Localising morphosyntactic variation in Twitter data

David Willis, Deepthi Gopal, and Tam Blaxter, University of Cambridge
dwew2@cam.ac.uk; dg537@cam.ac.uk; ttb26@cam.ac.uk

Twitter offers us, in principle, unprecedented amounts of relatively-accessible data on real-time natural language use. This abundance of information should then allow us to examine the propagation of morphosyntactic variants across the population in much finer detail than is straightforwardly possible using traditional variationist and dialectological methods; our intent in this paper is to demonstrate that this can be done, but that new methods of localisation are required before the spatial component of social-media data can be considered adequately reliable.

Previous social-media studies (see eg. for American English: Russ 2012; Bamman et al. 2014; Doyle 2012; Eisenstein et al. 2014) make direct use of the automated language identifications and geolocation tags that are provided with raw Twitter data. This severely limits the possible scope of study, for several reasons. Twitter's geographic metadata are only provided when a user explicitly opts in, and as such are available for an extremely small fraction of the whole (representing less than two percent of users, per Leetaru et al. 2013); from the point of view of the researcher, it is unclear whether this additionally implies the existence of significant statistical bias distinguishing those users responsible for location-tagged text from the population at large. Automatically-generated metadata correspond solely to the location of an individual at the moment of posting, which does not imply any certainty with respect either to their habitual place of residence, or (if distinct) their ultimate place of origin. Consider then the problem of language identification: without recourse to tagged coordinates, geographically-separated varieties of the same language, such as American English and British English, cannot be automatically differentiated from one another; but we cannot trust that coordinates placed within the British Isles imply a speaker of a British variety.

Our results employ two corpora of Tweets collected between October 2017 and April 2018, intended to represent different approaches to the identification and localisation of British and Irish English tweets. The first corpus uses the geolocation tags supplied by Twitter without further modification, and consists of all posts for which these coordinates are located within the UK and Ireland ($\sim 5,000,000$/month). The second corpus consists of the full timelines of approximately 1 million users, whose predicted location has been algorithmically determined based largely on the presence of gazetteer place-names within users' self-identified locations, biographical information, and post text. We demonstrate, for a range of British and Irish English variables known from traditional datasets—*was-were* levelling eg. *I was/I were*, *you/youse*, *amn't/aren't*, *X's not/X isn't*, and the dative alternation in *give it me/give me it/give it to me*—that the algorithmically-localised dataset appears to behave significantly more like a traditional dialectological dataset. Clear regional patterns of variation can be mapped therein, but are frequently obscured in the automatically-geocoded dataset by patterns of short-term movement and medium-term migration.

References

Bamman, David, Jacob Eisenstein & Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics* 18.

Doyle, Gabriel. 2012. Mapping dialectal variation by querying social media. Paper presented at the 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden. http://web.stanford.edu/ gdoyle/papers/doyle-2014-eacl.pdf (1 August 2016).

Eisenstein, Jacob, Brendan O'Connor, Noah A. Smith & Eric P. Xing. 2014. Diffusion of lexical change in social media. *PLoS ONE* 9(11). 0113114. doi: 10.1371/journal.pone.0113114.

Leetaru, Kalev, Shaowen Wang, Guofeng Cao, Anand Padmanabhan & Eric Shook. 2013. Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday* 18(5).

Russ, Brice. 2012. Examining large-scale regional variation through online geotagged corpora. Paper presented at the Annual Meeting of the American Dialect Society, Portland. http://www.briceruss.com/ADStalk.pdf.