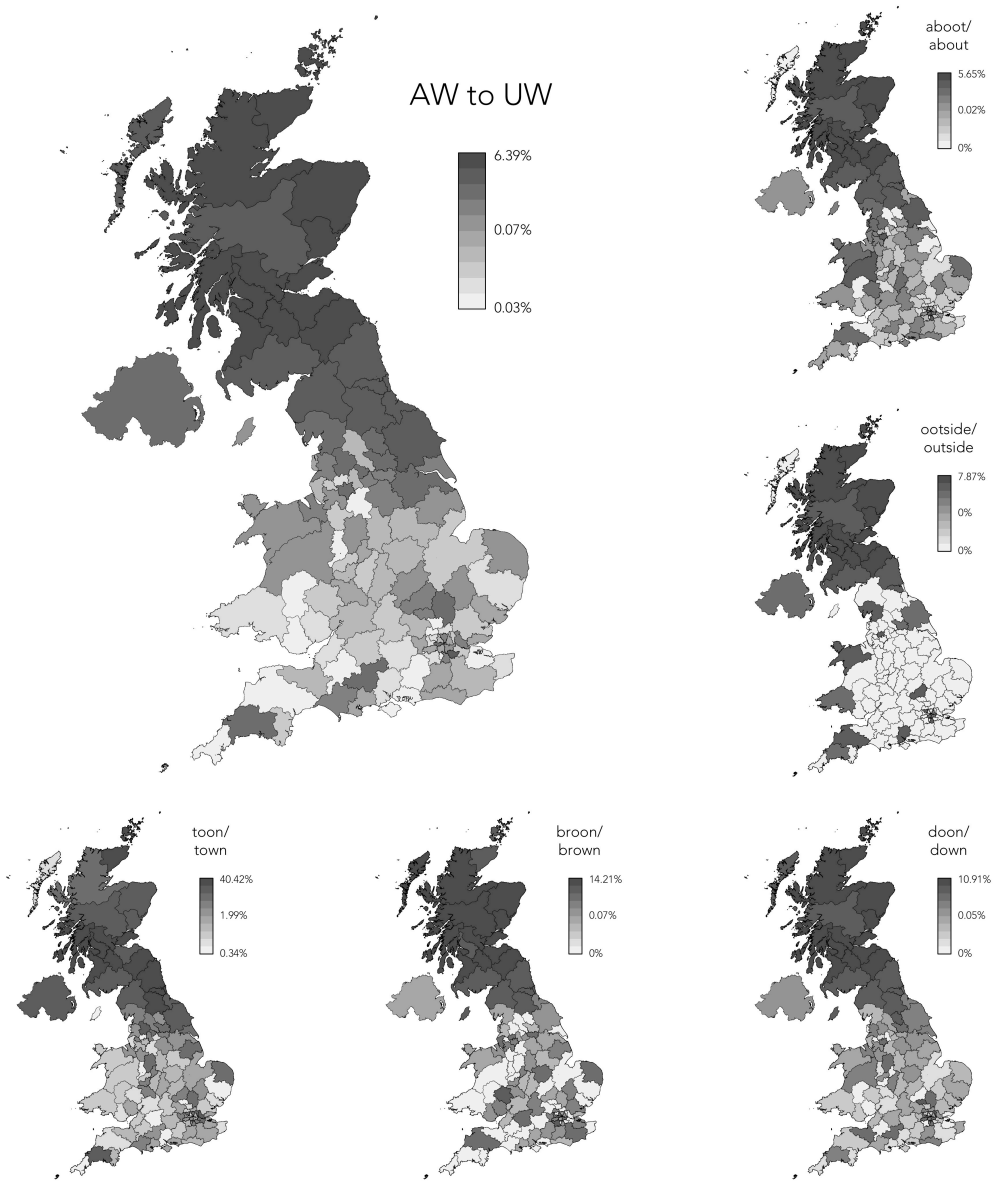


Figure 4 - Maps for the AW to UW feature and for a selection of its words (*aboot, ootside, doon, toon, broon*)

As predicted, some words are more widely adopted even outside of the area of origin of the phonetic variation, due to their general popularity as cultural stereotypes of this region (e.g. *broon, toon*). When the feature is used in these popular forms, it is very likely that the Twitter users are consciously portraying a northern identity, as in the following examples:

(11) I miss you too and the **doon toon** bantaaa
(12) How dare the **toon** lose another match
(13) Can't believe they ran out of **broon** ale last night

However, certain users do not restrict themselves to the most popular forms but extend the variation to other less common forms, such as *without, house, now,* and even *council*:

(14) my mum trying to read **withoot** her glasses is hilarious
(15) Finally got internet workin again in ma **hoose noo** to watch some game of thrones
(16) Was going to go for a nap but as always **cooncil** are cutting the grass

The non-standard spellings that represent presence or absence of a FOOT-STRUT split are mapped in Figure 5, revealing a somewhat inconsistent geographic distribution on Twitter.

