To appear in Similar Languages, Varieties, and Dialects (Studies in Natural Language Processing), edited by Marcos Zampieri and Preslav Nakov, Cambridge: Cambridge University Press

10

Dialect Variation on Social Media

Dong Nguyen

10.1 Introduction

Social media has changed our daily lives: We share our thoughts, opinions and news using social media and connect with people throughout the world. Social media has also radically changed a variety of research disciplines: It is both *massive*—we can now study potentially millions of people—and *microscopic*—we can carry out analyses at the level of individual interactions (Golder and Macy, 2014). Rather than relying on self-reports or elicited data, we can now observe *language in use* at scale in a variety of social contexts. The availability of social media data has been one of the driving factors of the emerging area of *computational sociolinguistics* (Nguyen et al., 2016).

There is no 'single' online language variety (Herring and Androutsopoulos, 2015). Instead, we find a multitude of linguistic varieties and styles in social media—even within a single social media platform. Still, the informal nature of social media platforms means that language in social media is often closer to everyday speech than the language typically found in many other data sources, such as newspapers. Social media is therefore a rich resource to study regional and social variation in language.

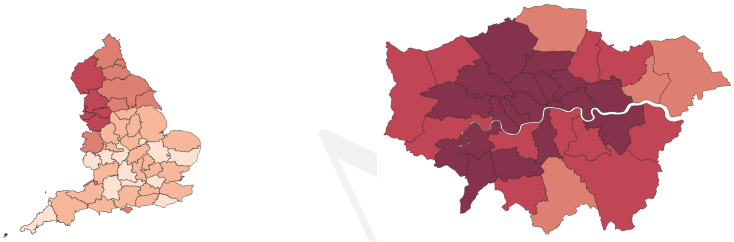
For example, here are two tweets from public Twitter accounts, one by Virgin Media and one by Cara Delevingne, an international model:

virginmedia: *Nice one bruv! Here if you need us. ^MK*

Caradelevingne: *Soo excited 2announce my first Novel titled Mirror Mirror!Pre order on Amazon!!Can't wait 2share story with you all! [LINK]*

The tweet posted from the Virgin Media account involves an interaction with a customer. The language is informal, perhaps to connect with the customer. We find an instance of *bruv*, an address term that has featured in a sociolinguistic study (Kerswill, 2013). The tweet posted by the supermodel is a promotional tweet, with orthographic variation (e.g., *soo* instead of *so* and *2* instead of *to*) and spacing and punctuation that automatic tools would be challenged by.

Patterns in language variation become more salient when we aggregate across a larger number of tweets, for example to study regional patterns. Figure 10.1a shows the relative frequencies of *pants* and *trousers* in England based on geo-tagged Twitter data from May-August 2014[1]. *Pants* has a higher usage in the north west of England, which matches the pattern observed through fieldwork carried out by undergraduate students of Linguistics and English Language at the University of Manchester.[2]



(a) *pants* vs. *trousers* in England

(b) Standard *were* vs. non-standard *was* in London

Figure 10.1 Geotagged tweets, May-August 2014

We can also zoom in on a particular region, for example London. London has been of interest in sociolinguistic studies, because of its multicultural character and the emergence of Multicultural London English. Cheshire and Fox (2009) studied *was* versus *were* variation in London by analyzing speech of adolescents and elderly speakers in the multicultural inner London area (Hackney) and in a less diverse outer London area (Havering). The use of non-standard *was* in standard *were* contexts was higher in outer London adolescents compared to Inner London adolescents. A similar trend is observed in Twitter (Figure 10.1b), by comparing the use of non-standard WAS (*we was, you was, they was*)

[1] Part of a larger dataset collected for UK election passive polling and analysis (Wang et al., 2017a,b).
[2] http://projects.alc.manchester.ac.uk/ukdialectmaps/lexical-variation/trousers/

to standard WERE (*we were, you were, they were*). Standard WERE has a higher occurrence in inner London.

The scale of social media data makes it possible to study rare phenomena, such as specific syntactic constructions or lexical variants. Furthermore, information on interaction patterns makes it possible to jointly analyze geographical variation with a variety of social factors, for example how linguistic choices relate to someone's online conversation partner. However, there are also many challenges: Social media data needs to be repurposed—social media platforms were not designed to study dialect variation—and processing language in social media can be challenging because many NLP tools are not robust to linguistic variation.

