Journal of Linguistic Geography (2017) **4**, 97–100. © Cambridge University Press 2017
doi:10.1017/jlg.2016.14

BOOK REVIEW

Last, half of Upper Midwesterners ought to get into this book: Written Grammatical Variation in the US

Regional Variation in Written American English
By Jack Grieve
Cambridge: Cambridge University Press, 2016. Pp. xi–335.
ISBN 978-1-107-03247-7; $110.00 cloth ($88.00 e-book)

Reviewed by
Kelly Abrams, *University of Wisconsin-Madison*
Thomas Purnell, *University of Wisconsin-Madison*

It is now commonplace for big data scholars to push methodological boundaries; the field of dialectology is no exception, with authors in this journal tapping into advanced technological and statistical methods. One example of big data dialectology is Jack Grieve's *Regional Variation in Written American English*. The seven chapters and five appendices deliver maps of 135 grammatical features of written American English gleaned from just under 37 million words from letters to the editor in newspapers in 240 city areas over a 13 year period at the beginning of the 21st century. This work presents readers with a thorough and meticulous account of Grieve's statistical mapping methods, as well as a simplified map for each grammatical variable represented in three different ways (raw values, spatial autocorrelation analysis, and multivariate factor analysis). Grieve outlines the gaps in research currently: grammatical variation, the status of the Midland dialect region, data since the turn of the century, written variation, and differences in register—arguing that this book fills those gaps in research. This work goes beyond the author's 2009 dissertation and 2001 paper that appeared in this journal. It is a volume well-suited for linguists interested in dialectology, linguistic geography, and grammatical variation.

Presenting a short history and overview of the three major large-scale dialectology projects and their influence in defining American dialects, the first chapter, "Introduction," places Grieve's work squarely within the field of American English dialectology. The three projects discussed are the *Linguistic Atlas of the United States* (Kurath et al., 1939-1943), the *Dictionary of American Regional English* (Cassidy & Hall, 1985-2013), and the *Atlas of North American English* (Labov, Ash, & Boberg, 2006). Kurath's work first used bundles of isoglosses plotted on maps to show the dialect boundaries of New England—a method that dominated the field for decades. In subsequent years, the *Atlas* project was expanded with more publications, ultimately presenting three major dialect regions based on lexical variation in the US: the North, Midland, and South. Kurath argued his regions were based on historical settlement patterns, a theory that would eclipse all others in dialectology for decades. Carver (1987) uses *DARE* data to identify two main dialect regions—the North and South—with a few sub-regions in each. Finally, Labov, Ash, and Boberg (2006) use new dialectology methods focusing on cities rather than rural areas, diverse sampling to get additional social demographics, and acoustic analysis of vowel formants. They identified four major dialect regions—the North, the Midland, the South, and the West, as well as a distinct Canadian region—while suggesting the regions were formed by both historical settlement and linguistic patterns (i.e. chain vowel shifts).

The second chapter, "Corpus," presents arguments for using a corpus-based approach in dialectology: scholars can collect a significantly larger sample without the travel expenses and the effort of finding and interviewing informants, which, Grieve argues, leads to more reliable and generalizable results. Furthermore, grammatical variation is often too abstract or obscure to reliably obtain from informants through direct questioning and may not appear in open-ended sociolinguistic interviews. Thus, Grieve argues, a large enough corpus must be curated to find grammatical variation. In the case of this work, the corpus consists of 211,949 letters to the editor from across the US (in 240 cities) between 2000 and 2013 by 166,083 authors. The letters contain a variety of topics including current events, corrections, and responses to newspaper articles, and public announcements or public thanks to the community. He chose this genre because the geographic placement of individuals is known, it is a genre produced by a wide number and variety of people, and they are freely available. Additionally, the similar format and genre can control for register variation and have similar communicative purposes. One issue with this genre, Grieve notes, is that editors may edit the letters before publication. Grieve had discussions with editorial page editors from various publications and found they are often edited for length, whole passages may be deleted, and "letters are also edited for grammatical, typographical, punctuation, and content errors, but according to editorial page editors, grammatically correct sentences are rarely altered" (p. 21–22). Another potential issue for this corpus is that length of residence—a key question when identifying dialect patterns—is not available and may skew the results.