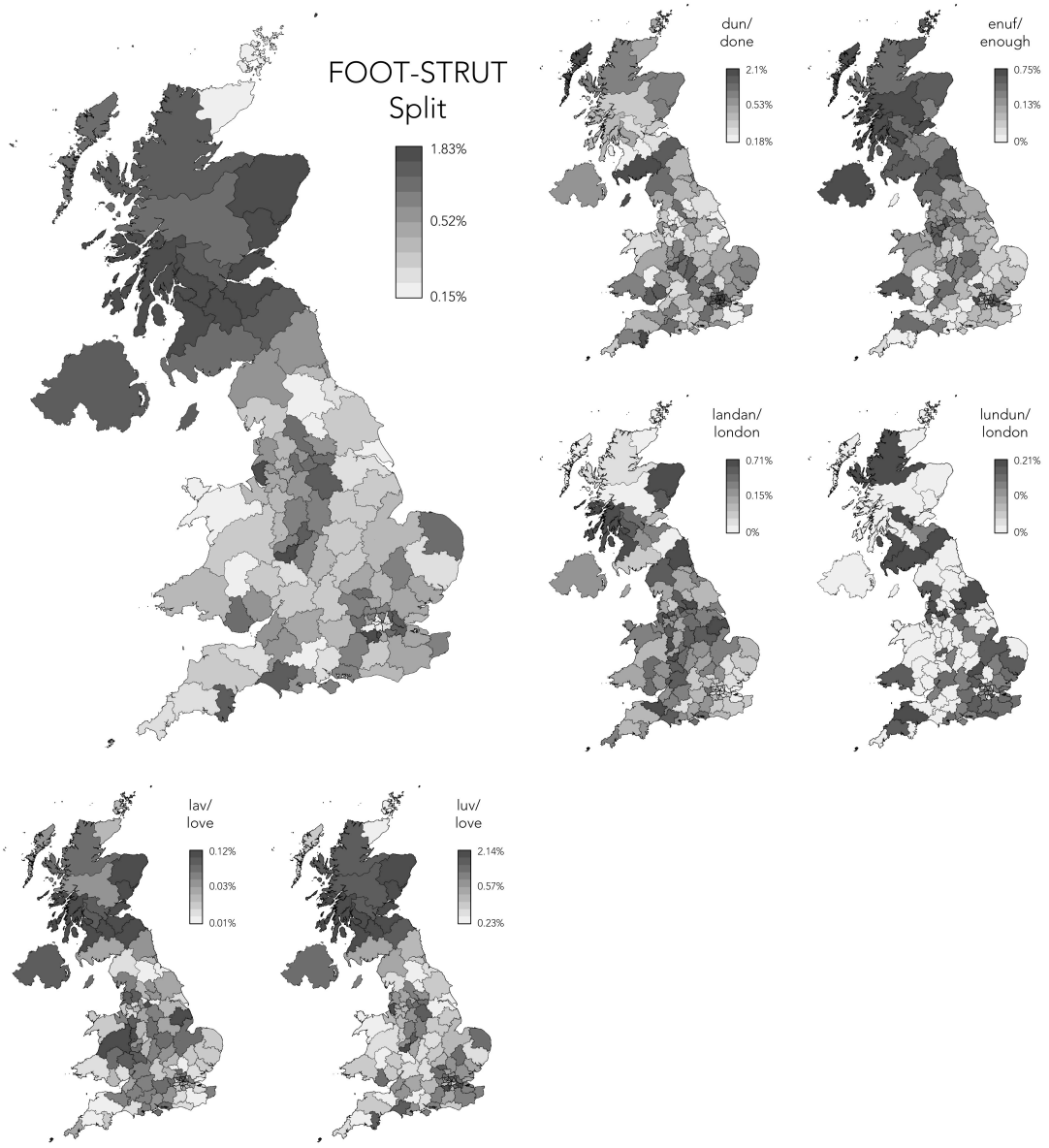


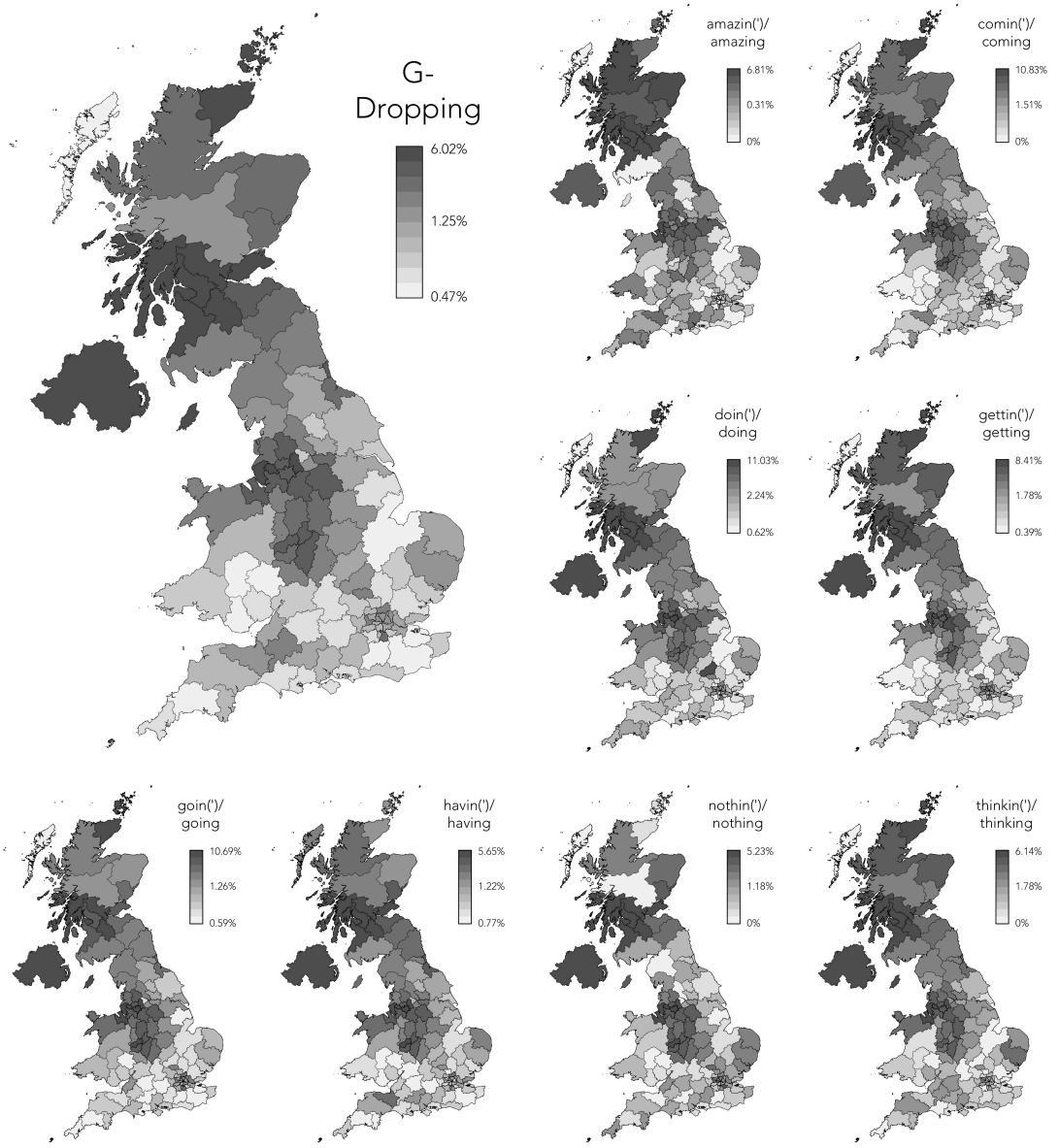
Figure 5 - Maps for a sample of FOOT-STRUT words (*dun, enuf, landan, lundun, lav, luv*)

The distribution of *landan* is particularly interesting, deserving specific mention. This non-standard spelling likely represents the southern STRUT vowel [ʌ], but its regional patterning in Figure 5 actually suggests that it is used most frequently outside of London and the South East of England. This is likely to be another case – discussed earlier in the context of LETTER-backing in *manchestoh* – in which speakers outside of a dialect region are using these variants in an imitative manner to stereotype speakers from that region. It is also not surprising to find this kind of dialectal imitation registered most strongly in place names. It is also interesting to note that in this Twitter corpus the collocation *landan town/taan/tarn* is particularly frequent, exemplified in (17), and that this non-standard spelling commonly co-occurs with other features stereotypical of Estuary English, as in (18).

(17) just touched down in **landan** town
(18) apparently I sound like I'm from **saaaaahf landan** when drinking

It is not easy to interpret the meaning of these results for dialectology because of the confounding factor of users portraying a particular identity or stereotype. However, it is possible that the patterns we observe indicate that Twitter users are more likely to focus on a lowering of this vowel (orthographically represented as <a>), therefore suggesting that this is the most salient dimension along which FOOT-STRUT words differ.

The analysis of the G-dropping feature presented in Figure 6 shows a remarkable degree of consistency with respect to the clear northern trend of g-dropping, thus indicating that this feature is commonly adopted in spelling in the north of the UK. As such, there are strong parallels between the regional patterning of the phonetic feature and its orthographic reflection on Twitter, corroborating the afore-mentioned claim that this feature is primarily associated with region, rather than social status, in the case of British English.



G-dropping feature and for a sample of its words (*amazin/amazin', comin/comin', doin/doin', gettin/gettin', goin/goin', havin/havin', nothin/nothin', thinkin/thinkin'*)

An interesting picture also emerges for TH-fronting and TH-stopping, mapped in Figure 7 and Figure 8, respectively. Both of these features have strong ties to the south of England, specifically London, but their phonetically-motivated spelling variants show a different geographic distribution on Twitter. While TH-fronting has clearly diffused throughout most of the UK, showing no clear regional pattern, TH-stopping is somewhat more restricted to London and its surrounding area. This is particularly evident in the case of *ting* and *dem,* which are arguably the most salient examples of this sociophonetic phenomenon.