This chapter focuses on geographical dialect variation in social media from the perspective of computational linguistics, but it also draws from sociolinguistics and dialectology to identify fruitful future research directions. First, I'll discuss opportunities and challenges that social media offers for analyzing dialects (Section 10.2). Next, I'll briefly discuss the processing of social media data (Section 10.3) and then I'll discuss computational studies on geographical variation in social media (Section 10.4). The chapter concludes with a future outlook (Section 10.5).

10.2 Social Media for Dialect Research

This section discusses aspects of using social media for dialect research.

Unobtrusive observation of language Everyday speech, for example when you are talking to your family or close friends, is of particular interest in the study of dialect. However, capturing everyday speech is difficult. Questionnaires have been fundamental to collect dialect data. For example, a question might be "*What do you call this plant*".[3] However, the way the questions are phrased, or the interaction with the researcher could influence the responses given. Furthermore, questionnaires usually do not support fine-grained measurements regarding the frequency of use of a certain variable and the analysis of *intra-speaker* variation, for example how the choice for a particular variant depends on the situational context. Observations and sociolinguistic interviews are also frequently used to collect data, but here the presence of the observant or the interviewer could again influence the language.

[3] See Llamas (2018) for a discussion on questionnaires for dialect research.

One of the key advantages of using social media for research is that it allows unobtrusive observation of language and behavior (Salganik, 2017). As (Golder and Macy, 2014, p. 133) point out: *"the social pressures and normative constraints on individuals are exerted by their peers rather than by the researchers*". Social media allows us to study how language is *used* in a variety of social contexts. Moreover, language and social behavior are recorded *in real time* and there is no need to specify beforehand which items are of interest, in contrast to questionnaires.

Social media users Traditional dialectology has focused on so-called *NORMs*, i.e. non-mobile, old, rural men (Chambers and Trudgill, 1998, p. 29), because they were believed to be more conservative in their speech. However, from about the 1960s attention has shifted from rural areas to urban areas, and widened to a variety of social groups. With social media, we often have data about both rural and urban areas. The use of social media means a radical shift away from NORMs as the target population. In a 2018 report by PEW Research on social media use by Americans,[4] 88% of 18- to 29-year-olds indicated that they use any form of social media. This declines to only 37% of Americans aged 65 and older. As another example, in a study focused on Dutch Twitter users (Nguyen et al., 2013), a fine-grained manual annotation effort revealed a heavily skewed distribution towards younger users. Furthermore, only 5 out of 2709 users (excluding profiles for which no annotations could be obtained) were annotated as *retired*. Moreover, in social media not all accounts belong to an individual, but sometimes accounts represent organisations, fake people, and bots.

When studying sociolinguistic variation in social media, *demographic information* about the users is often important to understand demographic biases in the data and how language varies across social groups. For example, in Nguyen et al. (2013), there were more females among the younger age category, but more males among the older age categories. When studying how language varies across demographic groups, it is important to control for these unbalanced gender distributions. Unfortunately, in many cases (almost) no demographic information is available. Different approaches have therefore been explored to derive demographic information, such as automatically inferring demographics from language use (see Nguyen et al. (2016) for an overview), combining location-tagged data with census data (e.g., Jacobo et al. (2018)

[4] http://www.pewinternet.org/2018/03/01/social-media-use-in-2018/

Jørgensen et al. (2015)) and deriving demographics from names (e.g., Bamman et al. (2014b)). A limitation of these approaches is that classifications are imposed on users, rather than asking the users themselves, which can be especially problematic when it involves variables like gender and ethnicity. Androutsopoulos (2013) mentions the alternative of moving away from socio-demographic categories and focusing on participant roles (e.g. admin, novice).

Critical for studying dialect variation in social media is location information. Social media content sometimes comes with fine-grained location information, such as GPS coordinates, and many studies rely on geo-tagged content alone. However, this is often only a small fraction of the number of posts produced. Studies sometimes estimate a 'home location' for social media users, for example based on the location of the first tweet (e.g., Eisenstein et al. (2010)), or by using the most frequent location (e.g., Jacobo et al. (2018)). Such aggregations, however, lose information about mobility patterns of users.