As to the potential problem that social demographic information was not available, Grieve argues that enough letters were obtained to null this issue: "the approach taken here was to obtain a large and nearly exhaustive sample of letters to the editor from each of the newspapers targeted for representation over a given period of time, which ensure that the demographic background of the population of letter to the editor writers in that city is represented accurately" (p. 26). However, keep in mind that letter to the editor writers are more likely middle or upper class, educated (or at least literate enough to write in English), and may skew towards older individuals (who may have more free time to write letters to editors). Moreover, they must be highly motivated and invested citizens who care enough to write letters, and editors provide a filter on the letters in selection as well as language. The consequence is that the claims are restricted to a highly specific population at the city level; the results are not generalizable to the general population and the absence of social demographic data available prohibits examining grammatical variation of sociolinguistic subgroups.

The third chapter, "Grammatical analysis," discusses the 135 grammatical alternations analyzed (summarized on Table 3.1, p. 48) and provides maps of the 295 variants. Grieve explains his mapping technique: variants are mapped based on percentages present, and the simplified maps indicate the strongest percentage of each variable. This presentation reveals regional variation and pattern recognition more easily than full maps cluttered with all alternations at once. The list of alternations is a good reference but the maps of raw values seem to indicate that the variation is random. Grieve admits: "[they] do not show clear patterns of regional variation. The percentages of variants often vary greatly at adjacent locations and rarely do the higher percentage locations form a single cluster on the map" (p. 98). Before turning to the following chapter, in which the reader sees Grieve's analytic prowess in turning what appears meaningless into some fairly meaningful regional patterns, we note an important didactic force in presenting raw maps. These visualizations are instructive in two ways, the first being that we often encounter linguistic forms appearing entirely unstructured in spite of our assent that variation reflects systematic homogeneity (Weinreich, Labov & Herzog, 1968); until we learn to weight the data, everything appears to be random. Second, the raw data is consistent with the zeitgeist of data honesty and full disclosure; without the inclusion of the raw maps, readers would be left wondering about the workings under the hood, leaving transformed results unconnected to the vast amounts of input data.

The fourth chapter, "Spatial analysis," initially discusses the limitations of traditional methodology in dialectology: (1) isoglosses are drawn by hand based on the judgment of the dialectologist, who estimates the approximate location where an alternation transitions between its variants, making replication difficult; (2) plotting isoglosses by hand is time-consuming, making large-scale projects more difficult; and (3) since an isogloss is a line, it does not represent gradual change. Thus, Grieve presents a method known as local spatial autocorrelation analysis, which identifies patterns of spatial clustering using a standardized and statistically grounded procedure. Unlike the limitations of isogloss assignment of earlier days, this method allows statistical software to conduct the analysis, thereby allowing for replicability. A score is assigned to each location so variation can be determined based on whether the variant is in the middle of many other locations with a high frequency of the same variant (a high G_i score, represented by a black dot on the maps) or with a low frequency of the same variant (a low G_i score; represented by a white dot). All of the maps are presented in figure 4.7 (pp. 123-140) with some clear patterns of regional distribution. Two conclusions are drawn. First, despite unclear variation in the raw maps, the majority of variants do show spatial clustering. Second, the patterns are diverse and complex, revealing regional grammatical variation in this register of written Standard Modern American English, although Grieve admits it is relatively weak.

The fifth chapter, "Multivariate analysis," discusses the methodology and three stages of a multivariate spatial analysis:

> First, individual patterns of spatial clustering are identified in the maps for each individual linguistic variable based on a local spatial autocorrelation analysis.… Second, sets of linguistic variables that exhibit similar patterns of spatial clustering are identified and mapped based on a factor analysis of the local spatial autocorrelation maps. Third, dialect regions are identified and mapped based on a cluster analysis of the factor maps (p. 147).