Nini, A., Bailey, G., Guo, D., Grieve, J. (2020) The graphical representation of phonetic dialect features of the North of England on social media. In Honeybone, P. & Maguire, W. (eds), *Dialect Writing and the North of England*, 266-296, Edinburgh: Edinburgh University Press

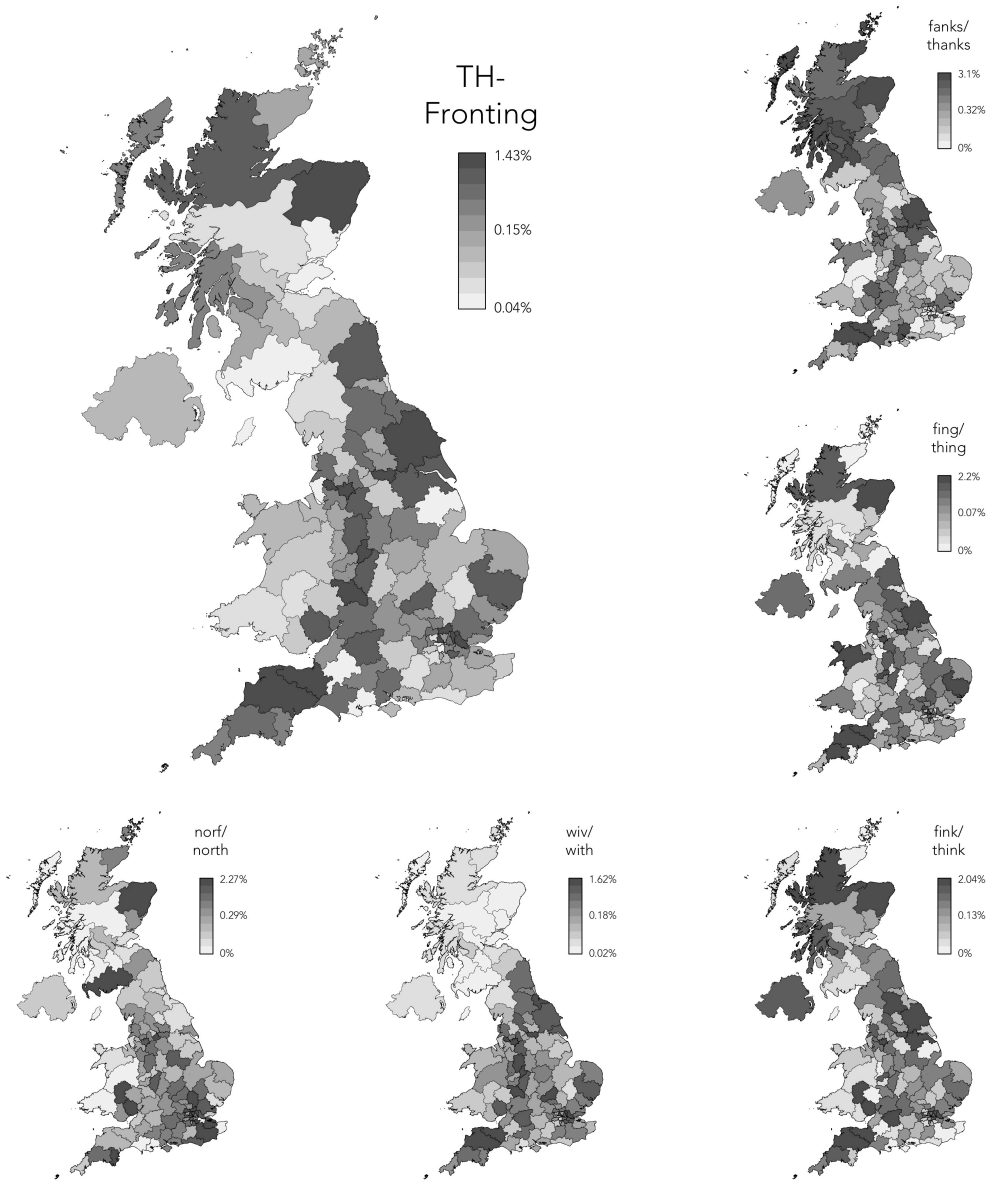


Figure 7 - Maps for the TH-fronting feature and for a sample of its words (*fanks, fing, fink, norf, wiv*)

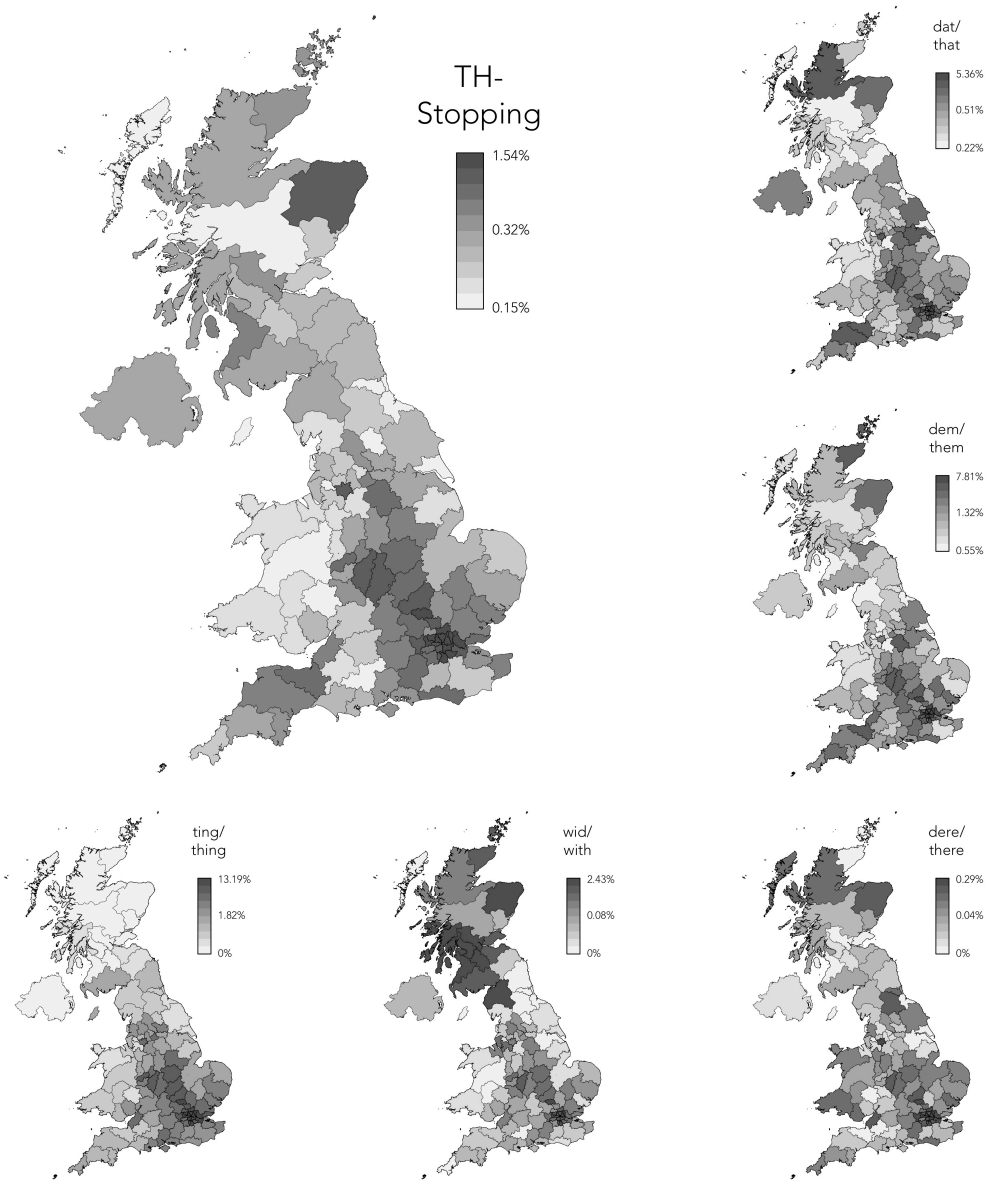


Figure 8 - Maps for the TH-stopping feature and for a sample of its words (*dat, dem, dere, ting, wid*)

The graphical representation of H-dropping seems also to be widespread across the UK on Twitter, despite being overall more common in the North (top left of Figure 9). Although other studies have found that H-dropping in content words is more typical of the North while H-dropping in function words is common everywhere, our results largely found the reverse of this pattern, with H-dropping in function words, such as *have* or *here*, seemingly more frequent in the North and particularly in Scotland.

