A statistical analysis of regional variation in adverb position in a corpus of written Standard American English

JACK GRIEVE

Abstract

This paper investigates whether the position of adverb phrases in sentences is regionally patterned in written Standard American English, based on an analysis of a 25 million word corpus of letters to the editor representing the language of 200 cities from across the United States. Seven measures of adverb position were tested for regional patterns using the global spatial autocorrelation statistic Moran's I and the local spatial autocorrelation statistic Getis-Ord Gi. Three of these seven measures were indentified as exhibiting significant levels of spatial autocorrelation, contrasting the language of the Northeast with language of the Southeast and the South Central states. These results demonstrate that continuous regional grammatical variation exists in American English and that regional linguistic variation exists in written Standard English.*

Keywords: *adverb position, corpus linguistics, dialectology, dialectometry, spatial autocorrelation, written American English*

1. Introduction

The primary method of data collection in regional dialectology is the linguistic interview.[1] The linguistic interview has been adopted by dialectologists because it is an effective method for observing lexical and phonological variation. To observe lexical variation, which often involves words that are rare in natural language, it is easiest to directly elicit vocabulary items from an informant through a linguistic interview. To observe phonological variation, it is necessary to record or transcribe an informant's utterances. The linguistic interview is therefore well suited for observing variation in vocabulary and accent. This traditional approach to data collection, however, does not always allow for other forms of regional linguistic variation to be observed as efficiently. Most notably, the linguistic interview is often unsuitable for observing

DOI 10.1515/cllt-2012-0003

grammatical variation because grammatical constructions can be difficult to elicit from an informant and are unlikely to be uttered spontaneously during an interview. It is also difficult to observe continuous linguistic variation through the linguistic interview because measuring a continuous variable requires that many tokens of the variable be observed in discourse so that the relative frequency of its variants can be estimated accurately. In addition, the traditional approach to data collection does not allow for regional linguistic variation to be observed across the many registers of a language, including written and standard varieties, because collecting data through the linguistic interview only allows for language to be observed in that one very specific context. In order to analyze continuous grammatical variation in a range of registers it is necessary to directly analyze large amounts of natural language discourse. This can be accomplished by adopting a corpus-based approach to data collection.

In addition to allowing for new types of research questions to be investigated, the corpus-based approach to regional dialectology also allows for the language of hundreds of informants to be observed at each location. In traditional dialectology, usually the language of only two or three informants is observed at each location because interviewing informants is such a laborious task. In order to ensure that regional linguistic variation is found in such small sample, traditional dialect surveys have focused on the language of long-term residents – often elderly members of families that have lived in a region for many generations. While this approach has been used to successfully identify regional linguistic patterns, it is unclear if these patterns exist in the language of the general population or only in the language of that small minority of speakers. This problem can be overcome by sampling the language of hundreds of informants at each location, which is possible using a corpus-based approach to data collection. A large sample allows for the language of informants from across the population to be observed resulting in a more complete picture of regional linguistic variation. It is important to note that this larger sample should include the language of both short- and long-term residents, despite the fact that traditional dialect surveys only consider the language of long-term residents. Including short-term residents does complicate the identification of regional linguistic variation, as regional patterns will not be as clear in the dataset, but a more inclusive sample increases the likelihood that any regional patterns that are found in the dataset are characteristic of the language produced currently by the entire population under analysis, not just some small historic subsection of the population. Only by analyzing the language of the entire population can current and pervasive regional linguistic patterns be identified.

Despite the advantages of the corpus-based approach, the study of regional dialect variation has rarely been based on natural language discourse.[2] This is especially true in the case of American English, where no major study of regional linguistic variation has ever been carried out using true corpus data;

