Site-Restricted Web Searches for Data Collection in Regional Dialectology

Abstract

This paper presents a new method for data collection in regional dialectology based on site-restricted web searches. The method allows for the values of many lexical alternation variables to be measured across a region of interest using common search engines such as Google or Bing. The method involves estimating the proportions of the variants of a lexical alternation variable over a series of cities by counting the number of webpages that contain these variants on newspaper websites originating from these cities through site-restricted web searches. The method is evaluated by mapping the 26 variants of 10 content word alternation variables with known distributions in American English. In almost all cases, the maps based on site-restricted web searches align closely with traditional dialect maps based on data gathered through questionnaires, demonstrating the accuracy of this method for the observation of regional linguistic variation. However, unlike collecting dialect data using traditional methods, which is a relatively slow process, the use of site-restricted web searches allows for dialect data to be collected from across a region as large as the United States in a matter of days.

1. Introduction

Regional dialect studies are generally based on language collected by surveying informants from across a region of interest. For example, most North American dialect studies have been based on language collected through questionnaires (e.g. Kurath, 1949; Warkentyne, 1974; Cassidy, 1985; Chambers, 1994; Vaux, 2003; Boberg, 2005; Labov et al., 2006) and sociolinguistic interviews (e.g. Labov et al., 2006). Although these methods have a long history of use in dialectology, eliciting language data from individual informants is a slow process, especially across a relatively large region. This is why data collection for all of the major North American dialect surveys has taken anywhere from a year to decades to complete, and why in most cases only a small number of informants were observed at each location. Recently, corpora of natural language have also been used as the basis of regional dialect studies (e.g. Grieve, 2011; Elspaß et al., 2007), but compiling corpora is still a relatively slow process and it is difficult to collect sufficient amounts of natural language from a sufficient number of locations to observe many lexical alternations. While all of these approaches to data collection are valid and have been successfully applied, none allow for data to be gathered quickly. Research on regional linguistic variation has therefore progressed relatively slowly compared to research on other forms of linguistic variation, including social and situational variation, for which data can more easily be obtained.

This paper presents a new method for data collection in regional dialectology that is based on site-restricted web searches (SRWSs). This method can be used to quickly measure regional variation in the values of many lexical alternation variables using common search engines such as Google or Bing. The basic method involves measuring the proportions of the variants of an alternation variable over a series of cities by counting the number of webpages in which these variants appear on websites originating from these cities using SRWSs. For example, the alternation between *sneakers* and other synonymous forms (e.g. *tennis shoes, running shoes, gym shoes*) could be measured in Anniston, Alabama by counting the number of webpages on the *annistonstar.com* newspaper website that contain

each of these variants based on the results of a series of SRWSs made on Google. The proportion of *sneakers* in Anniston newspaper writing would then be calculated by dividing the number of hits for *sneakers* by the total number of hits for *sneakers* and the other synonymous forms. This process would then be repeated for newspapers in many cities from across the United States to identify patterns of regional linguistic variation in the values of this alternation variable.

The goal of this paper is to introduce and evaluate this method for the observation of regional lexical variation based on an analysis of regional lexical variation in American newspaper writing. The method is first introduced through a detailed analysis of the alternation between the words *sneakers, tennis shoes, running shoes,* and *gym shoes*, including a discussion of the selection of newspaper websites and the use of the local spatial autocorrelation statistic Getis-Ord *Gi* to identify underlying patterns of regional variation in the map for each variant. The method is then evaluated by mapping ten lexical alternation variables with well-established patterns of regional variation in American English and by comparing these maps to the corresponding maps from the *Harvard Dialect Survey* (Vaux, 2003).

2. Newspaper Selection

The basic method for data collection being introduced here involves estimating the values of a linguistic alternation variable across a series of locations based on web searches that are restricted to websites originating from those locations. In this study, this method is evaluated by mapping content word alternations with known distributions in American English through web searches that are restricted to websites for city newspapers from across the United States.

In order to access American newspaper websites, a list of over 2,000 newspapers was taken from the website *refdesk.com,* along with the city, state and URL associated with each of these newspapers. This particular newspaper index was selected because it was well organized and simply

designed, which facilitated data harvesting, and because it appeared to list a relatively large number of newspaper websites compared to similar websites. After the list of websites was compiled, the *www.* prefix was stripped from each URL to allow for additional URLs associated with the newspapers to be accessed through SRWSs (e.g. allowing for *topics.nytimes.com* to be searched in addition to *www.nytimes.com*). Each URL was then tested online and approximately half of the URLs were discarded because they were inactive or because they were not associated with a sizable number of webpages. Business, entertainment, and university newspapers were also deleted from the list in order to focus the analysis on the typical city newspaper register of American English. The list was then checked by hand to see if the largest cities and most popular newspapers in the United States were represented. If a city or newspaper was missing, a newspaper URL was manually added to the list whenever possible. In addition, the cities represented by the newspapers were mapped and regional gaps were filled by adding newspapers from the largest cities in those regions whenever possible. In total, the final version of the list used for this paper contains 1,349 newspaper websites representing 1,232 cities from across the contiguous United States. In addition, the longitude and latitude for each of these 1,232 cities were obtained from the U.S. Postal Service.

3. The Measurement of *Sneakers/Tennis Shoes/Running Shoes/Gym Shoes* Alternation

In order to describe how to use SRWSs to collect regional linguistic data, the analysis of the alternation between *sneakers, tennis shoes, running shoes* and *gym shoes* is presented here in detail. This lexical alternation variable was selected to exemplify the application of the method because it allows for the measurement of both multi-word lexical items and alternations consisting of more than two variants to be discussed. These four variants were selected for analysis because they are the four most frequent variants for this alternation variable in American English according to the Harvard Dialect Survey (see Section 5). The exclusion of less frequent variants, including *trainers, runners* and *jogging shoes,* is

discussed below.