Extracting locations from profile information has also been explored. For the dataset used in Nguyen et al. (2015), my collaborator Dolf Trieschnigg collected a dataset focused on two Dutch provinces (Limburg and Friesland). Users were mapped to locations based on the text provided in the location field in their profiles. However, this turned out to be non-trivial. For example, users who lived in the city of *Leeuwarden*, the provincial capital of Friesland (the Netherlands), provided strings like *Leeuwarden*; *Leeuwarden, The Netherlands*; and *Leeuwarden, Friesland*. However, a long tail of profile locations only occurred once in our data, such as *leeuwarden de gekste* 'leeuwarden the craziest'; *leeuwarden# freeceland*; *LeeuwardenCity (L)*; and *Leeuwarden & Barcelona*.

Sampling In many cases a study involves selecting a sample from the data. Sampling approaches include random sampling, sampling by time period, by individuals/group, event, or by convenience (Herring, 2004). To analyze geographical variation in language using computational approaches, some studies have selected tweets or users based on geotags (Eisenstein et al., 2010), based on profile information (Nguyen et al., 2015), by searching for specific key words (Doyle, 2014; Jones, 2015), and by searching for specific hash tags (Shoemark et al., 2017). However, sampling approaches can introduce biases. For example, Pavalanathan and Eisenstein (2015b) found that GPS-tagged tweets were more often written by young people and women, in comparison to tweets with self-reported locations.

Operationalization of concepts The use of social media sometimes requires adapting operationalizations of concepts developed on other domains. An example is the concept of '*audience*'. The influence of audience on a speaker's style has been widely studied in sociolinguistics, for example using the framework of audience design (Bell, 1984). We tend to speak differently when talking to our boss as opposed to when talking to a close friend. However, in many social media platforms, e.g., Twitter, multiple audiences (e.g., friends, colleagues) are collapsed into a single context. While the audience is potentially limitless, users do often imagine an audience when writing messages and they may target messages to different audiences (Marwick and boyd, 2011). This means that when we want to apply a framework such as audience design to a social media context, we need to rethink how we operationalize audience (see Androutsopoulos (2014); Nguyen et al. (2015); Pavalanathan and Eisenstein (2015a)). For example, studies on Twitter and audience used the presence of hashtags and user mentions as proxies for the target audience (Nguyen et al. (2015); Pavalanathan and Eisenstein (2015a)). Messages with hashtags were assumed to target a broader audience, while messages with user mentions were assumed to target smaller audiences.

As another example, the units of analysis in social media do not always correspond to traditional units of analysis. For instance, posts are not one-to-one comparable to utterances or turns (Androutsopoulos, 2013). Similarly, looking at code-switching patterns may involve different units of analysis compared to studies on spoken data that have focused on analyzing turns or sentences.

Ethical concerns The use of social media also raises various ethical concerns. Social media has been particularly attractive because of its perceived public nature. While many platforms offer a binary choice regarding visibility (public vs. private posts), in reality privacy is not a binary notion but highly contextual and situational (Zook et al., 2017). In other words, there is a difference between what is legal,[5] and norms and expectations regarding privacy and the use of such data. Ethical concerns not only surround the collection of data, but also how such content is quoted in research output (Williams et al., 2017) and how the data is made available to other researchers.

[5] For example, Williams et al. (2017) point out that users consent that their public tweets will be made available to third parties.

10.3 Processing Data

Processing language in social media using automatic tools can be challenging. Many NLP tools have been developed on non-social media data, like newswire texts, and they might work less well on social media. Adapting and or designing NLP tools for social media can require quite some effort. For example, taggers designed for social media often include special tags for hashtags, @-mentions, and emoticons. To illustrate this, Table 10.3 shows two tweets tagged by the ARK Twitter Part-of-Speech (POS) tagger (Owoputi et al., 2013). A 'compound tag' is used to handle cases such as *lemme* ('let me').