Then exploratory factor analysis, a statistical process that is used to identify patterns of variation is applied to the data. Grieve identifies three factors for this data and presents the results: "overall, the final factor analysis accounts for 60.3% of the variance in the spatially autocorrelated linguistic data matrix, with Factor 1 accounting for 26.3% of the variance, Factor 2 accounting for 20.9% of the variance and Factor 3 accounting for 13.1% of the variance" (p. 166).

The sixth chapter, "Sources of regional linguistic variation" identifies the five dialect regions of the United States based on the aforementioned factor analysis: the Northeast, the Southeast, the Midwest, the South Central, and the West. Factor 1 differentiates the East from the West and includes 36 alternations; the strongest variants in the factor are: *do not* full vs. contraction, pronoun *have* full vs. contraction, and *may* vs. *might.*

https://doi.org/10.1017/jlg.2016.14 Published online by Cambridge University Press

Many of the variants on this factor have to do with formality of language, with the more full, formal, and complex forms (i.e. *therefore* has two syllables and *thus* has one) being found in the East and the contracted, less formal, and less complex forms being found in the West. Factor 2 includes 28 alternations that distinguish the South Central from the Northeast and the West Coast. This also includes a formality distinction, particularly related to academic and professional forms of writing, with the South Central preferring the less formal variants. The top alternations include pronoun/noun alternation and predictive/attributive adjective alternation, both of which refer to the density of noun phrases, with the shorter noun phrases preferred in the South Central. Factor 3 includes 19 alternations that differentiate the Midwest from the rest of the US, particularly the Southeast. In this set, there are many variants that differ between prescriptively correct and incorrect forms, with the Midwest preferring the more formal, prescriptive forms (reflecting Preston 1989: 59). The Midwest prefers the non-*by*-passive (vs. the *by*-passive), *because* rather than *since*, and *previous* rather than *prior*.

Grieve then assigns an overall formality score for each of the regions by combining the three factors. The Northeast region is the most formal, followed by a relative high formality in the Midwest. The South Central, Southeast, and West all have relatively low formality scores, although with differing degrees for different variables. He contends the following results: "Overall, the analysis therefore not only identifies three different patterns of *regional* variation but three different patterns of *linguistic* variation, all of which are related to formality. The analysis also shows more generally that formality is not a simple construct but rather that there are different types of formality, each of which is associated with a different regional pattern in American English" (p. 196). Grieve notes these findings show that formality is a significant internal correlate of regional variation in grammar, challenging Labov (2013) who maintained that grammatical variation is stable and that lexical variation is arbitrary. Grieve states that all three factors can also be explained by external factors (e.g. physical geography, cultural divides).

Although he had similar findings with earlier dialectology studies, particularly with the South and the West dialect regions, Grieve argues that his Northeast and Midwest regions are different. He prefers to divide things east-to-west (with Northeastern and Midwestern) rather than north-to-south (with Northern and Midland). Thus, settlement patterns cannot explain this difference as New England and Mid-Atlantic people settled both areas. He argues, then, that this can be accounted for with physical and cultural boundaries, especially the Appalachian Mountains. Finally, Grieve reanalyzes data from ANAE using the same methods of this study in order to provide a clearer chronology and link his results to earlier studies, arguing that Carver's analysis was correct: "A distinction between the Northeast and the Midwest has been taking its place in the second half of the twentieth century, which by the early twenty-first century appears to have become well-established" (p. 215).

The seventh and final chapter, "Conclusion," summarizes the results of the book:

> Regional linguistic variation is far more complex and pervasive than has previously been assumed, existing across registers and linguistic levels. Given these results, it is clear that regional variation is not a marginal phenomenon only operating on the edges of language where the environment is particularly suitable for its development and maintenance; rather, regional linguistic variation appears to be a general property of natural language. (p. 220)

Grieve notes that his study provides new methods in the following areas: (1) data collection: the first American dialect study that has been based on a corpus of naturally occurring written language produced in real communicative situations without linguist intervention; and (2) data analysis: a new approach to statistical analysis of regional linguistic variation using multivariate spatial analysis.