At the core of the method being introduced here is the use of SRWSs to count the number of webpages upon which the variants of an alternation variable appear in hundreds of websites from across a region of interest. When querying Google, Bing, or other search engines, a search can be restricted to websites whose URLs contain a particular string by including that URL prefixed with the "*site:*" tag in the search box in addition to the search string. For example, searching for *site:nytimes.com "tennis shoes"*[1] counts the number of webpages on the *nytimes.com* website that contain the search string "tennis shoes," including websites indexed by Google that may otherwise have restricted access. The number of results returned by the SRWS, as listed by the search engine on the top of the results page, is then recorded. This process is then repeated for each of the variants and for each of the websites under analysis.

It should be emphasized at this point that SRWSs only allow for the total number of webpages on a website that contain the variants of an alternation variable to be counted, as opposed the actual number of times that those variants occur on that website. Part of the goal of this study is to test if proportions calculated based on these hit counts are accurate estimates of the real proportions of the variants on that website. It should also be emphasized that SRWSs do not necessarily count all of the pages on a website that contain a particular search string and might also count the same page multiple times, depending on how that website is indexed by the search engine. In addition, web searches in general are unstable, because search engines are constantly updated and newspaper webpages are regularly modified. Issues such as these may or may not invalidate the use of SRWSs for data collection in regional dialectology; the evaluation that follows tests whether or not this is the case.

Before using this method to observe a particular variable, it is necessary to ensure that the variable is suitable for analysis using SRWSs. First, at least one of the variants must occur relatively frequently in the variety of language under analysis. In this case, all four of these variants occur

multiple times on most of the 1,349 newspaper websites, with the first two variants being particularly common. Second, the variants must generally be used synonymously in the variety of language under analysis. This can be checked by looking over some of the webpages listed on the results pages generated by the SRWSs. In this case, for example, nine of the first ten webpages found by searching for *sneakers* on *nytimes.com* linked to newspaper articles where the word *sneakers* could have been replaced by the other variants without any major change in referential meaning, including an article on sneakers that tone your leg muscles while you walk and an article on a pair of sneakers designed to commemorate the World Basketball Festival. The tenth webpage, however, contained information on the 1992 movie "Sneakers." This hit is problematic because *sneakers* is being used as a proper noun and therefore cannot be replaced with the other variants in this context. Nevertheless, *sneakers*, as well as the other three variants analyzed here, appear to be used primarily to refer to athletic shoes in newspaper writing, and are therefore interchangeable in the majority of contexts. This alternation therefore appears to be suitable for analysis in newspaper writing using SRWSs.

Alternatively, if the most common variants of an alternation variable are highly polysemous or commonly used as a part of idioms or proper nouns, then that variable probably cannot be analyzed using SRWSs, at least following the basic method being introduced in this paper (although see Section 6). For example, these four most common variants have relatively stable meanings in American English, but this is not true of the less common variants *runners* and *trainers,* which are usually used to refer to people who run or who train. These variants were therefore excluded from this analysis, as well as other low frequency variants such as *jogging shoes* and *athletic shoes,* although these infrequent but relatively monosemous variants could have been included in the analysis. Excluding uncommon variants is not a problem. Ideally, the proportion of a variant is calculated by dividing the frequency of that variant in a sample of discourse by the frequency of all variants of that variable in that sample of discourse. However, when calculating the proportion of a variant, it is often acceptable to ignore the

frequencies of very uncommon variants. This is because when uncommon variants account for a very small percentage of the total occurrences of a variable, their inclusion or exclusion cannot substantially change the proportions of other variants. For example, *gym shoes* accounts for 8% of the total hits for this variable, but if *gym shoes* had been excluded from the analysis, the maps for the other three variables would have been practically identical. G*ym shoes* is simply too infrequent to have a significant effect on the proportions of the other variants. The same is true of *runners, trainers, jogging shoes,* and *athletic shoes* and other less frequent variants. Although ignoring uncommon variants does violate the principle of accountability, which requires that all variants be considered when analyzing an alternation variable (Labov, 1972; Kretzschmar, 2009), if the focus of the analysis is only on the most common variants of that alternation variable, as is the case here, then the principle of accountability is unnecessarily conservative.

Once the variable has been checked to make sure it is suitable for analysis, its variants can be counted across the set of regionally-defined websites using SRWSs. For example, in this case, the four variants were searched for in 1,349 newspaper websites, totaling 5,396 Google searches. Although the search engine can be queried manually, it is much easier to query the search engine automatically using computer programs designed for harvesting information online. In this case, a Perl LWP script was written to automatically download the html source code from the URL associated with the results page for that web search (e.g. *http://www.google.com/search?&q=%22 tennis+shoes%22&sitesearch= nytimes.com*) and the number of hits were then extracted from this html code. Although the searches were made for this analysis using Google, it is easier and quicker to use Bing, which has a much simpler html code and places fewer restrictions on the frequency of searches. Furthermore, based on informal comparisons, both Google or Bing appear to produce very similar results.

Overall, *sneakers* was found to be the most common variant accounting for 54% of the total hits, *tennis shoes* was found to be the second most common variant accounting for 25% of the total hits,

running shoes was found to be the third most common variant accounting for 13% of the total hits, and *gym shoes* was found to be the least common variant accounting for 8% of the total hits. Before computing the proportion of the variants for the individual cities, the hit counts were combined for every city that was represented by two or more newspapers. The proportions of each variant were then calculated across all of the cities where at least one of these four variants occurred. In total, 352 cities were excluded from the analysis, because none of the four variants occurred on the newspaper websites representing these cities, making it impossible to compute a proportion for these cities. For the other 880 cities, the proportion of each variant was calculated by dividing the number of hits for that variant by the total number of hits for all of the variants of that variable. For example, on the *Anniston Star* website, *tennis shoes* was found to refer to athletic shoes 64% of the time and *sneakers* was found to refer to athletic shoes 36% of the time, while *running shoes* and *gym shoes* were not found to be used at all. Although there is a lack of agreement in dialectology and sociolinguistics on how to treat non-binary alternation variables (see Chambers and Trudgill, 1998), especially when measured quantitatively and especially when the variants cannot be arranged in a natural ranking, measuring the proportion of each variant relative to all of the variants is a simple solution to this problem.