Table 10.1

Nice	A	adjective
one	$	numeral
bruv	N	common noun
!	,	interjection
Here	R	adverb
		pre- or postposition, or subordinating
if	P	conjunction
you	O	pronoun
need	V	verb
us	O	pronoun
.	,	punctuation
MK	G	other abbreviations, foreign words, possessive endings, symbols, garbage

Table 10.2

Yes	!	interjection
!	,	punctuation
Lemme	L	nominal + verbal, verbal + nominal
know	V	verb
what	O	pronoun
u	O	pronoun
think	V	verb
:)	E	emoticon
#digitalart	#	hashtag

Table 10.3 *Assigned POS tags by the tool from Owoputi et al. (2013)*

Multilingual social media users sometimes use multiple languages in a single social media post, presenting another challenge to NLP tools. Although most NLP tools assume that the input text is written in a single language, there is an increasing interest in developing NLP tools for code-switched texts (e.g., Bhat et al. (2018)). Fine-grained language identification at the word level (e.g., Nguyen and Doğruöz (2013)) can be a useful step in processing code-switched texts.

Studies have found that the performance of NLP tools can vary based on the socio-demographic background of authors. Hovy and Søgaard (2015) observed performance differences with regard to the age of the authors when the POS tagger was trained on texts from newspapers

and Jørgensen et al. (2015) found that POS taggers are more likely to make mistakes on African-American Vernacular English (AAVE) sentences compared to non-AAVE sentences. Furthermore, Blodgett et al. (2016) found that messages from African-Americans were more likely to be erroneously classified as non-English by automatic language identification systems. Such disparities in performance can have ethical implications: Texts produced by certain social groups may be wrongly analyzed, or even excluded, from a variety of social media analyses.

The difficulty of processing language in social media also affects the analysis of linguistic variation. Computational approaches have mostly focused on lexical (e.g., *pants* vs. *trousers*) and orthographic (e.g., *going* vs. *goin*) variation. Analyzing syntactic variation typically requires the use of a tagger. For example, in a study on African-American English in Twitter, Blodgett et al. (2016) analyzed habitual *be* by tagging tweets with the ARK Twitter POS tagger (Owoputi et al., 2013) and searching for O-be-V and O-b-V sequences. A workaround is to search for lexical patterns instead that instantiate a syntactic variation of interest. For example, Doyle (2014) analyzed the occurrence of *needs done* (need + past participle) and *might could* (double modals). However, this would limit the analysis to specific strings.

10.4 Patterns in Social Media

The use of certain words or grammatical constructions can reveal where someone is from. There is a large body of work on text-based geocoding: Given a text, automatically predict the location of the user or message. These geocoding approaches sometimes identify dialect features (e.g., Rahimi et al. (2017)), but it is usually not their primary aim. For example, toponyms are useful features for these tasks, but they are usually not of interest for research on dialect variation. Work in this area is therefore not discussed here, but the interested reader is referred to Melo and Martins (2016) for an extensive overview.

This section focuses on the analysis of dialect variation. There is a lot of variety in the type of analyses that have been carried out, from analyzing individual linguistic features (e.g., usage of *yinz*) or alternations (e.g., usage of *soda* vs. *pop* vs. *coke*), to automatically discovering dialect regions. Section 10.4.1 discusses how findings from social media data have been compared to more conventional sources. Next, Section 10.4.2 looks at analyzing dialect variation at different linguistic levels.

10.4.1 Comparison against Other Sources

Findings from social media data have been compared against conventional sources in several ways, and so far they generally seem to match them quite well. Small differences are expected, of course, as there are often differences in demographics, the time period of data collection, and the type of variation studied (many studies using conventional sources focus on phonological variation).

One can compare individual patterns in social media against conventional sources. For example, Doyle (2014) studied syntactic patterns in US-geotagged tweets and found high correlations with patterns in the Atlas of North American English and the Harvard Dialect Survey. Individual patterns are sometimes aggregated to discover dialect regions, which can then in turn also be compared against previous sociolinguistic studies. For example, Huang et al. (2016) found that regions identified in Twitter were broadly similar to regions identified in previous studies based on phonetic variation. Some computational methods can be used to automatically identify dialect terms. These terms can also be evaluated by comparing against conventional sources. However, one should keep in mind that these conventional sources might not cover all relevant terms, for example, they might miss dialect terms specific to online language. The neural network approach by Rahimi et al. (2017) enables retrieving the k-nearest terms given the name of a region. They compared the identified terms to dialect terms in the DAREDS dataset, a dataset the authors have created based on the Dictionary of American Regional English (DARE).