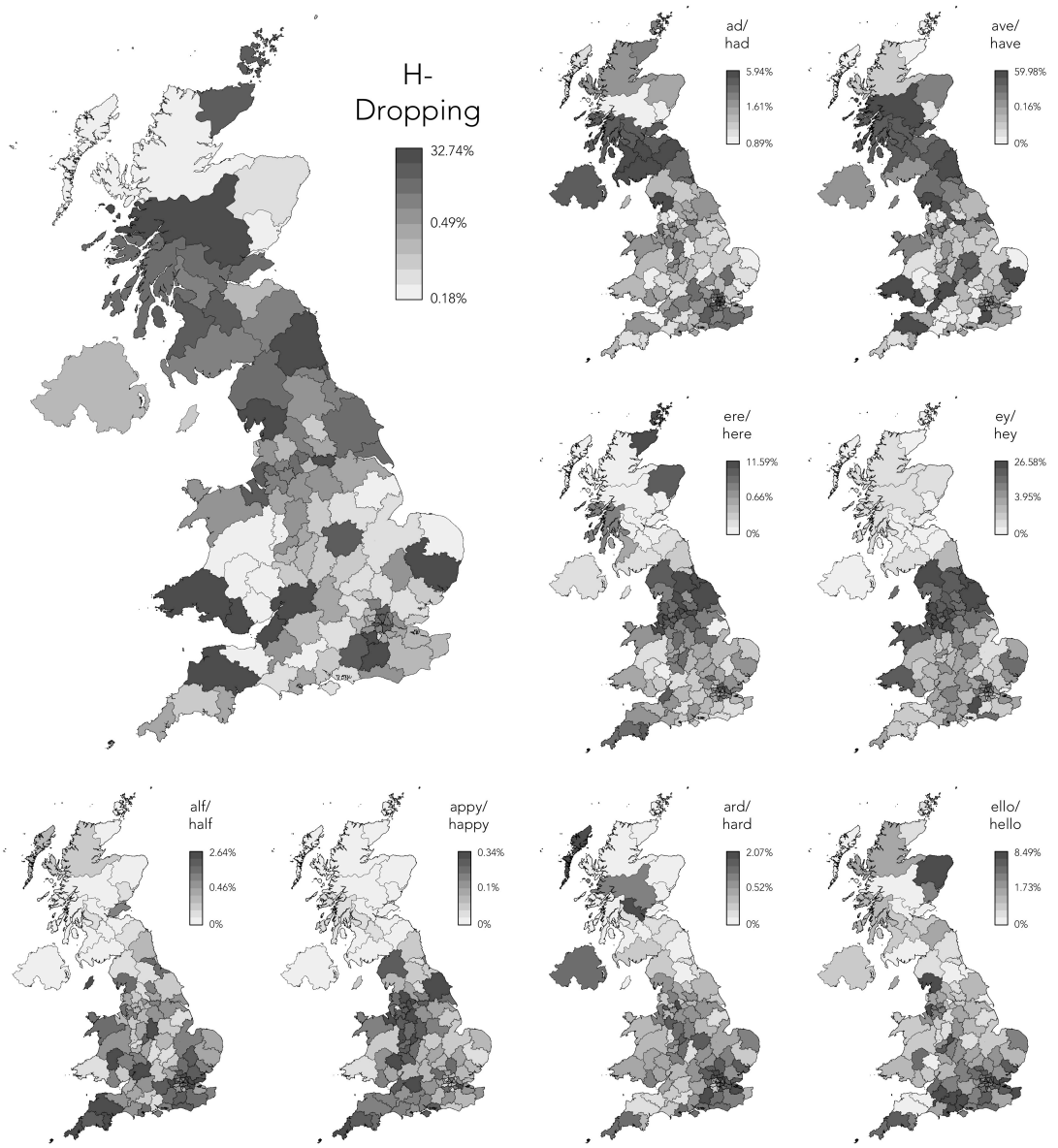


Figure 9 - Maps for the H-dropping feature and for a sample of its words (*ad, ave, ere, ey, alf, appy, ard, ello*)

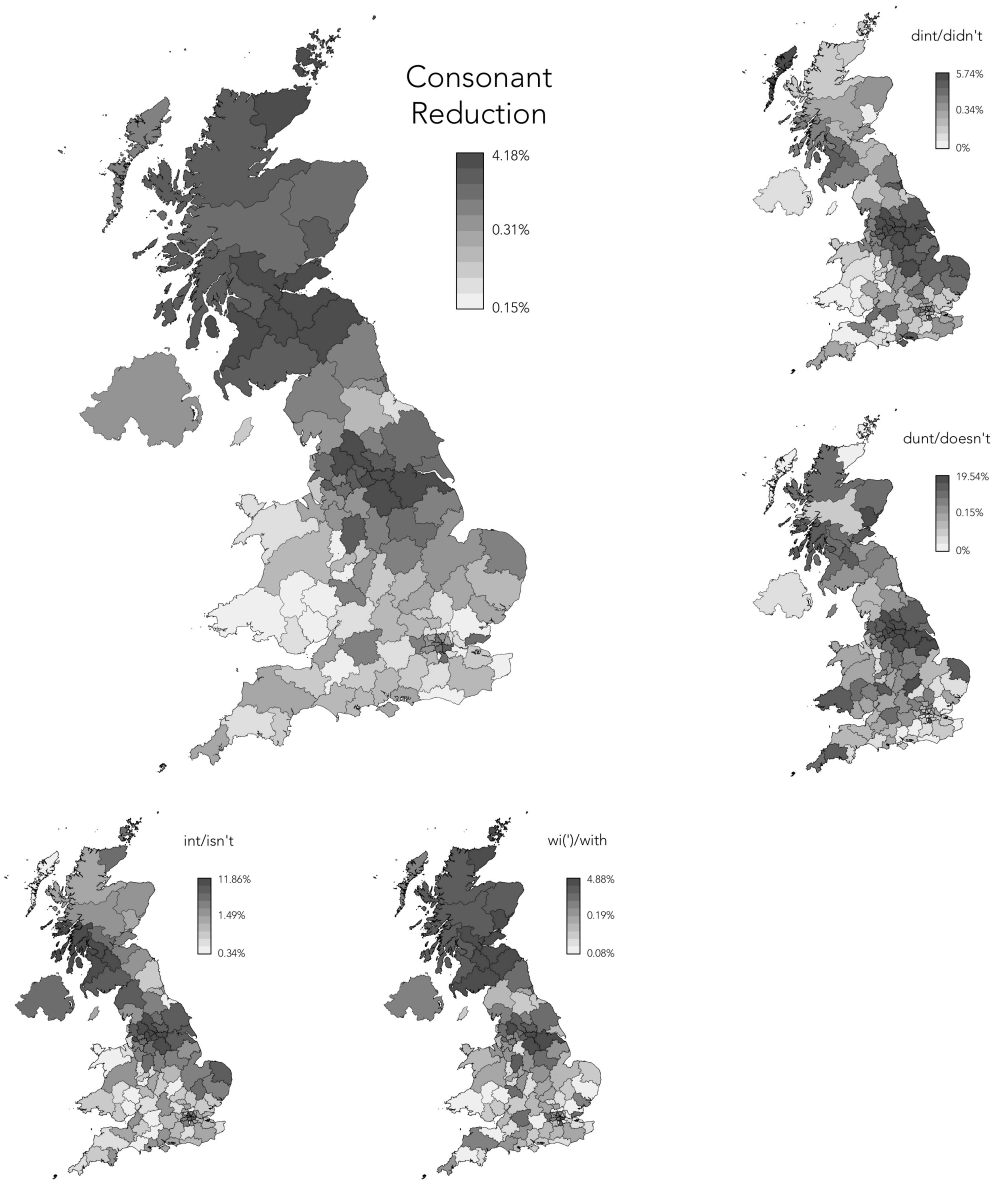


Figure 10 - Maps for the consonant reduction feature and for all of its words (*dint, dunt, int, wi/wi'*)