most corpus-based dialect surveys have focused on British English. One of the earliest and most important dialect corpora is the Helsinki Corpus of British English Dialects, which is based on 210 hours of recorded spontaneous speech collected in the 1970s and 1980s in six English counties (Inhalainen et al. 1987; Inhalainen 1990). Numerous dialect studies, which have primarily investigated continuous grammatical variation, have been based on the Helsinki Corpus (e.g., Ojanen 1985; Peitsara 1988), most notably by Ossi Inhalainen (e.g., Inhalainen 1976, 1980, 1985, 1991a, 1991b). More recently, dialect studies have been based on the Freiburg English Dialect Corpus, which contains 300 hours of recorded oral histories collected from 1968 to 2000 in England, Wales, Scotland, the Hebrides, and the Isle of Man (e.g. Kortmann et al. 2005; Hernandez 2006; Szmrecsanyi 2010). The Freiburg Corpus has also been used in functional-typological dialect studies (e.g., Kortmann 2004; Wagner 2004; Herrmann 2005), which are concerned with showing how language-internal dialect variation follows the same basic typological patterns that are found cross-linguistically. The British National Corpus has been used in dialect studies as well (e.g., Kortmann et al. 2005), and corpora have also been used in dialect studies focusing on historical non-standard varieties of English (Schneider and Montgomery 2001; McCafferty 2003; Van Herk and Walker 2005; Ingham 2006).

The goal of this study is to conduct a corpus-based analysis of continuous grammatical variation in written Modern Standard American English. The focus of this study is continuous grammatical variation in written Standard English because this type of linguistic variation and this type register of have received very little attention from dialectologists in the past. By searching for continuous grammatical variation in written Standard English it is thus possible to test the limits of regional linguistic variation – to investigate just how prevalent regional linguistic variation is in natural language. In particular, this paper describes an analysis of regional variation in the sentential position of adverbs in a 25 million word corpus of letters to the editor representing the language of 200 cities from across the United States. This paper is organized as follows. First, the design, compilation and dimensions of the corpus are presented. Second, the seven measures of adverb position are introduced and the algorithms used to compute these measures are described. Third, the two spatial autocorrelation statistics used to identify regional patterns in this dataset are presented; these statistics have not been used in past dialect studies but the voluminous and continuous nature of the data produced by a corpus analysis requires that advanced statistical techniques be adopted. Finally, the results of the analysis are reported and discussed. It is concluded that adverb position is regionally correlated in written Standard American English and that a corpus-based approach is an effective method for analyzing regional linguistic variation.

2. Corpus design

This section describes the design, compilation, and dimensions of the 25 million word corpus of letters to the editor that was the basis of this study of regional grammatical variation in written Standard American English. This section begins with a defense of the choice of the letter to the editor register and the selection of the 200 cities represented in the corpus. The process through which the letters to the editor were downloaded, cleaned and organized is then explained. Finally, the dimensions of the corpus are presented.

2.1 *Register selection*

The goal of this study is to determine if continuous grammatical variation exists in written Standard American English. The decision was made to analyze written Standard English because this is a variety of language that has not been the subject of previous analyses of regional linguistic variation. The letter to the editor register was selected in particular because it is a variety of written Modern Standard English that is very well suited to the analysis of regional linguistic variation. Most important, letters to the editor are annotated for their author's current place of residence, which allows letters to be sorted easily by geographical location. In addition, letters to the editor are published frequently and distributed freely online in machine-readable form, which allows for data to be collected quickly and cheaply. The frequency of publication also allows for letters to be gathered from a relatively short span of time, which minimizes the likelihood that temporal linguistic variation will confound a regional linguistic analysis of the corpus. Finally, letters to the editor are written by many non-professional authors from across the United States, which allows for a corpus containing the writings of a large and diverse sample of American authors to be compiled.

Despite the clear advantages of analyzing letters to the editor, there is a potential problem with this choice of registers: letters to the editor are presumably subject to editing by an editorial page editor. In order to address this issue a brief and informal questionnaire was sent by email to editorial page editors from many of the newspapers sampled in this study. The questionnaire asked whether or not letters to the editor were edited. Editors replied that they do edit letters to the editor, but mainly for clarity, spelling, fact, and length, none of which tend to have any direct effect on the grammar of a letter to the editor. Most editors also said that they do occasionally edit letters for grammar, although minimally: generally nothing is changed that is written in grammatically correct English; only obvious mistakes and ungrammaticalities are corrected, such as agreement errors and run-on sentences. The only exception is certain function words that can be optionally omitted from a text with no loss of information, such as complementizer *that* (e.g. *he thinks that he is right*). It appears that these types of words are sometimes deleted by editors, especially in longer letters, so as to reduce word counts. Otherwise it appears that editing does not generally affect the grammar of letters to the editor, including the position of adverbs, and it will therefore be assumed that the newspaper editing did not confound the results of this study.