10.4.2 Analyzing Variation

Most studies so far have focused on lexical variation or variation that can be captured using lexical patterns. Eisenstein et al. (2010) proposed a topic model that incorporates topics and regions as latent variables to model lexical variation. They found that dialect regions were characterized by various dialect words, locally-specific abbreviations, and named entities. Huang et al. (2016) take an approach that is more common in conventional dialect studies by looking at lexical alternations: different ways of saying the same thing, such as *automobile* vs. *car* and *holiday* vs. *vacation*, so that topic is controlled for. Using statistical testing with methods such as Global Moran's I, they identified lexical alternations that exhibited significant spatial autocorrelation in a large corpus of

geotagged tweets from the US. Rahimi et al. (2017) proposed a neural network approach with one hidden layer to predict the location of a user given tweets. The locations (latitude/longitudes coordinates) are discretized using k-d tree leaf nodes or k-means. The model also learns an embedding of the terms this way, which can be used to detect dialectal terms.

Many studies on dialect are based on speech data and focus on phonological variation. In contrast, with social media the focus has been on written data. Orthographic variation therefore provides an interesting opportunity to bring different strands of research together. Language in social media tends to be closer to spoken language, and Eisenstein (2015) suggests a strong connection between orthographic and phonological variation. He finds that orthographic variation is sensitive to phonological and grammatical contexts and mirrors to some extent patterns in speech. However, the link between orthography and phonology is complex. The pronunciation of a word is not always obvious from its spelling. Jones (2015) gives examples of this and points out that one has to be careful when using written social media data alone to make claims about phonology.

Jones (2015) studied regional patterns in AAVE by analyzing nonstandard spellings on Twitter linked to six phonological phenomena, such as glottal stops and nasal assimilation. He found that the identified dialect regions aligned with patterns of movement during the Great Migrations. Jørgensen et al. (2015) also focused on AAVE. They studied three phonological features based on how they are manifestated as orthographic variations on Twitter (e.g., *brotha* vs *brother*). The orthographic variations were correlated with demographic variables obtained from census data, as well as with geographical variables.

Grammatical variation has received less attention so far, possibly because of challenges related to parsing Twitter data. Grammatical variation has been analyzed using POS taggers as well as by searching for specific strings. Stewart (2014) analyzed African American English syntax in US-geotagged tweets. Regular expressions and two different part-of-speech taggers were used to find patterns such as habitual *be* and copula deletion. Doyle (2014) searched for strings like *needs done* (need + past participle) and *might could* (double modals) and found strong correlations with existing dialect sources. Haddican and Johnson (2012) studied regional effects on the English particle verb alternation using both acceptability judgments and a Twitter study. To collect Twitter data, they searched for specific strings (*turn on the light* vs. *turn the*

light on and variations). They found no regional effects in the UK, but they did find trans-Atlantic differences (UK and Ireland vs. US and Canada). Jacobo et al. (2018) analyzed a large French Twitter corpus. Among the variables studied, they looked at the omission of the French negative particle *ne*, which is considered optional in spoken French but obligatory in writing. They found that in the north of France there was a higher use of non-standard language.

There is also geographical variation at the semantic level. Word embeddings, which represent words as low-dimensional vectors (e.g., 100 dimensions), are effective approaches to capture the meaning of words and therefore to study semantic variation. Bamman et al. (2014a) present an extension of the skip-gram model to learn how the meaning of words varies geographically. While the skip-gram model has a single embedding matrix with embedding vectors for each word, the model proposed by Bamman *et al.* learns a global embedding matrix as well as additional matrices for different contexts, which in their study were the geographical states in the United States. These context-specific embeddings capture how the global representation should shift for specific contexts (e.g., when the word is used in Kansas). Based on a large geotagged Twitter corpus, their model learned that *wicked* in Kansas is close to terms like *evil*, *pure* and *gods*. And that in contrast, *wicked* in Massachusetts is most similar to intensifiers like *super*, *ridiculously* and *insanely*.

10.5 Future Outlook

This chapter concludes with discussing several open research directions.