The top left map in Figure 10 confirms that, even in social media writing, consonant reduction on Twitter seems to be more common in the north of England and in Scotland. The morphosyntactic function of this phonetic feature also has an effect on the regional distribution, with the contracted forms of auxiliaries (*dint, dunt,* and *int*) largely patterning together in the North of England while the abbreviation of the preposition (*wi/wi'*) is more frequent in Scotland. Examples from the data reveal that these reduced variants commonly co-occur with other non-standard spellings: note the representation of *right* as <reyt> in (19), reflecting the [ɛɪ] realisation typical of varieties spoken in Yorkshire, as well as the presence of TH-fronting (*bovva*) and vocalic reduction (*sez*) in (20).

(19) Get so use to people calling me by my last name; then when I get called by my first name it **dunt** sound reyt!
(20) **dunt** bovva me but mum sez I'll have no jacket left!

(21) Always have the urge ti start singing along **wi** ma music when am on public transport

The variants for vowel reduction altogether show a northern pattern, being favoured particularly in Scotland and in the North West of England, as can be seen in the maps in Figure 11.

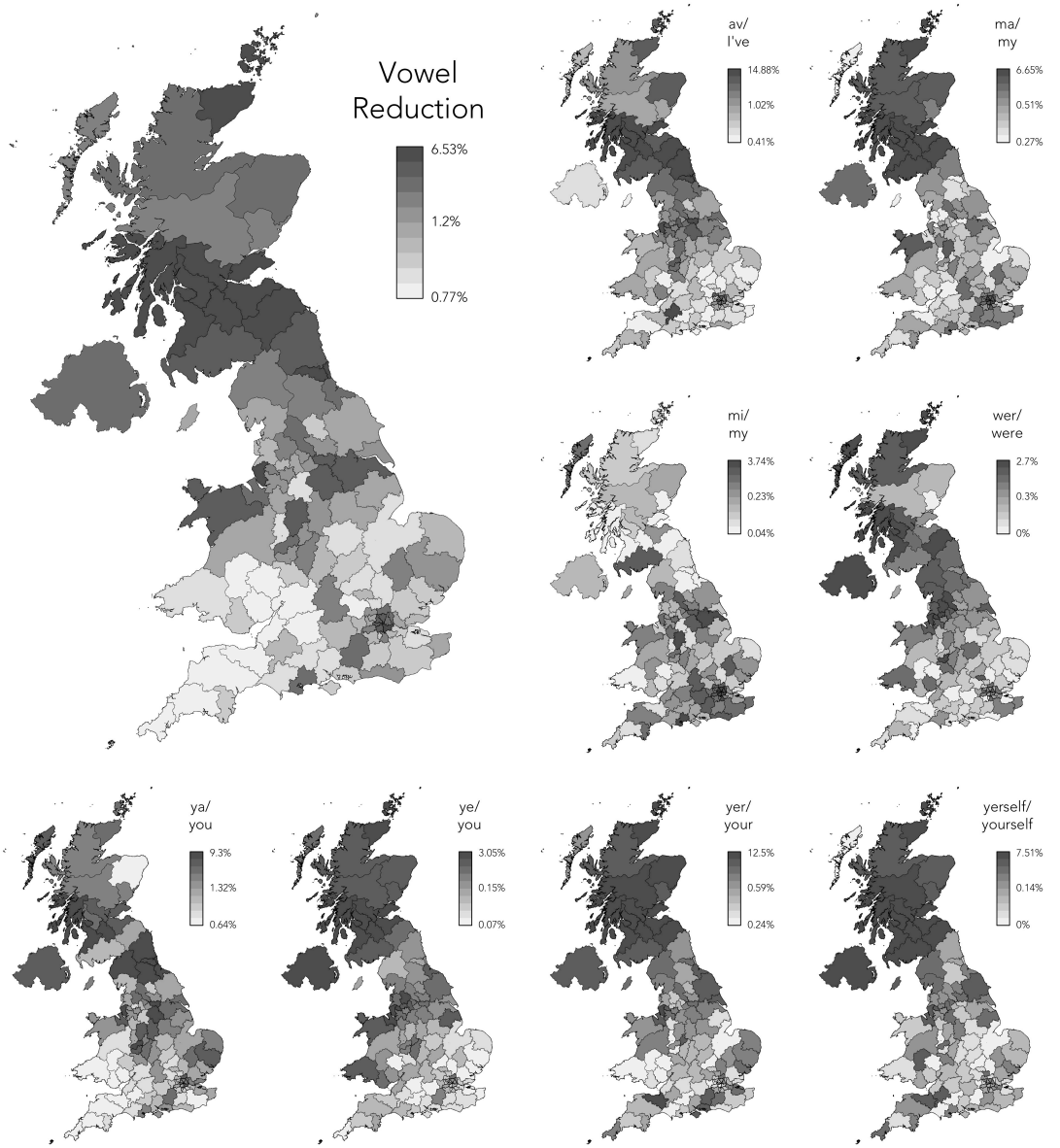


Figure 11 - Maps for the vowel reduction feature and for a sample of its words (*av, ma, mi, wer, ya, ye, yer, yerself*)

This feature is particularly interesting since the justification of its use cannot even be found in its length, which is often equal to its standard counter-part. In other words, Twitter users are not using these reduced variants because they are quicker to type or because of character restrictions in tweets, but rather because they wish to convey a particular identity or stance. For example, consider:

(22) Some ppl are just so rude sort **yerself** out
(23) i guess that means **am** buying **yer** pints all night

(24) think i've got food poisoning from **ma** chippy tea

12.7. Discussion

In this chapter we have shown that despite being a written medium of communication, Twitter can be used as a rich source of natural language data for contemporary studies of phonetic dialectal variation, and the results presented here have a number of implications both for theoretical and methodological issues in this field of study.

As discussed above, third wave approaches to language variation and change foreground the importance of indexicality and the active way in which speakers employ socially-meaningful variants in acts of identity construction and stance-taking. While there is a sizeable body of work adopting this line of inquiry in the domain of speech production (e.g. Labov, 1963; Eckert, 2000; Zhang, 2005; Pharao et al., 2014), there is comparatively little consideration of how these same forces are at play in other modalities, such as orthographic variation in written forms on social media. Our results present strong evidence that users of social media often employ creative non-standard spellings that reflect the phonetic realisation of the same words in their own spoken dialect, at least in the case of Northern English. For many of the northern features under study, the regional distribution of these non-standard forms matches well with the regional stratification of their phonetic equivalents. This suggests that in a large majority of cases we are indeed seeing phonetically-motivated orthography and a clear relationship between how dialect is projected across both speech and writing.