The proportions of each variant were then mapped across the 880 cities[2]. These maps are presented in Figure 1. The shading of a dot represents the proportion of the variant at that location: a darker dot indicates that the variant is relatively common at that location and a lighter dot indicates that the variant is relatively uncommon at that location. Figure 1 shows that *sneakers* is most common in the Northeast and *tennis shoes* is most common across the rest of the United States, especially in the Southeast and the North. The maps for the other variants are less clear but appear to show that *running shoes* is most common in the West, and that *gym shoes* is most common in the Midwest. There is, however, no need to rely on a subjective analysis to determine if these variables are regionally patterned; the statistical analysis presented in Section 4 will allow for these preliminary observations to

be verified by identifying the locations of significant high- and low-value clusters for each set of proportions.

Figure 1 Proportion of *Sneakers, Tennis Shoes, Running Shoes, Gym Shoes* (SRWS)

4. Spatial Autocorrelation Analysis

The lack of clear regional patterns in the maps presented in Figure 1 is not unusual. As will be demonstrated below, mapping linguistic alternation variables based on SRWSs generally produces noisy results[3]. In part this is because the proportions calculated through SRWSs are only estimates of the true proportions of the variants of alternation variables in newspaper writing. In part this is because temporal, functional and social sources of linguistic variation cannot be closely controlled when gathering data through SRWSs, as would be possible if a traditional approach to data collection had been adopted. This is because newspaper websites will vary in terms of the chronological depth of their archives, the range and proportion of registers that they publish online, and the demographic background of their authors, including the percentage of syndicated columnists and other non-local authors. All of these factors can affect the proportions of the the variants of an alternation variable, and may therefore obscure spatial patterns in data collected through SRWSs.

The maps for each variant were therefore subjected to a Getis-Ord *Gi* local spatial autocorrelation analysis (Ord and Getis, 1995) in order to identify underlying patterns of regional variation (see also Grieve, 2011, 2012; Grieve et al., 2011, 2013)[4]. A Getis-Ord *Gi* analysis is a geostatistical technique that identifies significant patterns of spatial clustering in the values of a variable that is measured over a series of locations. By comparing the value of a variable at each location to its values at nearby locations, a Getis-Ord *Gi* analysis identifies clusters of locations where the values of that variable are significantly higher or lower than would be expected if these values were distributed across the locations at random. In order to conduct a Getis-Ord *Gi* analysis, it is necessary

to define a *spatial weighting function*, which is a set of rules that assigns a weight to every pair of locations, so that comparisons between locations that are close together are given greater weight than comparisons between locations that are far apart (Odland, 1988). Various spatial weighting functions are possible, but in this case a reciprocal weighting function was used, which is a common spatial weighting function that weighs the comparison of the value of the variable at two locations based on the reciprocal of the distance between those two locations (Odland, 1988; Grieve et al, 2013)[5]. Given a specific spatial weighting function, the Getis-Ord *Gi* analysis generates a *z*-score for each location indicating the degree to which that location is part of a region of predominantly high values (a significant positive *z*-score), a region of predominantly low values (a significant negative *z*-score), or a region of transition or variability (an insignificant *z*-score approaching zero). These Getis-Ord *Gi* z-scores are then mapped across the locations in the dataset in order to identify the regions where the values of that variable tend to be particularly high or low.

The local spatial autocorrelation maps for *sneakers, tennis shoes, running shoes* and *gym shoes* are plotted in Figure 2. In these maps, clusters of darker dots represent regions where the variant under analysis is relatively common and clusters of lighter dots represent regions where the variant under analysis is relatively uncommon. All of these maps identify clear and significant patterns of regional variation and confirm the analysis of the raw maps in Figure 1 presented above. Figure 2 shows that *sneakers* is most common in the Northeast, *tennis shoes* is most common in the rest of the United States, especially in the Southeast and the North, *running shoes* is most common in the western Midwest and the West, outside of California, and *gym shoes* is most common in the Midwest and to a lesser extent in the Pacific Northwest, especially Oregon. Although the maps presented here are based on a reciprocal spatial weighting function, various other spatial weighting functions were tested, including nearest neighbor weighting functions and binary weighting functions. However, varying the spatial weighting function had very little effect on the results of the analysis, only causing minor shifts

in the values of the Getis-Ord z-scores on the edges of the major clusters, while preserving the basic patterns identified in the analysis presented here.[6] Crucially varying the spatial weighting function did not change the basic results of the evaluation of this method, which are presented below.

Figure 2 Local Autocorrelation Map for *Sneakers, Tennis Shoes, Running Shoes, Gym Shoes*

5. Evaluation

To evaluate this method for data collection, content word alternations with known distributions in American English were mapped, and these maps were compared to the results of previous American dialect surveys. There are, however, very few content word alternations that have been mapped in previous American dialect surveys that are suitable for analysis in newspaper writing. This is because there have been only three American dialect surveys that have mapped numerous content word alternations and because most of the alternations that were mapped in these surveys are very uncommon in newspaper writing, as well as in Modern English more generally.