Bottom-up discovery of features So far, most studies have focused on linguistic features that are selected based on intuition, manual inspection, or findings from previous studies on dialect. This is in fact similar to when using questionnaires, for which the researcher has to specify target items beforehand. However, the scale of social media also supports bottom-up *discovery* of linguistic features. Approaches to automatically identify variables that exhibit geographical variation tend to identify many proper nouns (Pavalanathan and Eisenstein, 2015a; Nguyen and Eisenstein, 2017; Rahimi et al., 2017), such as names of cities, regions, and companies. Additional filtering is therefore necessary to find the ones that are meaningful for sociolinguistic analyses. The next step would be identifying alternations. For example, Shoemark et al. (2017) manually

selected variables for which users can produce either a Standard English or Scottish variant. Combining methods to identify variables that exhibit geographical variation (e.g., Nguyen and Eisenstein (2017)) with methods to identify lexical variants (e.g., Gouws et al. (2011)) could be interesting to explore.

Geographical *and* social factors In both sociolinguistics and dialectology, geographical variation has often been studied separately from social variation (Britain, 2010). However, the integration of social factors, such as socio-demographic variables and social network structures, is increasingly receiving more attention (Kristiansen, 2018; Wieling et al., 2011). Social media affords studying language use in a variety of social settings. The availability of information about social network structures, conversation partners, etc., supports a further integration of social aspects in the study of dialects. Examples of this include work on audience design (Nguyen et al. (2015)) and work on combining socio-demographic factors with geographical variation (Jacobo et al. (2018)). For example, Nguyen et al. (2015) studied the use of two minority languages in the Netherlands. Tweets directed to larger audiences were more often written in Dutch, while within conversations users often switched to the minority languages. Recent studies have looked at the relation between sociolinguistic variation and political views in the context of the Scottish Independence referendum (Shoemark et al., 2017) and in the context of the Catalonian referendum (Stewart et al., 2018). Finally, social factors could also be integrated in analyses on how innovations spread (Eisenstein et al., 2014).

Level of analysis and treatment of place So far, most work in computational linguistics has focused on broad patterns of geographical variation, e.g., across the whole of the Netherlands (Nguyen and Eisenstein, 2017), or across the whole of the US (Eisenstein et al., 2010; Doyle, 2014; Huang et al., 2016). Less work has focused on variation in specific regions or cities. The scale of the data and the fine-grained location information allows us to study geographical patterns —quantitatively— on a detailed level, such as neighborhoods in urban cores. A challenge when zooming in on such levels is that for some fine-grained levels the data might become too sparse in certain areas. Further work could also explore more socially constructed approaches towards space (Johnstone, 2004) (for example, the view that being Texan is culturally defined rather than geographically defined). The study by Cocos and Callison-Burch

(2017) is a first step in this direction. They explore modeling language with respect to attributes of a location (e.g., residential landuse, movie theater) instead of absolute physical locations, and use this as context when training word embeddings.

Dialect perception Perception studies make up a core part of sociolinguistic research, but using computational methods to study the social values that people place on linguistic forms is an underexplored area. Rymes and Leone-Pizzighella (2018) motivate that Web 2.0 enables studying processes through which linguistic forms gain social value and demonstrate this with a qualitative analysis of YouTube comments of videos taking part in the accent competition. The scale of social media allows studying such processes quantitatively over time. Work in this space could also draw from recent studies on fairness and how computational methods encode social biases (Garg et al., 2018).

Variation *and* change Social media also provides the opportunity to study language change across space and time. For example, Eisenstein et al. (2014) analyzed patterns in diffusion of linguistic change over the United States. Geographical proximity and population size were important factors, but the study also found that demographic similarity (especially with regard to race) played a central role. There has been more work on language change and social media, e.g., Grieve et al. (2017) studied the emergence of new words, but this study only focused on the diachronic dimension. A challenge is teasing apart true language change from confounding factors, especially since social media data is generated in an uncontrolled setting. For example, the population of a social media platform need not stay the same, e.g., younger users might migrate to another platform over time.

Multiple platforms, multiple modalities Social media studies usually focus on a *single* data source, e.g., Twitter, Facebook or YouTube. However, different social media platforms have different mechanisms shaping language use and behavior. Research that compares patterns across different social media platforms would thus support more robust interpretations of the findings and help us answer questions about generalizability. Furthermore, social media also allows us to extend our focus to other modalities, like speech or video, which could shed further light on the relation between phonological and orthographic variation.

Acknowledgements

This work was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1 through an Alan Turing Institute Fellowship (TU/A/000006). I would like to thank Bo Wang and Maria Liakata for help with accessing the Twitter data.