2.2 *City selection*

Cities were selected for inclusion in the corpus based primarily on the availability of a free online archive containing a large number of letters to the editor that were recently published by a major newspaper in that city. No special interest or alternative newspapers were sampled. If a suitable archive could not be located, the city was not included in the corpus. Usually, if the archive did not contain at least 50,000 words of recently published letters to the editor, the city was not included in the corpus. In a few cases, however, smaller archives were sampled for large or geographically isolated cities. In cases where more than one newspaper with suitable archives were available for a city, the newspaper which had the larger archive, the more well-organized archive, and the higher circulation rate was sampled. In a few cases, multiple newspapers from a single city were sampled in order to increase the size of a city sub-corpus. When suitable newspaper archives were available, cities were selected based on numerous other criteria: populous cities, capital cities, isolated cities and historically important cities were primarily targeted for representation in the corpus. Overall, the basic approach to city selection was to locate newspaper archives for the largest and most important cities in every state in the contiguous United States and to then select smaller cities with suitable newspaper archives in order to fill in any regional gaps.

The geographical distribution of the 200 cities included in the corpus is presented in Figure 1. The cities in the corpus are relatively evenly distributed across the United States: the Northeast and Midwest are very well represented and the Southeast, Texas and the West Coast are well represented. While cities from all states are included, there are gaps in the Mountain States and the Northern Plains. These gaps are due primarily to sparse settlement in these regions. The ramification of having these regional gaps is that any dialect patterns that encompass these areas must be interpreted with care, especially in the sparsely populated region encompassing eastern Montana, northeastern Wyoming, western South Dakota and Western Nebraska, where no cities are represented. The corpus also includes most major cities in the United States. According to the 2000 census, the top 30 metropolitan areas in the United States are included in the corpus and the top 50 metropolitan areas in the United States are included in the corpus, except Providence, Rhode Island, Jacksonville, Florida, and Birmingham, Alabama. Other major cities missing from the corpus include New Haven, Connecticut, Worcester, Massachusetts, Baton Rouge, Louisiana, Springfield, Massachusetts, Harrisburg, Pennsylvania, Jackson, Mississippi, and Chattanooga, Tennessee. All of these cities were excluded from the corpus because of the unavailability of free and sizeable newspaper archives.

2.3 *Data collection*

For each city newspaper, letters to the editor were primarily obtained using the online service *Newsbank,* which provides complete archives for many American newspapers, or, when the newspaper for a particular city was not available on *Newsbank,* the online archives provided by the newspaper. Once an online archive was located and a searching strategy was devised to find letters to the editor in that archive, letters to the editor were then copied and pasted manually into a text editor. Whenever possible, letters from the years 2005–2008 were targeted for download. However, when necessary, letters from 2000–2004 were also sampled in order to increase the size of the sub-corpora. Once approximately 50,000–200,000 words of texts were downloaded for a city (depending on the size of the archive and the speed at which letters could be downloaded), collection for that city was stopped. In some cases where letters could be downloaded very quickly or where many letters appeared to be written by non-local authors, as is the case for the national newspapers published in New York City and Washington, D.C., more data was downloaded. Depending on the average number of letters to the editor that a newspaper archives per web page (i.e. some newspapers archive an entire days worth of letters on a single web page, whereas others archive each letter on a separate web page), this process took between twenty minutes and three hours for each city. Using this approach, approximately 35 million words were collected from newspapers from across the United States.

2.4 *Corpus cleaning*

Each text file, which contained the letters downloaded from a single newspaper archive, was subjected to four rounds of cleaning.[3] First, the text was split into individual letters and the author's name and place of residence and the date of publication for each letter was recorded. Second, boilerplate text was removed. Third, punctuation and spacing was standardized. Fourth, repeated letters were deleted. In addition, after each text file was cleaned, it was inspected by hand.

