Regional Variation of General Extender Usage in a Geo-tagged Microblog Corpus of American English

Joseph Kessler (*University at Buffalo*)

The usage of a class of English expressions known as general extenders (GEs), which includes items such as “and stuff” and “or whatever,” has been found to vary by features including speakers’ social class (Cheshire 2007), age (Tagliamonte and Denis 2010), and gender (Levey 2012). However, no study to date has explored the issue of geographic variation in GE usage beyond the general finding that different GEs are preferred in different English-speaking countries (Pichler and Levey 2011). Thus, although these expressions are known to vary across populations, the literature on that variation is somewhat disconnected from research on American English dialect regions. The present study aims to fill this gap with the first quantitative research into the regional variation of general extender usage, as represented in a corpus of electronically-mediated communication (EMC) messages that have been tagged with each user’s geographic coordinates at the time of writing. EMC is a largely unedited, generally informal mode of written communication with many vernacular and speech-like qualities (Tagliamonte and Denis 2008, Baron 2008, etc.), and the present study adds to a growing body of literature exploring how regional variation of language can be reflected and modeled with data from EMC (Eisenstein et al. 2010, Russ 2012, etc.). For this study, an American English corpus of roughly 380,000 public messages posted to the EMC microblogging website Twitter in March 2010 and geotagged with each author's latitude and longitude at the time of writing has been searched for a number of general extenders that have been previously attested in the literature. This corpus, originally compiled by Eisenstein et al. (2010), has been used in a number of follow-up studies (Wing and Baldridge 2011, Yuan et al. 2013, etc.), and thus represents a useful benchmark for analyzing regional variation in a microblogging context. Eleven general extenders with more than 10 tokens each have been found in this corpus, for a total of 786 individual instances. These eleven expressions are found to be used at differential rates, with “and shit” and “etc” as the most commonly used types, despite both being relatively infrequent in previous corpus research involving American English telephone conversations (Overstreet and Yule 1997) and instant-messaging interactions (Fernandez and Yuldashev 2011). This finding may suggest that speakers have different general extender preferences across disparate modes and genres. The geotagging metadata indicate that the eleven GE types in the microblogging corpus display differential geographic distributions that do not fall into the traditional dialect regions of American English (Wolfram and Schilling-Estes 2006), although contemporary research indicates that individual linguistic variables often show regional patterns that diverge from such larger trends (Grieve 2014, etc.). The most significant regional association found in this study links the general extender “and all that” to New York City, thereby representing a previously undocumented feature of that dialect region. This study demonstrates the potential strengths in applying this methodology to the issue of variable general extender usage, as the patterns it reveals may be harder to detect through more traditional methods.

References

Baron, N. S. (2008). *Always on: Language in an online and mobile world*. New York, NY: Oxford University Press.

Cheshire, J. (2007). Discourse variation, grammaticalisation and stuff like that. *Journal of Sociolinguistics, 11*(2), 155–193.

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2010). A latent variable model for geographic lexical variation. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 1277-1287.

Fernandez, J., & Yuldashev, A. (2011). Variation in the use of general extenders and stuff in instant messaging interactions. *Journal of Pragmatics, 43*(10), 2610–2626.

Grieve, J. (2014). A comparison of statistical methods for the aggregation of regional linguistic variation. In B. Szmrecsanyi & B. Wälchli (Eds.), *Aggregating dialectology, typology, and register analysis: Linguistic variation in text and speech* (53-88). Berlin/New York: Walter de Gruyter.

Levey, S. (2012). General extenders and grammaticalization: Insights from London preadolescents. *Applied Linguistics, 33*(3), 257-281.

Overstreet, M., & Yule, G. (1997). On being inexplicit and stuff in contemporary American English. *Journal of English Linguistics, 25*(3), 250-8.

Pichler, H., & Levey, S. (2011). In search of grammaticalization in synchronic dialect data: General extenders in northeast England. *English Language and Linguistics, 15*(3), 441–471.

Russ, B. (2012). Examining large-scale regional variation through online geotagged corpora. Presented at the 2012 American Dialect Society Annual Meeting.

Tagliamonte, S. A., & Denis, D. (2008). Linguistic ruin? Lol! Instant messaging and teen language. *American Speech, 83*(1), 3–34.

Tagliamonte, S. A., & Denis, D. (2010). The stuff of change: General extenders in Toronto, Canada. *Journal of English Linguistics, 38*(4), 335-368.

Wing, B. P., & Baldridge. J. (2011). Simple supervised document geolocation with geodesic grids. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, 955-964.

Wolfram, W., & Schilling-Estes, N. (2006). *American English: Dialects and variation* (2nd ed.). Malden, MA: Blackwell.

Yuan, Q., Cong, G., Ma, Z., Sun, A., & Magnenat-Thalmann, N. (2013). Who, where, when and what: Discover spatio-temporal topics for Twitter users. Proceedings of the 19th ACM SIGKDD International Conference on Knowledge, Discovery and Data Mining.