These results also lend insight into the relative salience of different dialectal features, which is an important concept in contemporary approaches to language variation and change but is difficult to operationalise and – as a result – often poorly defined (Auer, Barden and Grosskopf, 1998; see also discussion in Jaeger and Weatherholtz, 2016). Although, as discussed earlier, the third wave variationist approach emphasises the active role of speakers in using particular variants, it remains the case that many dialectal features are used in a relatively subconscious manner as a natural consequence of speakers' own linguistic systems. However, the use of orthographic equivalents on Twitter presents an interesting contrast in that most of the features explored in this study require a much more deliberate action. These considerations imply that most of the words and features we observed are subject to a high level of indexicality in the speech community and can therefore be considered *stereotypes* (Labov, 1978). To the extent that the nonstandard spellings found are made with full awareness of their social meanings, this study shows that the process of dialect *enregisterment* (Agha, 2005) that has been attested in literary and artistic contexts, especially by authors in this volume (e.g. see Beal; Clark,; Cooper) is also used with similar purposes by the speech community in social media.

In turn, we can gain insight into the social salience of a dialectal feature by looking closely at the extent to which it is registered orthographically on Twitter. In other words, which features do people tend to focus on when attempting to construct and project a particular dialectal identity through writing? Of course, it is important to note that in some cases this is confounded by the correspondence – or lack thereof – between sound and spelling in English, and how certain features are simply not possible to reflect through graphemic replacement or substitution. However, it nevertheless provides an interesting and novel approach to the study of social and linguistic salience in the context of non-standard dialectal features. The relevance of salience is arguably most notable in cases of mismatch between the regional patterns

of graphemic and phonetic variation. As shown in the results in Section 12.6, this occurs most frequently with place names when users perform a kind of dialect imitation. In other words, quantitative and qualitative analysis of the data suggests that when certain sociophonetic features can be registered in place names, they tend to be used not by speakers of those dialects but by other speakers drawing on such features in stereotypes of that variety. Although this is problematic if the primary goal is to use these methods of analysis to replace conventional methods in dialectology, which would rely on the assumption that a person's graphemic forms reflect their phonetic forms, it provides yet more insight into the features that are socially salient in stereotypes of certain dialects. These findings therefore also provide new evidence for the understanding of salience and the influencing or constraining factors of phonetic dialect writing (see Honeybone, this volume).

Another important point to highlight is the way in which these phonetically-motivated orthographic forms often co-occur with other non-standard spellings, as seen in many of the examples given in Section 12.6. Although this calls for a more nuanced approach to the covariation between different dialectal features, the fact that these spelling variants do not occur as isolated examples suggests that they are indeed used as part of a wider linguistic style tailored to a user's own dialectal identity.

Finally, from a methodological point of view, it is very important to note that, as in previous sociolinguistic or dialect studies using Twitter, the effects found in this research are largely consistent with previous findings. This confirmation is particularly important because geo-coded Twitter data has at least two major limitations. Firstly, the population of geo-coded Twitter users is not a representative sample of the population of the UK, as discussed above. Secondly, although a geocoded tweet contains exact information of the location from which it was sent, it does not contain information of the dialect background or area of origin of the person who is writing it. For these reasons, social media data should always be treated carefully, as these representativeness biases are well known. However, despite these problems, the results presented in this chapter contribute to suggesting that, if a very large data set has been collected, the geographical signal underlying dialect patterns can be still detected in social media data of this kind even through the noise generated by these two confounding factors.

12.8. Conclusions

In conclusion, the present chapter offered an analysis of a large corpus of geo-coded tweets for graphological variation reflecting dialectal phonetic variation of features found in the dialects of the north of England. The results of the analysis provide a new angle on both the study of the dialects of the north and on wider issues related to methods in dialectology. The findings of this analysis reveal that users of social media adopt spelling variants that reflect their dialects and, more specifically, to represent their identity in social media. This phenomenon is therefore consistent with third wave sociolinguistics considerations that stress the importance of identity performance. Moreover, these results suggest that the analysis of social media like Twitter, especially if they are geo-coded, can not only offer a lot of useful real data in much less time than a survey would take, but that they can also offer an altogether new perspective. As the qualitative and quantitative analysis revealed, naturally occurring corpus data, which is not affected by the observer paradox, can uncover how these variants can either have a stereotyped function within a speech community or be used to portray an identity. This new approach to the study of dialectology and

sociolinguistics not only can lead to interesting new findings but, altogether, lead to new fundamental questions on the nature of linguistic variation.

12.9. References

Agha, Asif (2005) Voice, footing, enregisterment, *Journal of Linguistic Anthropology*, **15**(1), pp. 1–5.

Asprey, Esther (n.d.) Black Country dialect literature and what it can tell us about Black Country dialect, In *Dialect Writing and the North of England*, Honeybone, Patrick and Maguire, Warren (eds.), Edinburgh, Edinburgh University Press.

Auer, Peter, Barden, Birgit and Grosskopf, Beate (1998) Subjective and objective parameters determining 'salience' in long-term dialect accommodation, *Journal of Sociolinguistics*, John Wiley & Sons, Ltd, **2**(2), pp. 163–187.

Bailey, George (2019) Emerging from below the social radar: Incipient evaluation in the North West of England, *Journal of Sociolinguistics*, **23**(1), pp. 3–28.

Bailey, George (2016) Regional variation in 140 characters: Mapping geospatial tweets, In *Workshop on Using Twitter for Linguistic Research*, University of Kent.

Baranowski, Maciej (2017) Class matters: The sociolinguistics of GOOSE and GOAT in Manchester English, *Language Variation and Change*, **29**, pp. 301–339.

Baranowski, Maciej and Turton, Danielle (2015) Manchester English, In *Researching Northern English*, Hickey, Raymond (ed.), Amsterdam, John Benjamins, pp. 293–316.

Beal, Joan C. (n.d.) Dialect and the construction of identity in the ego-documents of Thomas Bewick, In *Dialect Writing and the North of England*, Honeybone, Patrick and Maguire, Warren (eds.), Edinburgh, Edinburgh University Press.

Beal, Joan C. (2004) English dialects in the North of England: Phonology, In *A handbook of varieties of English. Vol 1: Phonology*, Schneider, Edgar W., Burridge, Kate, Kortmann, Bernd, Mesthrie, Rajend, and Upton, Clive (eds.), Berlin, Mouton de Gruyter, pp. 113–133.

Beal, Joan C. (2000) From Geordie Ridley to Viz: Popular literature in Tyneside English, *Language and Literature*, Sage PublicationsSage CA: Thousand Oaks, CA, **9**(4), pp. 343–359.

Beal, Joan C., Burbano-Elizondo, Lourdes and LLamas, Carmen (2012) *Urban North-Eastern English: Tyneside to Teesside*, Edinburgh, Edinburgh University Press.

Braber, Natalie (n.d.) Nottingham: City of Literature. Dialect Literature and Literary Dialect, In *Dialect Writing and the North of England*, Honeybone, Patrick and Maguire, Warren (eds.), Edinburgh, Edinburgh University Press.

Britain, David (2002) Diffusion, levelling, simplification and reallocation in past tense BE in the English Fens, *Journal of Sociolinguistics*, Wiley/Blackwell (10.1111), **16**(1), pp. 16–43.

Buchstaller, Isabelle, Corrigan, Karen P., Holmberg, Anders, Honeybone, Patrick and Maguire, Warren (2013) T-to-R and the Northern Subject Rule: Questionnaire-based spatial, social and structural linguistics, *English Language and Linguistics*, Cambridge University Press, **17**(01), pp. 85–128.

Cardoso, Amanda (2015) Variation in nasal–obstruent clusters and its influence on PRICE and MOUTH in Scouse, *English Language and Linguistics*, Cambridge University Press, **19**(03), pp. 505–532.

Carter, Paul (2002) Structured variation in British English liquids, University of York.

Clark, Lynn and Watson, Kevin (2011) Testing claims of a usage-based phonology with Liverpool English t-to-r, *English Language and Linguistics*, Cambridge University Press, **15**(03), pp. 523–547.