The *Linguistic Atlas of the United States and Canada* was the first American dialect survey to map numerous lexical alternation variables in American English. Although the survey was never completed, the results of numerous smaller regional dialect surveys were published, including Hans Kurath's *A Word Geography of the Eastern United Sates* (1949) and E. Bagby Atwood's *The Regional Vocabulary of Texas* (1962), which focus on lexical variation, and the dialect atlases for New England (Kurath et al., 1939), the Upper Midwest (Allen, 1973), and the Gulf Coast (Pederson, 1984-1993), which map phonological and grammatical variation as well. The rest of the United States, however, was never mapped, and almost all of the alternations that were mapped are very rare in modern newspaper writing (e.g. words for *hay stacks*, *dragon flies*, and *clabbered milk*). Similarly, the *Dictionary of American Regional English* (DARE; Cassidy & Hall, 1985, 1991; Hall & Cassidy, 1996; Hall, 2002, 2012, 2013; Carver, 1987) also focuses on rare vocabulary items (e.g. words for *fishing worms*,

cigarette butts, and *silver dollars*). DARE also does not provide maps for most alternations, and the maps that they do provide have been adjusted so that the relative size of the states is proportional to the number of informants, making it difficult to compare DARE to the maps generated here. For these reasons, in addition to being somewhat dated at this point in time, neither of these datasets could be the basis of an evaluation, although DARE in particular was consulted whenever possible.

The lexical alternation variables used to evaluate the method were therefore drawn from the *Harvard Dialect Survey* (HDS; Vaux, 2003), which is the only dialect survey that has mapped everyday content word alternations in modern American English. The HDS began as a paper questionnaire distributed in Bert Vaux's "Dialects of English" class at Harvard in 1999, but it was then expanded and placed online by Vaux in 2002, where the survey was completed by more than 47,000 informants over the next year. The online questionnaire elicited 122 phonological, grammatical and lexical alternation variables, including 53 content word alternations of the type being analyzed here. Although the results of the study were never formally published (but see Vaux, 2003), the maps for all 122 alternations are available online, where for each variable a map is provided that plots the occurrences of all variants across the United States based on the reported hometown of each informant.[7]

Although the HDS is the only American dialect survey suitable for evaluating the method being introduced here, it is important to acknowledge that there are some limitations with this dataset. Perhaps most notably, because the HDS was conducted online, there was relatively little control over the regional and social backgrounds of informants, compared to traditional American dialect surveys. Nevertheless, because it was conducted online, the HDS was able to sample many more informants than traditional American dialect surveys, which to some extent offsets issues of control. The HDS was also presumably biased toward young, affluent, and urban informants, but while this perhaps limits the generalizability of the survey, the same is true of traditional dialect surveys that focused on non-mobile, old, and rural males. Finally, the approach to mapping was relatively simple, with infrequent variants

and fine detail potentially being obscured by the massive amounts of superimposed data points; however, the raw dataset from the HDS was made available by Bert Vaux, allowing for the data to be remapped for comparison.

Despite these limitations, all of the content word alternation variables analyzed here were therefore from the HDS, as it is only dataset that is suitable for this purpose. In particular, out of the 53 content word alternations elicited by the HDS, 10 alternations were selected to evaluate the method. The ten lexical alternation variables analyzed here are *bag/sack, carry out/take out, casket/coffin, drinking fountain/water fountain, frosting/icing, garbage can/trash can, cut the grass/mow the grass/mow the lawn, grandma/granny/nana*[8]*, garage sale/rummage sale/tag sale/yard sale,* and *gym shoes/running shoes/sneakers/tennis shoes.* These ten variables were selected because they are the only variables whose most frequent variants are both relatively common and monosemous in newspaper writing. Almost all of the other lexical alternation variables from the HDS had to be excluded because they are very rare in newspaper writing (e.g. words for the *night before halloween*, *daddy long legs*, and the *end of a loaf of bread*).

Each of these 26 variants were then mapped based on the HDS dataset so that these maps could be compared to the corresponding maps produced using SRWSs. In order to make these two sets of maps as comparable as possible, the HDS informants were pooled by city before being mapped. Specifically, for each alternation variable, informants were pooled by computing the proportion of informants from each city that preferred each variant of that variable. All cities represented by fewer than five informants were then deleted from the dataset, leaving 1,162 cities representing 29,240 informants. For each of the variants, a map was then produced showing the proportions of informants in each city that selected that variant. These raw maps were then subjected to a local spatial autocorrelation analysis, as described above, and the Getis-Ord *Gi z*-scores were mapped so that the results of the HDS could be directly compared to the results obtained here.

6. Results

The 26 variants of the 10 lexical alternation variables were measured across the 1,349 newspaper websites based on 35,074 SRWSs made on Google between October 14th and November 2nd, 2011 and between December 14th and December 29th, 2011. The proportion of each variant was then calculated for each city relative to all of the variants that were measured for that variable. These proportions were then subjected to a local spatial autocorrelation analysis and mapped in order to identify underlying pattens of regional variation in the values of each variant. Finally, these maps were visually compared to the corresponding maps from the HDS in order to evaluate the accuracy of this method for mapping regional linguistic variation.

For each of the 10 alternation variables, Table 1 lists the figure number, the variants of this variable that are being analyzed, the overall proportions of each variants in both the SRWS and HDS datasets, the total number of cities over which each variable was measured in each dataset, and any additional variants from the HDS that were excluded from the analysis due to polysemy or infrequency in newspaper writing. For some variables, the proportions of the variants are very similar across the two datasets, whereas for other variables the proportions of the variants are quite different. Aside from variables where one or more variants appear to be relatively polysemous (e.g. *carry out/take out, frosting/icing*), it is unclear why this would be the case. However, the maps for the variants may or may not align across the two datasets regardless of whether or not their overall proportions align in the two datasets. The total number of cities over which each variable was measured also varies within the two datasets. In the SRWS dataset, this is because sometimes none of the variants occur on the websites from a particular city, in which case a proportion cannot be computed for that city. In the HDS dataset, this is because sometimes all of the informants from a particular city leave a question unanswered or select variants that were excluded from this study.