References

Androutsopoulos, Jannis. 2013. Online data collection. Pages 236–249 of: Mallinson, Christine, Childs, Becky, and Herk, Gerard Van (eds), *Data Collection in Sociolinguistics: Methods and Applications*. Routledge.

Androutsopoulos, Jannis. 2014. Languaging when contexts collapse: Audience design in social networking. *Discourse, Context & Media*, **4–5**, 62 – 73.

Bamman, David, Dyer, Chris, and Smith, Noah A. 2014a. Distributed representations of geographically situated language. Pages 828–834 of: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

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

Bell, Allan. 1984. Language style as audience design. *Language in Society*, **13**(2), 145–204.

Bhat, Irshad, Bhat, Riyaz A., Shrivastava, Manish, and Sharma, Dipti. 2018. Universal dependency parsing for Hindi-English code-switching. Pages 987–998 of: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

Blodgett, Su Lin, Green, Lisa, and O'Connor, Brendan. 2016. Demographic dialectal variation in social media: A case study of African-American English. Pages 1119–1130 of: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Britain, David. 2010. *An International Handbook of Linguistic Variation*. Handbooks of Linguistics and Communication Science, no. 30. Berlin: Mouton de Gruyter. Chap. Language and space: The variationist approach, pages 142–163.

Chambers, Jack K., and Trudgill, Peter. 1998. *Dialectology*. Cambridge University Press.

Cheshire, Jenny, and Fox, Sue. 2009. Was/were variation: A perspective from London. *Language Variation and Change*, **21**(1), 1–38.

Cocos, Anne, and Callison-Burch, Chris. 2017. The language of place: Semantic value from geospatial context. Pages 99–104 of: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*.

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

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

Eisenstein, Jacob, O'Connor, Brendan, Smith, Noah A., and Xing, Eric P. 2010. A latent variable model for geographic lexical variation. Pages 1277–1287 of: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*.

Eisenstein, Jacob, O'Connor, Brendan, Smith, Noah A., and Xing, Eric P. 2014. Diffusion of lexical change in social media. *PLoS ONE*, **9**(11), e113114.

Garg, Nikhil, Schiebinger, Londa, Jurafsky, Dan, and Zou, James. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*.

Golder, Scott A., and Macy, Michael W. 2014. Digital footprints: Opportunities and challenges for online social research. *Annual Review of Sociology*, **40**(1), 129–152.

Gouws, Stephan, Hovy, Dirk, and Metzler, Donald. 2011. Unsupervised mining of lexical variants from noisy text. Pages 82–90 of: *Proceedings of the First workshop on Unsupervised Learning in NLP*.

Grieve, Jack, Nini, Andrea, and Guo, Diansheng. 2017. Analyzing lexical emergence in Modern American English online. *English Language and Linguistics*, **21**(1), 99–127.

Haddican, Bill, and Johnson, Daniel Ezra. 2012. Effects on the particle verb alternation across English dialects. *University of Pennsylvania Working Papers in Linguistics*, **18**(2), 5.

Herring, Susan C. 2004. *Designing for virtual communities in the service of learning*. Vol. 338–376. New York: Cambridge University Press. Chap. Computer-mediated discourse analysis: An approach to researching online behavior.

Herring, Susan C, and Androutsopoulos, Jannis. 2015. *The Handbook of Discourse Analysis*. Chichester: John Wiley & Sons. Chap. Computer-mediated discourse 2.0, pages 127–151.

Hovy, Dirk, and Søgaard, Anders. 2015. Tagging performance correlates with author age. Pages 483–488 of: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*.

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*, **59**, 244 – 255.

Jacobo, Levy Abitbol, Karsai, Márton, Magué, Jean-Philippe, Chevrot, Jean-Pierre, and Fleury, Eric. 2018. Socioeconomic dependencies of linguistic patterns in Twitter: a multivariate analysis. Pages 1125–1134 of: *Proceedings of the 2018 World Wide Web Conference WWW '18*.

Johnstone, Barbara. 2004. Place, globalization, and linguistic variation. *Sociolinguistic variation: Critical reflections*, 65–83.