Each text file was first split into individual letters. In order to facilitate this process, some of the header newspaper information that preceded the letter on the archive web page was copied when the letters were downloaded. These lines also included the date of publication for all of the letters presented on that archive web page; this information was recorded. In those cases where the newspaper archive provided each letter on a separate web page, individual letters were identified by splitting on this header information. In those cases where the newspaper archive provided multiple letters on the same webpage, the text file was first split based on this header newspaper information and then split again based on the byline markings, which at this point was the most consistent letter delimiter. Unfortunately, different newspapers identify bylines differently and often inconsistently. Usually a byline was marked by a line initial dash (or some other symbol) or by the lack of line final punctuation, but sometimes more subtle patterns had to be identified, often involving line internal punctuation and capitalization. Once the byline was identified, the author's name and the author's place of residence were recorded, the text was split on the byline, and the byline was deleted. A header was then added to the start of each letter containing the author's name, the author's place of residence, and the date of publication, as well as the newspaper archive from which it originated. In addition, header information was standardized: names and cities were converted to capital letters, spacing and punctuation for initials in names were standardized, and the format of dates was standardized. If any of this information was missing, the letter was deleted.

Next, boilerplate text was deleted. Common strings of boilerplate text that were deleted included editor notes, instructions for letter writers, captions, abstracts, and advertisements. Certain conventionalized parts of the letter to the editor register – specifically the salutation and the closing – were also deleted when present because only some newspapers include these lines. All of these strings were removed from the corpus by identifying lines beginning with certain word and punctuation patterns. In addition, titles were removed from the letters to the editor because they are not always present and are presumably written by the newspaper and not the author. Titles were identified by looking for unpunctuated lines or short lines ending with a question mark or an exclamation mark at the very beginning of a letter.

Numerous features primarily related to spacing and punctuation were also standardized in order to facilitate data analysis. Overall, spacing was minimized. Extra whitespace between words, sentences and paragraphs was deleted, as were all tabs. Various punctuation marks were also standardized, including apostrophes, quotation marks, dashes, and ellipses. Bullets were standardized and sentence final punctuation was added to the end of bulleted lines. In addition, sentence final punctuation marks were moved outside of apostrophes, quotation marks, parentheses and brackets, and punctuation marks were deleted from the end of abbreviations in order to facilitate the identification of sentence boundaries. HTML punctuation codes (e.g. *&apos*, *&ndash, &ldquo*) were also replaced with their orthographic equivalent.

Finally, repeated letters were deleted. These letters resulted from either downloading error or from multiple postings of a letter in a newspapers archive. These letters were deleted automatically by comparing every pair of letters in a sub-corpus in order to find duplicate letters. Ultimately, the entire cleaning process resulted in the loss of approximately 8 million words of data, dropping the size of the entire corpus from 35 million words to 27 million words.

2.5 *Corpus organization*

Once every text file was cleaned, the individual letters were sorted into sub-corpora based primarily on the *core based statistical area* (CBSA) where their author currently resides. A CBSA is a term used by the United States Census Bureau to denote a region consisting of a county containing a core urban area with a population of at least 10,000 people and any adjacent counties with a high degree of socioeconomic integration with the core urban area, as indicated by the number of commuters. There are two types of CBSAs: *metropolitan areas*, which contain a core urban area of at least 50,000 people, and *micropolitan areas*, which contain a core urban area of between 10,000 and 50,000 people. Basically, a CBSA corresponds to a city and its suburbs. Letters were sorted primarily by CBSA, rather than by municipality, in order to increase the size of the corpus: had letters been sorted by municipality, there would have been many more letters that would have had to have been deleted because they originate from municipalities that had contributed too few letters to form a separate sub-corpus. However, because letters were sorted by CBSA, letters from many underrepresented cities could be included in the corpus, as they were part of larger CBSAs. Letters were automatically sorted by CBSA by cross-referencing a list of all the cities and counties in the United States (compiled by the United States Postal Service) with a list of all CBSAs in the United States and their associated counties (compiled by the United States Census Bureau). A sub-corpus was then formed for each CBSA for which at least 25,000 words had been collected.