Table 1 Lexical Alternation Variables

		Site-restricted Web Searches		Harvard Dialect Survey		
Variants	Fig.	Prop.	Cities	Prop.	Cities	Excluded Variants
bag	3	82%	1,217	87%	1,162	*poke* (polysemous)
sack		18%		13%		
carry out	4	48%	1,123	11%	1,160	
take out		52%		89%		
casket	5	44%	655	24%	1,162	
coffin		66%		76%		
drinking fountain	6	10%	337	39%	1,161	*bubbler* (rare), *water bubbler* (rare)
water fountain		90%		61%		
frosting	7	20%	934	65%	1,157	
icing		80%		35%		
garbage can	8	41%	703	42%	1,162	*rubbish bin* (rare), *waste basket* (rare)
trash can		59%		58%		
cut the grass	9	66%	273	23%	1,162	*cut the lawn* (rare)
mow the grass		12%		8%		
mow the lawn		22%		69%		
grandma	10	61%	1,205	86%	1,162	*gramma* (rare), *grammy* (polysemous), *mimi* (rare)
granny		26%		6%		
nana		12%		8%		
garage sale	11	57%	1,122	59%	1,162	*car boot sale* (rare), *carport sale* (rare), *jumble sale* (rare), *patio sale* (rare), *sidewalk sale* (rare), *stoop sale* (rare), *thrift sale* (rare)
rummage sale		6%		4%		
tag sale		3%		2%		
yard sale		34%		35%		
gym shoes	12	8%	880	6%	1,162	*athletic shoes* (rare), *jumpers* (polysemous), *runners* (polysemous), *sand shoes* (rare), *trainers* (polysemous)
running shoes		13%		2%		
sneakers		54%		37%		
tennis shoes		25%		55%		

The local spatial autocorrelation maps for the variants of the ten variables, based on both the SRWS and HDS datasets, are presented in Figures 3-12. It is important to note that when an alternation variable consists of more than two variants, each variant is mapped separately, as was exemplified above for *sneaker/tennis shoes/running shoes/gym shoes* alternation. In these maps, clusters of darker dots represent regions where that variant is relatively common and clusters of lighter dots represent regions where that variant is relatively uncommon. Alternatively, when an alternation variable consists of only two variants, it is only necessary to map one set of proportions, because the map for the second variant is always the inverse of the map for the second variant. In these maps, clusters of darker dots represent regions where the first variant is relatively common and clusters of lighter dots represent regions where the second variant is relatively common.[9]

Figure 3 *Bag/Sack* Alternation

Figure 4 *Carry Out/Take Out* Alternation

Figure 5 *Casket/Coffin* Alternation

Figure 6 *Drinking Fountain/Water Fountain* Alternation

Figure 7 *Frosting/Icing* Alternation

Figure 8 *Garbage Can/Trash Can* Alternation

Figure 9 *Cut the Grass/Mow the Grass/Mow the Lawn* Alternation

Figure 10 *Grandma/Granny/Nana* Alternation

Figure 11 *Garage Sale/Rummage Sale/Tag Sale/Yard Sale* Alternation

Figure 12 *Gym Shoes/Running Shoes/Sneakers/Tennis Shoes* Alternation

Bag/sack alternation follows a similar pattern in both maps (see Figure 3). According to the SRWS map, the use of *bag* is relatively common on the East Coast, especially in the Northeast and the Middle Atlantic States, and to a lesser extent in the Southwest, while the use of *sack* is relatively

common in the Midwest and the Central States. The West is identified as a region of variability. The HDS map identifies a similar pattern, except that the eastern Midwest, which is identified as a *sack* region in the SRWS map, and California, which is identified as a region of variability in the SRWS map, are identified as *bag* regions.

Carry out/take out alternation follows a nearly identical pattern in both maps (see Figure 4), although both variants are relatively polysemous in newspaper writing compared to the other variables being analyzed here. According to the SRWS map, the use of *carry out* is relatively common in the Central States and the Midwest, while the use of *take out* is relatively common in the Northeast, Florida, and the West. The South is identified as a region of variability. The HDS map identifies almost the exact same regional pattern, including small details like the identification of Minneapolis and Atlanta as cities of variability. The only differences are that Colorado and New Mexico are identified as a region of transition in the HDS map, but as a *carry out* region in the SRWS map, and Washington State is identified as a region of variability in the SRWS map, but as a *take out* region in the HDS map.

Casket/coffin alternation follows a nearly identical pattern in both maps (see Figure 5). According to the SRWS map, the use of *casket* is relatively common in the Central States and the Southeast, while the use of *coffin* is relatively common in the Northeast and to a lesser extent in the West. The HDS map identifies the same basic regional pattern, except that the western *coffin* region is stronger in California, the eastern *coffin* region is extended into Virginia and North Carolina, and the central *casket* region is extended into Ohio and Michigan.

Drinking fountain/water fountain alternation follows the same pattern in both maps (see Figure 6), although the SRWS map is based on far fewer locations, because these words do not occur at all on most of the newspaper websites. According to both maps, the use of *drinking fountain* is relatively common in the Midwest and the West, while the use of *water fountain* is relatively common in the East and the South Central States.

Frosting/icing alternation follows a similar pattern in both maps (see Figure 7), as well as in DARE (Hall, 2013), despite the fact that *icing* is relatively polysemous in newspaper writing. According to the SRWS map, the use of *frosting* is relatively common in New England and most of the Midwest and the Northwest, while the use of *icing* is relatively common in the Southeast, including southern Ohio and Indiana. The West is identified as a region of variability. The HDS map identifies the same basic regional pattern, except that the West is identified as a *frosting* region, and the Southeast *icing* region is extended outward, including cities such as Dallas, Kansas City, St. Louis and New York.

Garbage can/trash can alternation follows a nearly identical pattern in both maps (see Figure 8). According to the SRWS map, the use of *garbage can* is relatively common in the North, while the use of *trash can* is relatively common in the South, with the border between the two regions running through Pennsylvania, the Lower Midwest, and across the country into California. The HDS map identifies almost the exact same regional pattern, except that Philadelphia and eastern Massachusetts are identified as *trash can* regions in the HDS map, but as regions of transition or variability in the SRWS map.