Clark, Urszula (n.d.) Enregistering dialect representation in Staffordshire Potteries' cartoons, In *Dialect Writing and the North of England*, Honeybone, Patrick and Maguire, Warren (eds.), Edinburgh, Edinburgh University Press.

Cook, Paul, Han, Bo and Baldwin, Timothy (2014) Statistical Methods for Identifying Local Dialectal Terms from GPS-Tagged Documents, *Dictionaries: Journal of the Dictionary Society of North America*, **35**, pp. 248–271.

Cooper, Paul (n.d.) Russian dolls and dialect literature: the enregisterment of nineteenth century 'Yorkshire' dialects, In *Dialect Writing and the North of England*, Honeybone, Patrick and Maguire, Warren (eds.), Edinburgh, Edinburgh University Press.

Doyle, Gabriel (2014) Mapping dialectal variation by querying social media, In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Wintner, Shuly, Goldwater, Sharon, and Riezler, Stefan (eds.), Gothenburg, Sweden, pp. 98–106.

Drummond, Rob (2012) Aspects of identity in a second language: ING variation in the speech of Polish migrants living in Manchester, UK, *Language Variation and Change*, Cambridge University Press, **24**(01), pp. 107–133.

Drummond, Rob (2018) Maybe it's a grime [t]ing: TH-stopping among urban British youth, *Language in Society*, Cambridge University Press, **47**(02), pp. 171–196.

Eckert, Penelope (2000) *Linguistic variation as social practice*, Oxford, Blackwell.

Eckert, Penelope (2012) Three Waves of Variation Study: The Emergence of Meaning in the Study of Sociolinguistic Variation, *Annual Review of Anthropology*, Annual Reviews , **41**(1), pp. 87–100.

Eisenstein, Jacob (2018) Identifying regional dialects in online social media, In *The Handbook of Dialectology*, Boberg, Charles, Nerbonne, John, and Watt, Dominic (eds.), Hoboken, NJ, Wiley-Blackwell, pp. 368–383.

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

Eisenstein, Jacob (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.*, Stroudsburg, Pennsylvania, Association for Computational Linguistics, pp. 359–369.

Eisenstein, Jacob, O'Connor, Brendan, Smith, Noah A. and Xing, Eric P. (2012) Mapping the geographical diffusion of new words, *arXiv*, **1210.5268**, pp. 1–13.

Foulkes, Paul and Docherty, Gerard (2007) Phonological variation in England, In *Language in the British Isles*, Britain, David (ed.), Cambridge, Cambridge

University Press, pp. 52–74.

Gonçalves, Bruno and Sánchez, David (2014) Crowdsourcing Dialect Characterization through Twitter, Physics and Society; Computation and Language; Machine Learning, p. 10.

Grieve, Jack, Montgomery, Chris, Nini, Andrea, Murakami, Akira and Guo, Diansheng (2019) Mapping lexical dialect variation in British English using Twitter, *Frontiers in Artificial Intelligence*, **2**(11).

Grieve, Jack, Nini, Andrea and Guo, Diansheng (2017) Analyzing lexical emergence in Modern American English online, *English Language and Linguistics*, Cambridge University Press, **21**(01), pp. 99–127.

Guy, Gregory R. (1991) Explanation in variable phonology: An exponential model of morphological constraints, *Language Variation and Change*, Cambridge University Press, **3**(01), pp. 1–22.

van Heuven, Walter J. B., Mandera, Pawel, Keuleers, Emmanuel and Brysbaert, Marc (2014) SUBTLEX-UK: A New and Improved Word Frequency Database for British English, *Quarterly Journal of Experimental Psychology*, Routledge, **67**(6), pp. 1176–1190.

Honeybone, Patrick (2007) New-dialect formation in nineteenth century Liverpool: A brief history of Scouse, In *The Mersey Sound: Liverpool's Language, People and Places*, Grant, A. and Grey, C. (eds.), Liverpool, Open House Press, pp. 106–140.

Honeybone, Patrick (n.d.) Which phonological features get represented in dialect writing? Answers from three types of Liverpool English texts, In *Dialect Writing and the North of England*, Honeybone, Patrick and Maguire, Warren (eds.), Edinburgh, Edinburgh University Press.

Honeybone, Patrick, Watson, Kevin and van Eyndhoven, Sarah (2017) Lenition and T-to-R are differently salient: The representation of competing realisations of /t/ in Liverpool English dialect literature, In *Perspectives on Northern Englishes*, Beal, Joan C. and Hancil, Sylvie (eds.), Berlin, Mouton de Gruyter, pp. 83–108.

Houston, Ann Celeste (1985) Continuity and change in English morphology: The variable (ING), University of Pennsylvania.

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

Hughes, Arthur, Trudgill, Peter and Watt, Dominic (2012) *English accents and dialects*, London, Routledge.

Jaeger, T. Florian and Weatherholtz, Kodi (2016) What the Heck Is Salience? How Predictive Language Processing Contributes to Sociolinguistic Perception, *Frontiers in Psychology*, Frontiers, **7**, pp. 1–5.

Johnston, Paul (1997) Regional variation, In *The Edinburgh history of the Scots language*, Jones, Charles (ed.), Edinburgh, Edinburgh University Press, pp. 433–513.

Jones, Taylor (2015) Toward a Description of African American Vernacular English Dialect Regions Using 'Black Twitter', *American Speech*, Duke University Press, **90**(4), pp. 403–440.

Kerremans, D., Stegmayr, S. and Schmid, H. (2012) The NeoCrawler: Identifying and retrieving neologisms from the Internet and monitoring ongoing change, In *Current Methods in Historical Semantics*, Allan, K. and Robinson, J. A. (eds.), Berlin, de Gruyter Mouton, pp. 59–96.

Kilgarriff, Adam and Grefenstette, Gregory (2003) Introduction to the special issue on web as corpus, *Computational Linguistics*, **29**(3), pp. 1–15.

Kulkarni, Vivek, Perozzi, Bryan and Skiena, Steven (2016) Freshman or Fresher? Quantifying the Geographic Variation of Internet Language, In *Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM 2016)*, Strohmaier, Markus and Gummadi, Krishna P. (eds.), Palo Alto, California, The AAAI Press, pp. 615–618.

Labov, William (2001) *Principles of Linguistic Change. Volume 2: Social Factors*, Malden, MA; Oxford, UK, Blackwell Publishers.

Labov, William (1978) *Sociolinguistic Patterns*, Oxford, Blackwell.

Labov, William (1989) The child as linguistic historian, *Language Variation and Change*, **1**, pp. 85–97.

Labov, William (1963) The social motivation of a sound change, *Word*, **19**, pp. 273–309.

Levon, Erez and Fox, Sue (2014) Social salience and the sociolinguistic monitor: A case study of ING and TH-fronting in Britain, *Journal of English Linguistics*, **42**(3), pp. 185–217.

Longley, Paul A., Adnan, Muhammad and Lansley, Guy (2015) The geotemporal demographics of Twitter usage, *Environment and Planning A*, **47**, pp. 465–484.

MacKenzie, Laurel, Bailey, George and Turton, Danielle (2014) Crowdsourcing dialectology in the undergraduate classroom, In *Methods in Dialectology XV*, University of Groningen.

McColl Millar, Robert (2007) *Northern and Insular Scots*, Edinburgh, Edinburgh University Press.

Mellon, Jonathan and Prosser, Christopher (2017) Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users, *Research and Politics*, **4**(3), pp. 1–9.

Milroy, James, Milroy, Lesley, Hartley, Sue and Walshaw, David (1994) Glottal stops and Tyneside glottalization: Competing patterns of variation and change in British English, *Language Variation and Change*, **6**, pp. 327–357.