Not every sub-corpus in the corpus, however, represents a CBSA: there are two exceptions. First, when a sufficient number of letters were available, sub-corpora were created containing all the letters from the same *metropolitan division* – a term used by the United States Census Bureau to denote one or more contiguous counties that constitute a distinct employment region within a metropolitan CBSA with a population core of at least 2.5 million people. For example, despite the fact that both San Francisco and Oakland are part of the same CBSA, distinct sub-corpora were compiled for each of these metropolitan divisions, because a sufficient number of letters originated from each. Creating corpora that represent metropolitan divisions increased the number of cities represented in the corpus and thereby increased the resolution of the

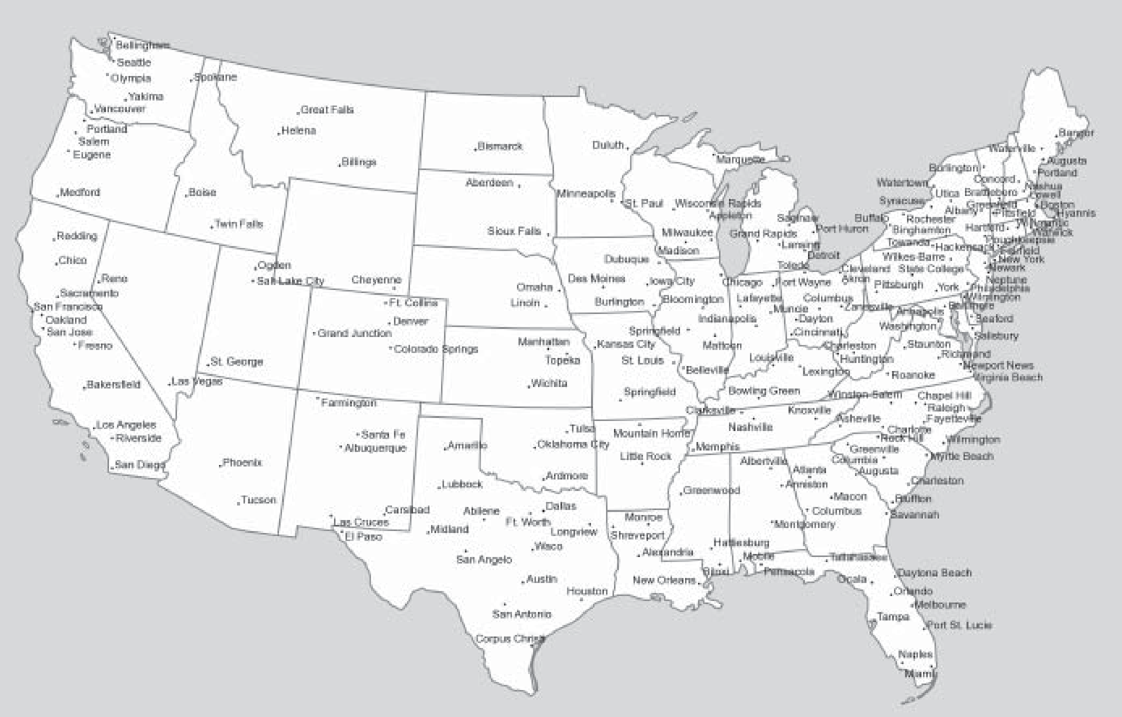

Figure 1. *Geographical Distribution of City Sub-Corpora*

study. Second, one sub-corpus was compiled containing letters written by the residents of Brattleboro, Vermont, even though it is not a part of any CBSA (due to its isolation and small population), because over 25,000 words of letters to the editor were downloaded from that town's newspaper – one of the few newspapers in Vermont that was accessible.

After the letters were sorted into the 200 city sub-corpora, each sub-corpus was analyzed by hand once again in order to ensure that there were no errors or misclassified letters included in the corpus. Ultimately, this procedure produced sub-corpora for each of the 200 cities identified in Figure 1. In total, this procedure resulted in a loss of approximately 2 million additional words from letters whose author's place of residence was not in any CBSA, metropolitan division, or city whose residents had contributed a sufficient number of letters to form a separate sub-corpus.

2.6 *Corpus dimensions*

The entire corpus contains 25,794,656 words. The mean sub-corpus size in words is 128,973 words. The size of the city sub-corpora ranges from 26,885 words (Omaha) to 317,592 words (Nashville). The entire corpus contains 154,269 letters to the editor. The mean sub-corpus size in letters is 771 letters. The size of city sub-corpora ranges from 119 letters (Springfield, Missouri) to 3,154 letters (Los Angeles). The entire corpus contains letters written by

126,422 different authors. The mean number of authors per sub-corpus size is 632 authors. The number of authors per city sub-corpus ranges from 105 authors (Springfield, Missouri) to 1621 authors (Dallas).