Cut the grass/mow the grass/mow the lawn alternation follows similar patterns in both sets of maps (see Figure 9), although the SRWS map in this case is based on far fewer locations. According to the SRWS maps, the use of *cut the grass* is relatively common in the East, aside from New England, the use of *mow the lawn* is relatively common in the West and in New England, and the use of *mow the grass* is relatively common in the region in between—in the Lower Midwest, the Upper South, and Texas. The HDS maps identify the same basic patterns for the two most common variants, except that the *cut the grass* region is extended into the Midwest, and the *mow the lawn* region includes New York City. The HDS map for *mow the grass* also identifies a similar pattern, except that the *mow the grass* region is much larger, extending downward to include the entire the Deep South.

Granny/Grandma/Nana alternation follows similar patterns in both sets of maps (see Figure

10). According to the SRWS maps, the use of *grandma* is relatively common in the West and the Midwest, the use of *granny* is relatively common in the Southeast and the southern Midwest, and the use of *nana* is relatively common in New England, the Southwest, and to a lesser degree along the East Coast and across the South. The HDS maps, which are based on the combined results for maternal and paternal grandmothers, which were separate questions on the HDS, identify the same basic patterns for the two most common variants, except that the border between the *grandma* region and the *granny* region is slightly further south, following the Ohio river. The HDS map for *nana* also identifies a similar pattern, except that the variant is much less widely distributed, only being relatively common in the Northeast and in the Southern California.

Garage sale/rummage sale/tag sale/yard sale alternation follows similar patterns in both sets of maps (see Figure 11). According to the SRWS maps, the two most frequent variants are in complementary distribution, with the use of *garage sale* being relatively common in the West, aside from California, and with the use of *yard sale* being relatively common in the East, with the approximate border between the two regions running along the Ohio and Lower Mississippi rivers. In addition, the use of *rummage sale* is relatively common in the Midwest, the northern Mountain States, and the West Coast, and the use of *tag sale* is relatively common in New England, in addition to a few isolated clusters. The HDS map identifies the same basic pattern, except that California is identified as a *garage sale* region, the Middle Atlantic States are identified as a *tag sale* region, and the *rummage sale* region is limited to the Midwest and northern Mountain States.

Finally, *gym shoes/running shoes/sneakers/tennis shoes* alternation follows similar patterns in both sets of maps (see Figure 12). As discussed above, according to the SRWS maps, *sneakers* and *tennis shoes* are in complementary distribution, with the use of *sneakers* being relatively common in the Northeast, and to a lesser extent along the entire the East Coast, and with the use of *tennis* shoes being relatively common across the rest of the United States. In addition, the use of *running shoes* is

relatively common in Illinois, Missouri, Iowa and the West, except for California, and the use of *gym shoes* is relatively common in the eastern Midwest and to a lesser extent the Northwest. The HDS map identifies the same basic pattern, especially for *sneakers* and *tennis shoes*. *Running shoes* and *gym shoes* also show similar patterns in the HDS maps, except that California was identified as a *running shoes* region, the Midwest was identified as a stronger *gym shoes* region, and the Northwest was not identified as a *gym shoes* region.

6. Discussion

Overall, the maps generated through SRWSs match the maps based on the HDS quite well. There are certainly some differences between these two sets of twenty maps (especially between the maps for *mow the grass, bag/sack, nana, rummage sale, frosting/icing*), but every pair of maps exhibits the same basic pattern, and in many cases these maps are almost identical.

The most consistent difference between the two sets of maps is that the HDS maps generally identify stronger and more categorical patterns than the SRWS maps, with smaller regions of transition and fewer regions of variability. For example, California is identified as a stronger *bag* region and the Midwest is identified as a stronger *gym shoes* region in the HDS maps. In addition to differences in the relative strength of the maps, some pairs of maps are characterized by shifts in the locations of regions associated with a particular variant. For example, the eastern *bag* region is extended into the Midwest and the midland *mow the grass* region is extended into the Deep South in the HDS maps. Furthermore, in some pairs of maps the regions associated with a particular variant in one map are identified as regions of variability in the other map. For example, California is identified as a *garage sale* region and the West is identified as a *frosting* region in the HDS maps but as regions of variability in the SRWS maps. Alternatively, in a few cases, the SRWS maps identified more extensive regional patterns. For example, the South is identified as a *nana* region and the West is identified as a *rummage sale* region in

the SRWS maps but as regions of variability in the HDS maps.

There are various possible explanations for the differences between the two sets of maps, some of which point to weaknesses in the SRWS method as it was applied here. Most notably, some differences may be due to the difficulty of analyzing variables whose main variants are infrequent in newspaper writing. For example, the SRWS maps for *mow the grass, rummage sale,* and *nana* may not align as well with the HDS maps as most of the other variants analyzed here, because these are three of the least frequent variants under analysis. Furthermore, although the SRWS maps for these variants still do largely align with the corresponding HDS maps, the method as it was applied here cannot be used to measure variables whose main variants are very infrequent in modern newspaper writing, including most of the variables analyzed in previous American dialect surveys, which is why these variables were excluded from the evaluation. It should be noted, however, that this is not necessarily an inherent problem with the basic approach of using SRWSs for data collection in regional dialectology: the infrequency of these variables is a characteristic of the variety of language under analysis. It may be possible to use SRWSs to measure many of these alternations in websites representing other varieties of language, but further research is needed to test such an extension of the method.