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

Jørgensen, Anna Katrine, Hovy, Dirk, and Søgaard, Anders. 2015. Challenges of studying and processing dialects in social media. Pages 9–18 of: *Proceedings of the Workshop on Noisy User-generated Text*.

Kerswill, Paul. 2013. *Space in language and linguistics: Geographical, interactional, and cognitive perspectives*. Walter de Gruyter. Chap. Identity, ethnicity and place: the construction of youth language in London, pages 128–164.

Kristiansen, Tore. 2018. *The Handbook of Dialectology*. Wiley-Blackwell. Chap. Sociodialectology, pages 106–122.

Llamas, Carmen. 2018. *The Handbook of Dialectology*. Wiley-Blackwell. Chap. The dialect questionnaire, pages 253–267.

Marwick, Alice E., and boyd, danah. 2011. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, **13**(1), 114–133.

Melo, Fernando, and Martins, Bruno. 2016. Automated geocoding of textual documents: A survey of current approaches. *Transactions in GIS*, **21**(1), 3–38.

Nguyen, Dong, and Doğruöz, A. Seza. 2013. Word level language identification in online multilingual communication. Pages 857–862 of: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*.

Nguyen, Dong, and Eisenstein, Jacob. 2017. A kernel independence test for geographical language variation. *Computational Linguistics*, **43**(3), 567–592.

Nguyen, Dong, Gravel, Rilana, Trieschnigg, Dolf, and Meder, Theo. 2013. "How old do you think I am?" A study of language and age in Twitter. Pages 439–448 of: *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*.

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

Nguyen, Dong, Doğruöz, A. Seza, Rosé, Carolyn P., and de Jong, Franciska. 2016. Computational sociolinguistics: A survey. *Computational Linguistics*, **42**(3), 537–593.

Owoputi, Olutobi, O'Connor, Brendan, Dyer, Chris, Gimpel, Kevin, Schneider, Nathan, and Smith, Noah A. 2013. Improved part-of-speech tagging for online conversational text with word clusters. Pages 380–390 of: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

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

Pavalanathan, Umashanthi, and Eisenstein, Jacob. 2015b. Confounds and consequences in geotagged Twitter data. Pages 2138–2148 of: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Rahimi, Afshin, Cohn, Trevor, and Baldwin, Timothy. 2017. A neural model for user geolocation and lexical dialectology. Pages 209–216 of: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Rymes, Betsy, and Leone-Pizzighella, Andrea. 2018. YouTube-based accent challenge narratives: Web 2.0 as a context for studying the social value of accent. *International Journal of the Sociology of Language*, **2018**(250), 137–163.

Salganik, Matthew J. 2017. *Bit by bit: social research in the digital age*. Princeton University Press.

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

Stewart, Ian. 2014. Now we stronger than ever: African-American English syntax in Twitter. Pages 31–37 of: *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*.

Stewart, Ian, Pinter, Yuval, and Eisenstein, Jacob. 2018. Sí o no, què penses? Catalonian independence and linguistic identity on social media. In: *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Wang, Bo, Liakata, Maria, Zubiaga, Arkaitz, and Procter, Rob. 2017a. TDParse: Multi-target-specific sentiment recognition on Twitter. Pages 483–493 of: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*.

Wang, Bo, Liakata, Maria, Tsakalidis, Adam, Georgakopoulos Kolaitis, Spiros, Papadopoulos, Symeon, Apostolidis, Lazaros, Zubiaga, Arkaitz, Procter, Rob, and Kompatsiaris, Yiannis. 2017b. TOTEMSS: Topic-based, Temporal Sentiment Summarisation for Twitter. Pages 21–24 of: *Proceedings of the IJCNLP 2017, System Demonstrations*.

Wieling, Martijn, Nerbonne, John, and Baayen, R. Harald. 2011. Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PLOS ONE*, **6**(9), 1–14.

Williams, Matthew L, Burnap, Pete, and Sloan, Luke. 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.

Zook, Matthew, Barocas, Solon, boyd, danah, Crawford, Kate, Keller, Emily, Gangadharan, Seeta Peña, Goodman, Alyssa, Hollander, Rachelle, Koenig, Barbara A., Metcalf, Jacob, Narayanan, Arvind, Nelson, Alondra, and Pasquale, Frank. 2017. Ten simple rules for responsible big data research. *PLOS Computational Biology*, **13**(3), 1–10.