Moore, Samuel, Meech, Sanford B. and Whitehall, Harold (1935) *Middle English dialect characteristics and dialect boundaries*, Ann Arbor, University of Michigan Language and Literature Series.

Nguyen, Dong, Doğruöz, A. Seza, Rosé, Carolyn P. and De Jong, Franciska (2016) Computational Sociolinguistics: A Survey, *Computational Linguistics*, **42**(3), pp. 537–593.

Pavalanathan, Umashanthi and Eisenstein, Jacob (2015) Confounds and Consequences in Geotagged Twitter Data, In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, Association for Computational Linguistics, pp. 2138–2148.

Petyt, Keith M. (1985) *Dialect and accent in industrial West Yorkshire*, Amsterdam, John Benjamins.

Pharao, Nicolai, Maegaard, Marie, Møller, Janus Spindler and Kristiansen, Tore (2014) Indexical meanings of [s+] among Copenhagen youth: Social perception of a phonetic variant in different prosodic contexts, *Language in Society*, Cambridge University Press, **43**(01), pp. 1–31.

Rahimi, Afshin, Cohn, Trevor and Baldwin, Timothy (2017) A Neural Model for User Geolocation and Lexical Dialectology, *arXiv*, (1704.04008).

Ramsammy, Michael and Turton, Danielle (2012) Higher or lower? An investigation of the unstressed vowel system in Mancunian English, In *Manchester and Salford New Researchers Forum in Linguistics*.

Shoemark, Philippa, Kirby, James and Goldwater, Sharon (2017) Topic and audience effects on distinctively Scottish vocabulary usage in Twitter data, In *Proceedings of the Workshop on Stylistic Variation*, Copenhagen, Denmark, pp. 59–68.

Tagliamonte, Sali (2004) Somethi[ŋ]'s goi[n] on! Variable (ing) at Ground zero, In *Language Variation in Europe: Papers from the Second International Conference on Language Variation in Europe (ICLaVE 2)*, Gunnarson, Britt-Louise, Bergström, Lena, Eklund, Gerd, Fridell, Staffan, Hansen, Lise H., Karstadt, Angela, Nordberg, Bengt, Sundgrenand, Eva, and Thelander, Mats (eds.), Uppsala, Uppsala University, pp. 390–403.

Tagliamonte, Sali and Temple, Rosalind (2005) New perspectives on an ol' variable: (t,d) in British English, *Language Variation and Change*, Cambridge University Press, **17**(03), pp. 281–302.

Tanner, James, Sonderegger, Morgan and Wagner, Michael (2017) Production planning and coronal stop deletion in spontaneous speech, *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, Ubiquity Press, **8**(1), pp. 1–39.

Tatman, Rachael (2016) 'I'm a spawts guay': Comparing the Use of Sociophonetic Variables in Speech and Twitter, *University of Pennsylvania Working Papers in Linguistics: Selected Papers from NWAV 44*, **22**, pp. 160–170.

Trudgill, Peter (1986) *Dialects in contact*, Oxford, Blackwell.

Trudgill, Peter (1999) *The dialects of England*, Oxford, Blackwell.

Turton, Danielle (2017) Categorical or gradient? An ultrasound investigation of /l/-darkening and vocalization in varieties of English, *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, Ubiquity Press, **8**(1), pp. 1–31.

Wakelin, Martyn (1977) *English Dialects: An Introduction*, London, The Athlone Press.

Watson, Kevin (2007) Liverpool English, *Journal of the International Phonetic Association*, **37**, pp. 351–360.

Watt, Dominic (2002) 'I don't speak with a Geordie accent, I speak, like, the Northern accent': Contact-induced levelling in the Tyneside vowel system, *Journal of Sociolinguistics*, Wiley/Blackwell (10.1111), **6**(1), pp. 44–63.

Watt, Dominic and Milroy, Lesley (1999) Variation in three Tyneside vowels: Is this dialect levelling?, In *Urban voices: Accent studies in the British Isles*, Foulkes,

Paul and Docherty, Gerard (eds.), London, Arnold, pp. 25–46.

Watts, Emma (2005) Mobility-induced dialect contact: A sociolinguistic investigation of speech variation in Wilmslow, Cheshire, University of Essex.

Wells, John C. (1982) *Accents of English vol. 2: The British Isles*, Cambridge, Cambridge University Press.

Williams, Ann and Kerswill, Paul (1999) Dialect levelling: Change and continuity in Milton Keynes, Reading and Hull, In *Urban voices: Accent studies in the British Isles*, Foulkes, Paul and Docherty, Gerard (eds.), London, Arnold, pp. 141–162.

Zhang, Qing (2005) A Chinese yuppie in Beijing: Phonological variation and the construction of a new professional identity, *Language in Society*, Cambridge University Press, **34**(03), pp. 431–466.

12.10. Appendix

List of words searched in the corpus:

T to R	*lorra, gerra, gerrout, gerroff, gerrup, shurrup*
HAPPY-laxing	*vereh, realleh, onleh, aneh, maneh, moneeh, actualleh, evereh, famileh, countreh, loveleh, sorreh, probableh, absoluteleh, happeh, parteh, readeh, pretteh, alreadeh, storeh, citeh, everybodeh, exactleh, properteh, obviousleh, certainleh, earleh, easeh, babeh, definiteleh, somebodeh, historeh, centureh, companeh, bodeh, economeh, ladeh, completeleh, finalleh, worreh, communiteh, quickleh, hopefulleh, nearleh, nobodeh, particularleh, opportuniteh, funneh, anybodeh, luckeh, yesterdaeh, secretareh, basicalleh, clearleh, energeh, especialleh, plenteh, qualiteh, buseh, slightleh*
LETTER-backing	*manchestah, manchestoh*
AW to UW	*aboot, oot, doon, aroond, hoose, roond, foond, withoot, sooth, ootside, hoors, toon, groond, hoor, soond, cooncil, soonds, amoont, thoosands, broon*
FOOT-STRUT	*dun, enuf, enaf, lundun, landan, luv, lav*
G-dropping	*goin|goin', somethin|somethin', bein|bein', doin|doin', lookin|lookin', comin|comin', gettin|gettin', nothin|nothin', havin|havin', tryin|tryin', makin|makin', sayin|sayin', workin|workin', mornin|mornin', talkin|talkin', takin|takin', amazin|amazin', playin|playin', thinkin|thinkin', livin|livin', feelin|feelin', interestin|interestin', runnin|runnin', buildin|buildin', usin|usin', durin|durin', movin|movin', waitin|waitin', evenin|evenin', givin|givin', seein|seein', watchin|watchin', puttin|puttin', hopin|hopin', happenin|happenin', startin|startin', meetin|meetin', sellin|sellin', sittin|sittin', spendin|spendin'*
TH-stopping	*dat, dis, dey, wid, dere, tink, dem, deir, dese, someting, oder, ting, tings, tought, anoder, anyting, everyting, togeder, widout, wheder, monts, sout, tinking, nort, moder, wort, fader, widin, furder, eider, tird, demselves, deat, healt, oders, weader, aldough, nordern, mont, tousands, broder, eart*
TH-fronting	*wiv, fink, somefing, fing, fings, anyfing, everyfing, nofing, togever, wivout, fanks, finking, norf, furver*
H-dropping	*ave, ere, ad, ome, ouse, elp, aving, ello, ard, appy, ope, alf, appened, ey, eard, appen, uge, istory, imself, ealth, appens, opefully, ospital, oping, uman, appening*
consonant reduction	*dunt, dint, int, wi|wi'*
vowel reduction	*yer, yerself, ye, ya, av, ar, mi, ma, wer*