3. Corpus analysis

Seven measures of adverb position were tested for regional variation across the 200 city sub-corpora. Adverb position was chosen because the goal of this study is to test if continuous grammatical variation is regionally patterned in written Standard English and adverb position is a form of continuous grammatical variation that is relatively frequent and variable in written Standard American English. Continuous grammatical variables such as adverb position are not usually analyzed in American dialectology, which has focused instead on categorical forms of lexical and phonological variation. In a traditional categorical analysis of regional linguistic variation, each location is associated with only one of the variants of a linguistic variable (e.g. a location uses the term *spigot* or *faucet,* but not both). However, in a continuous analysis of regional linguistic variation, in a manner similar to sociolinguistic studies (Labov 1966a, 1966b, 1972; Wolfram 1969, 1991), each location is associated with a continuous value computed by calculating the proportion of one variant (V_a) relative to the total tokens of the variable – i.e., the total number of both the first (V_a) and second (V_b) variant of the variable (Equation 1).

(1)
$$V = \frac{V_a}{V_a + V_b}$$

This equation is the basis for all of the measures of adverb position analyzed in this study, which are of two types. First, the proportions of sentences with split modals, auxiliaries and infinitives were measured relative to the frequency of sentences with an adverb and a non-split modal, auxiliary or infinitive. Second, the proportion of various adverbs occurring sentence initially was measured relative to the frequency of these adverbs occurring sentence internally and sentence finally. These seven measures have not been the subject of previous studies of regional linguistic variation in American English.

3.1 *Split adverbs*

In this study, three variables based on adverb splitting were analyzed: *infinitive splitting, non-modal auxiliary splitting,* and *modal auxiliary splitting*. These three variables and the algorithms used to measure their values across the 200 city sub-corpora are described below.

In Standard English, an adverb can be placed between the infinitive marker *to* and an infinitive verb or it can occur elsewhere in a sentence without changing the meaning of that sentence. For example, the adverb *just* can be placed before or after the infinitive marker *to*, as illustrated in Sentence (1) and Sentence (2).

(1) Again, I encourage you ***to just go*** by and see what a wonderful place this is (*Baxter Bulletin*, Mountain Home, December 9, 2006).

(2) True, plastic bags should not be thrown into the regular garbage ***just to go*** to a landfill (*Arizona Republic*, Phoenix, June 6, 2007).

In order to calculate the value of *infinitive splitting* in each city sub-corpus, all strings consisting of *to* followed by a common adverb[4] or a word larger than five characters ending in -*ly* followed by a common verb in the infinitive form[5] were counted as instances of infinitive splitting, and all strings consisting of *to* followed by a common verb in the infinitive form were counted as instances of non-infinitive splitting if that string occurred in a sentence that also contained a common adverb or a word larger than five characters ending in -*ly*.

Similarly, in Standard English, an adverb can be placed between a non-modal auxiliary verb and a main verb or it can occur elsewhere in a sentence without changing the meaning of that sentence. For example, the adverb *often* can be placed between an auxiliary verb and a main verb or elsewhere in a sentence, as illustrated in Sentence (3) and Sentence (4).

(3) But I just have to correct two misperceptions about evolution that ***are often repeated*** (*Duluth News-Tribune*, December 15, 2005).

(4) Testing is a good idea, if 16- to 25-year-olds ***are tested often*** (*Milwaukee Journal Sentinel*, May 1, 2005).

In order to calculate the value of *non-modal auxiliary splitting* in each city sub-corpus, all strings consisting of a form of the auxiliary verbs *be* or *have* followed by a common adverb or a word larger than five characters ending in -*ly* followed by a common irregular verb[6] or a word larger than five characters ending in -*ed*/-*en* were counted as instances of auxiliary splitting, and all strings consisting of the auxiliary verbs *be* or *have* followed by a common irregular verb or a word larger than five characters ending in -*ed*/-*en* were counted as instances of non-auxiliary splitting if that string occurred in a sentence that also contained a common adverb or a word larger than five characters ending in -*ly*.