One test of a book is to consider specific ways in which it could add to one's own research, teaching and service. Our specific research interests led us to ask this of Grieve's book: what do we learn about the features and extent of the boundaries of Upper Midwestern English (UME) and the three major, yet narrow, regions internal to Maryland? Reviewing the transformed maps, we observe that, in addition to the Midwest forms mentioned above, the narrower UME region is somewhat uniquely identified (black dots most occurring in the Minnesota, Wisconsin region) by the determiner *half of*, the prepositions *no matter* and *get into*, the modal *ought to* and some other features shared with other regions (e.g., the adverbial *last* instead of *lastly*). The cross factor figures (5.6 to 5.11) inform the overall map Grieve provides for UME (figure 5.13, p. 180), splitting the Midwestern region into three subgroups: a core UME region of Iowa, Minnesota and Wisconsin; a western subregion of Kansas, Nebraska, and North and South Dakota; and an eastern subregion of Illinois, Indiana, and Michigan.

Turning to our other region of interest, we see that the overall regional map sadly positions all of Maryland within the Northern region and does not plot Appalachian English as a region at all; although, in defense of this model, much of Appalachia's core (e.g., West Virginia) and the Ozark region separating

https://doi.org/10.1017/jlg.2016.14 Published online by Cambridge University Press

the South and the Midwest regions of the US are free of data points, and Grieve acknowledges that transitional ("intermediate") points are not plotted (p. 182). The three regions in Maryland (Appalachia, central or middle Maryland, and the Eastern Shore) are difficult to identify from the printed maps, partly because of the maps' resolution (18 per page) and partly due to what Kurath and McDavid (1961) noted about Maryland, that many features are shared with some surrounding dialect. This generalization seems borne out (aided by a bit of wishful squinting) by a number of features that appear to stop at a Maryland boundary. One feature, the sentence initial *and,* has a cluster in Maryland (and a few points out west) that does not include the rest of the Northern region. Because these are computer-generated maps, having them online would be a boon for examining dialect overlap in Maryland in more detail.

In short, Grieve's analysis of these two regional interests compelled us to ask questions about our own work and the social domains in which his findings are applicable or not. This information has the potential to inform our research, stimulate great in-class discussion at both the undergraduate and graduate level, and drive students to formulate their own line of inquiry about UME and Maryland English. In these ways, the effect of *Regional Variation in Written American English*—while paralleling that of other national-level reference works (*Linguistic Atlas*, *DARE*, *ANAE*)—appears as if it will most likely sit on our shelves next to Carver (1987).

References

Carver, Craig M. 1987. *American regional dialects: A word geography*. Ann Arbor, MI: University of Michigan Press.

Cassidy, Frederic G. & Hall, Joan H. 1985–2013. *Dictionary of American Regional English*. Cambridge, MA: Harvard University Press.

Kurath, Hans, Hansen, Marcus L., Bloch, Bernard, & Bloch, Julia. 1939. *Linguistic Atlas of New England*. Providence, RI: Brown University Press.

Kurath, Hans. & McDavid, Raven. 1961. *The Pronunciation of English in the Atlantic States: Based upon the Collections of the Linguistic Atlas of the Eastern United States*. Ann Arbor: University of Michigan Press.

Labov, William., Ash, Sharon & Boberg, Charles. 2006. *Atlas of North American English*. Berlin: Mouton.

Preston, Dennis. 1989. *Perceptual dialectology: Nonlinguists' views of areal linguistics*. Berlin: Walter de Gruyter.

Weinreich, Uriel., Labov, William., & Herzog, Marvin I. 1968. *Empirical foundations for a theory of language change*. Austin: University of Texas Press.

https://doi.org/10.1017/jlg.2016.14 Published online by Cambridge University Press