Other differences between the two sets of maps may be due to the difficulty of using SRWSs to measure alternation variables whose main variants are polysemous in newspaper writing. For example, the SRWS map for *frosting/icing* alternation may not align as well with the HDS map as most of the other variants analyzed here, because *icing* is one of the most polysemous variants under analysis, often being used in articles that discuss weather related topics. Furthermore, although the SRWS map for this variant still does largely align with the corresponding HDS map, the method cannot be used to measure variables whose main variants are highly polysemous, especially when the most common meaning of that variant is not the same as the other variants of that variable. For example, the method was unable to *map soda/pop* alternation correctly, because the use of the word *pop* to refer to a soft drink is much

less common in newspaper writing than the use of this word to refer to *pop culture*. Such variables are not suitable for analysis using the method as it is presented here, although it may be possible to use SRWSs to analyze variables with highly polysemous variants by counting variants in specific contexts where they are generally interchangeable (*e.g. drink a soda*/*pop*) or by excluding the variants in specific contexts where they are not interchangeable (*e.g. pop music, soda cracker*). Further research is needed to test such an extension of the method.

Although some of the variation between the two sets of maps is likely due to issues with the SRWS method, some of this variation may reflect small but real differences between the type of data that the two approaches are being used to collect. For example, the reason that the HDS maps in general show stronger and more categorical patterns than the SRWS maps is probably because the HDS maps are based on a survey where each informant was asked to select the variant that they use, whereas the SRWS maps are based on the measurement of the frequency of these variants in natural language. Alternatively, the maps for *bag/sack* alternation may differ because the HDS asked informants for the word for a "paper container in which you might bring home items you bought at the store," whereas the SRWS measured a more general alternation by counting all occurrences of these words regardless of context. But perhaps most important, the maps for some variants may differ because these two methods were used to measure regional variation from two slightly different varieties of language—from two different decades, registers, and social groups. It is therefore possible that both sets of maps are correct, and that divergences between the two sets of maps reflect small differences in the variables and varieties of language under analysis.

Despite the differences between these two sets of maps, all 26 variants exhibit the same basic regional patterns. This paper has therefore shown that using SRWSs to measure lexical variation is a practical and powerful approach to data collection in regional dialectology that is capable of mapping lexical alternation variables with variants that are relatively frequent and monosemous in newspaper

writing with a relatively high degree of accuracy. Furthermore, the use of SRWSs allows for these lexical alternation variables to be mapped with far greater efficiency than is possible using traditional approaches to data collection. The speed at which data can be collected using this new method for data collection is the main reason why this method is an important addition to the dialectologist's toolbox. Standard approaches to data collection in regional dialectology are relatively slow and labour intensive, often requiring years to map a variable across a region as large as the United States. By using SRWSs, however, it is possible to map a variable across the United States in a matter of hours.

Although the use of SRWSs allows for dialect data to be collected much more quickly than is possible using traditional approaches to data collection, the method has a much more limited scope than these traditional approaches and must be applied with care. Most notably, the method as it was applied here is not suitable for mapping lexical alternations variables whose variants are infrequent or polysemous in newspaper writing, including most of the lexical alternation variables that have been analyzed in previous American dialect surveys, which in particular tend to be very rare in modern American English. The method also cannot be used to measure phonological alternations, although the method could be extended to analyze certain grammatical alternation variables. Nevertheless, there are undoubtedly hundreds and perhaps even thousands of lexical alternation variables that are relatively common in newspaper writing, as well as in American English more generally, that can now be mapped for the first time using the basic method introduced in this paper. The method can also be used in conjunction with traditional methods for data collection. For example, SRWSs could be used to test a large set of lexical alternation variables to determine which variables should be included on a traditional questionnaire. Although an actual corpus was not compiled for this study, this method could also be the basis for compiling a corpus, either by following the links to the webpages identified through the SRWSs, or by directly downloading the extracts from those webpages returned by the search engine. The method could also be extended to analyze other varieties of language, assuming that

a large enough set of regionally defined websites representing that variety of language can be identified.

Finally, it is important to acknowledge that the method introduced here appears to be one of the most successful applications of commercial search engines for the collection of linguistic data—a practice that has recently been criticized in the literature (Kilgarriff, 2006; Lüdeling, Evert & Baroni, 2006; Baroni and Kilgarriff, 2006; Fletcher, 2012). Among other issues, mining Google hit counts has been criticized on the grounds that register variation cannot be controlled, that webpages can be repeated and thus counted more than once, that the number of searches that can be made per day is limited, that webpages are not annotated for grammatical information, and that search engines count pages containing particular strings rather than the strings themselves. Some of these issues have been addressed here. The use of SRWS in particular has allowed for register variation to be largely controlled. Analyzing the proportions of synonymous forms rather than analyzing the raw hit counts directly also largely neutralized the problem of counting repeated web-pages: while repeated pages will inflate the raw frequency of search strings, in general repeated pages will not effect the frequency of search strings when measured relative to other synonymous search strings. Other issues raised in these critiques have not been dealt with directly, but given the success of the method, they do not appear to be as serious as has been previously assumed. For example, search engines do limit the number of searches per day, but it is still much quicker to search Google than to travel across a region interviewing individual informants. Similarly, although it is not possible to check the part-of-speech of strings being counted or to retrieve actual string frequencies rather than page counts, these sources of noise can be overcome through the application of advanced statistical methods, as applied here. This paper has therefore shown that it is both possible and productive to use commercial search engines to collect linguistic data, especially when search engines allow for linguistic data to be collected with far greater efficiency than is possible using traditional approaches. The power of web searches for data

collection in dialectology has been demonstrated in this paper, but there are certainly many other fields of linguistics where similar methods could be applied.

Notes

1. The search string must be enclosed by quotation marks to avoid searching for synonyms. Note also that punctuation marks and capitalization are ignored when included in the search.

2. All maps were made in R using functions from the *maps, maproj* and *maptools* and *sp* packages (Bivand et al., 2008).

3. Note that mapping linguistic variables does not generally result in clear spatial patterns even if questionnaires (e.g. Kurath, 1949), sociolinguistic interviews (e.g. Labov et al., 2006; see Grieve et al., 2013), or corpus-based methods (e.g. Grieve et al., 2011) are used for data collection.

4. The spatial autocorrelation analysis was conducted in R using functions from the *spdep* package (Bivand et al., 2008).