Finally, in Standard English, an adverb can be placed between a modal auxiliary verb and a main verb or it can occur elsewhere in a sentence without changing the meaning of that sentence. For example, the adverb *often* can be placed between a modal and a main verb or elsewhere in a sentence, as illustrated in Sentence (5) and Sentence (6).

(5) They ***will often have*** only a split second to make their life-or-death decisions (*Billings Gazette*, August 28, 2005).

(6) ***Often*** it ***will be*** just as easy for customers to go to a different area or out of state to make purchases (*Utica Observer-Dispatch,* March 17, 2005).

In order to calculate the value of *modal splitting* in each city sub-corpus, all strings consisting of a modal followed by a common adverb or a word larger than five characters ending in *-ly* were counted as instances of modal splitting, and all strings consisting of a modal not followed by a common adverb or a word not ending in *-ly* were counted as instances of non-modal splitting if that string occurred in a sentence that also contained a common adverb or a word larger than five characters ending in *-ly*.

3.2 *Sentence-initial adverbs*

In this study, four variables based on adverbs position (i.e. sentence initial vs. sentence internal/final) were also analyzed: *temporal adverb position, however position, also position,* and *instead position*. These specific adverbs were selected for analysis because they are the most common adverbs in the corpus. These four variables and the algorithms used to measure their values across the 200 city sub-corpora are described below.

In Standard English, a temporal adverb can be placed sentence initially or sentence internally without changing the meaning of the sentence. For example, *currently* can be placed sentence initially or sentence internally, as illustrated in Sentence (7) and Sentence (8).

(7) ***Currently,*** adolescents are transferred to facilities in Raleigh, Jacksonville, etc. (*Wilmington Star News,* November 30, 2005).

(8) There are only about a dozen ***currently*** (*Chico Enterprise Record,* December 6, 2005).

In order to calculate the value of sentence initial temporal adverbs, a list of 19 temporal adverbs[7] was used to count all tokens of sentence initial temporal adverbs in each city sub-corpus, which begin with a capital letter, and all tokens of sentence internal temporal adverbs, which begin with a lower case letter.

Similarly, in Standard English, the linking adverb *however* can be placed sentence initially or sentence internally without changing the meaning of the sentence. For example, *however* can be placed at the beginning of a sentence or sentence internally, as illustrated in Sentence (9) and Sentence (10).

(9) ***However***, it lacked very important details explaining why the no building zones around streams and wetlands are important (*The Olympian,* Olympia, August 24, 2005).

(10) What baffles me, ***however,*** is that we have some of the most righteous needs right here in our own community that are not being addressed (*Spokane Spokesman Review,* September 30, 2005).

Similarly, the adverb *also* can be placed sentence initially or sentence internally without changing the meaning of the sentence. For example, *also* can be placed at the beginning or the end of a sentence, as illustrated in Sentence (11) and Sentence (12).

(11) ***Also***, there is a reuse center at the Bloomington interchange for your plant materials (*St. George Spectrum,* May 21, 2004).

(12) If a person is in the United States illegally, then any minor children – regardless of where they were born – should be considered illegal immigrants, ***also*** (*Houston Chronicle*, October 1, 2006).

Finally, the adverb *instead* can also be placed sentence initially or sentence internally without changing the meaning of the sentence. For example, *instead* can be placed at the beginning or the end of sentence, as illustrated in Sentence (13) and Sentence (14).

(13) ***Instead***, we should focus on using the water we have more efficiently (*Salt Lake City Tribune*, August 10, 2008).

(14) Perhaps they should come down from on high and offer to help out ***instead*** (*Roanoke Times*, September 14, 2006).

In order to calculate the value of all three of these measures, these three words were counted in each city sub-corpus when they occur sentence initially (when these words begin with a capital letter) and sentence internally (when these words begin with a lower case letter).