5. In addition, because of the large number of locations under analysis, comparison were limited to the closest 300 locations.

6. For example, numerous spatial weighting functions were tested when mapping *sneakers/tennis shoes/running shoe/gym shoes* alternation, and under all reasonable parameter settings the maps were almost identical, with the same basic regions being identified in all cases, and with most of the maps being almost indistinguishable.

7. See http://www4.uwm.edu/FLL/linguistics/dialect or http://www.tekstlab.uio.no/cambridge_survey for the maps for the HDS, although note that the base dataset used here is contains more informants and the data has been mapped differently.

8. The variant “grandmother” was excluded from the analysis because the HDS asked for the “nickname” used for one's female grandparents.

9. Only the local autocorrelation maps for these 10 variables are provided in the paper for comparison, but all of the raw maps and locally autocorrelated SRWS and HDS maps are presented in color in the supplemental materials.

References

Allen, Harold B. (1973). *The Linguistic Atlas of the Upper Midwest*. Minneapolis: University of Minnesota Press.

Atwood, E. Bagby. (1962). *The Regional Vocabulary of Texas*. Austin: University of Texas Press.

Baroni, Marco and Kilgarriff, Adam. (2006). Large linguistically-processed web corpora for multiple languages. *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics*: 87–90.

Bivand, Roger S., Pebesma, Edzer J. and Gomez-Rubio, Virgilio. (2008). Applied Spatial Data Analysis with R. New York: Springer.

Boberg, Charles. (2005). The North American Regional Vocabulary Survey: New variables and methods in the study of North American English. *American Speech* 80: 22-60.

Carver, Craig. (1987). *American Regional Dialects*. Ann Arbor: University of Michigan Press.

Cassidy, Frederic G. and Hall, Joan Houston. (1985). *Dictionary of American Regional English, Volume I: Introduction and A-C*. Harvard University Press.

Cassidy, Frederic G. and Hall, Joan Houston. (1991). *Dictionary of American Regional English: Volume II: D-H.* Harvard University Press.

Chambers, J.K. (1994). An introduction to dialect topography. *English World-Wide* 15: 35-53.

Chambers, J.K. and Trudgill, Peter. (1998). *Dialectology*. 2nd Edition. Cambridge University Press.

Davis, Alva L. (1948). *A Word Atlas of the Great Lakes Region.* Ph.D. Dissertation. University of Michigan.

Elspaß, Stephan, Langer, Nils, Scharloth, Joachim, and Vandenbussche, Wim. 2007. *Germanic Language Histories 'from Below' (1700-2000)*. Berlin: Walter de Gruyter.

Fletcher, William H. (2012). Corpus analysis of the world wide web. In Carol A. Chapelle (Ed.) *Encyclopedia of Applied Linguistics.* Hoboken, New Jersey: Wiley-Blackwell.

Grieve, Jack. (2011). A regional analysis of contraction rate in written Standard American English. *International Journal of Corpus Linguistics* 16: 514-546.

Grieve, Jack. (2012). A statistical analysis of regional variation in adverb position in a corpus of written Standard American English. *Corpus Linguistics and Linguistic Theory* 8: 39-72.

Grieve, Jack, Speelman, Dirk, and Geeraerts, Dirk. (2011). A statistical method for the identification and aggregation of regional linguistic variation. *Language Variation and Change* 23: 193-221

Grieve, Jack, Speelman, Dirk, and Geeraerts, Dirk. (2013). A multivariate spatial analysis of vowel formants in American English. Forthcoming in the *Journal of Linguistic Geography* 1.

Hall, Joan Houston. (2002). *Dictionary of American Regional English, Volume IV: P-Sk*. Harvard University Press.

Hall, Joan Houston. (2012). *Dictionary of American Regional English, Volume V: Sl-Z.* Harvard University Press.

Hall, Joan Houston. (2013). *Dictionary of American Regional English, Volume VI: Contrastive Maps, Index to Entry Labels, Questionnaire, and Fieldwork Data.* Harvard University Press.

Hall, Joan Houston and Cassidy, Frederic G. (1996). *Dictionary of American Regional English, Volume III: I-O.* Harvard University Press.

Kilgarriff, Adam. (2006). Googlelology is bad science. *Computational Linguistics* 33: 147–151.

Kretzschmar, William. (2009). *The Linguistics of Speech*. Cambridge University Press.

Kurath, Hans. (1949). *A Word Geography of the Eastern United States*. University of Michigan Press.

Kurath, Hans, Hansen, Marcus L., Bloch, Bernard, and Bloch, Julia. (1939). *Handbook of the*

Linguistic Geography of New England. Providence: Brown University Press.

Labov, William, (1972). *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Labov, William, Ash, Sharon, and Boberg, Charles. (2006). *Atlas of North American English: Phonetics, Phonology, and Sound Change*. New York: Mouton de Gruyter.

Lüdeling, Anke, Evert, Stefan and Baroni, Marco, (2006). *Using web data for linguistic purposes*. In Hundt, Marianne, Nesselhauf, Nadja, and Biewer, Carolin (Eds.) *Corpus Linguistics and the Web*. Amsterdam: Rodopi.

Odland, John D. (1988). *Spatial Autocorrelation*. Thousand Oaks, CA: Sage Publications.

Ord, J. K. and Getis, Arthur. (1995). Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis* 27: 286-306.

Pederson Lee, McDaniel, Susan L., and Adams, Carol M. (1986-93). *Linguistic Atlas of the Gulf States* (7 Volumes). Athens, Georgia: University of Georgia Press.

Vaux, Bert. 2003. American dialects. In Steven Goldberg (Ed.) *Let's Go USA 2004*. Upper Saddle River, NJ: Prentice Hall.

Warkentyne, Henry J. (1973). Contemporary Canadian English: A report of the Survey of Canadian English. *American Speech* 46: 193-199.