4. Statistical analysis

Once the values of the seven adverb phrase position variables were computed for each of the 200 city sub-corpora, the 200 values for each variable were analyzed for regional patterns using two measures of spatial autocorrelation: global Moran's *I* and local Getis-Ord *Gi**. Spatial autocorrelation is a measure of the degree to which the values of a variable are similar at nearby locations (Cliff and Ord 1973). In order to determine the degree to which high and low values cluster across the entire distribution of each variable, global spatial autocorrelation was measured using global Moran's *I* (Moran 1948). In order to determine the location of high and low value clusters in the distribution of each variable, local spatial autocorrelation was measured using local Getis-Ord *Gi** (Ord and Getis 1995).[8]

Calculating both measures of spatial autocorrelation involves comparing pairs of values in the spatial distribution of a single variable. These comparisons are weighted based on the location of the values that are being compared, so that comparisons between locations that are closer together are given greater weight than comparisons between locations that are farther apart. This is accomplished by using a *spatial weighting function*, which is a set of rules that assigns a weight to every pair of locations in the spatial distribution of a variable based on proximity (Odland 1988).[9] Various spatial weighting functions are possible, although two functions are most common. A *binary weighting function* assigns a weight of 1 to all pairs of locations that are within a certain distance and a weight of 0 to all other pairs of locations (Odland 1988). A *reciprocal weighting function* assigns a weight to all pairs of locations by taking the reciprocal of the distance between the locations, so that weighting decreases with distance (Odland 1988). In this study, a binary weighting function with a 500 mile cutoff was used.

A binary weighting function was selected because it is simpler and more basic than a reciprocal weighting function. The results of calculating spatial autocorrelation using a binary weighting function are thus easier to interpret because the analysis has a clear level of resolution, which is equivalent to the cutoff distance. Nonetheless, other spatial weighting functions were tested. A 500 mile cutoff was selected because the cutoff distance needs to be large enough so that each city can be compared to nearby cities. Since the most spatially isolated city in the corpus is Billings, Montana, it was necessary to select a cutoff distance that would allow Billings to be compared to nearby cities. A 500 mile radius was selected because this allows Billings to be compared to 12 other cities. A 400 mile radius, on the other hand, would only allow for 6 cities to be compared to Billings. A 500 mile cutoff was also selected because it is large enough to allow for cities in the same traditional dialect region (Carver 1987; Labov et al. 2006) and cultural region (Zelinsky 1973) to be compared. For example, the distance between Savannah and Biloxi (on the edges of the Deep South) is approximately 470 miles, the distance between Bellingham and Medford (on the edges of the Pacific Northwest) is approximately 440 miles, and the distance between Bismarck and Duluth (on the edges of the Upper Midwest) is approximately 410 miles.

In order to determine the degree to which high and low values cluster in the distribution of each variable, each variable was tested individually for global spatial autocorrelation. Significant global spatial autocorrelation exists when the values of a variable are distributed in a spatial pattern, such as the regional clustering of high and low values (Cliff and Ord 1973, 1981; Odland 1988). Measuring global spatial autocorrelation is a way to objectively determine if the values of a variable are distributed in a regional pattern. The global spatial autocorrelation statistic global Moran's *I* (Moran 1948) was used to test each of the variables individually for spatial patterns at the global level.[10] The value of Moran's *I* ranges from −1 to 1, where a significant negative value indicates that neighboring data points tend to have different values, an insignificant value (approaching zero) indicates that neighboring data points tend to have random values, and a significant positive value indicates that neighboring data points tend to have similar values. In order to interpret the value of global Moran's *I*, a standardized *z*-score was obtained.[11] The z-score was interpreted as significant if it was larger than or equal to ±2.69, because this z-score corresponds to a .007 alpha level, which was selected because there are seven variables being analyzed, requiring that a standard .05 alpha level be adjusted using the Bonferroni correction (.05/7 = .007143). A significant positive z-score for global Moran's *I* indicates that neighboring locations have similar values at a greater degree than would be expected by chance.

In addition to measuring global spatial autocorrelation, local spatial autocorrelation was measured for each variable, in order to identify specific regional patterns. Unlike global spatial autocorrelation, which returns one value for each variable indicating the degree of spatial clustering across the *entire* regional distribution of that variable, a measure of local spatial autocorrelation returns one value for *each location* for each variable indicating the degree to which that particular location is a part of a high or low value cluster in the distribution of that variable. The statistic local Getis-Ord *Gi** (Ord and Getis 1995) was used to measure local spatial autocorrelation for each location for each variable by comparing the value of the variable at that location to the value of the variable at other locations in order to determine the degree to which that location is a member of a cluster of high or low values.[12] The z-score was interpreted as significant if it was larger than or equal to ±2.69 because this is the corresponding z-score value for a 0.007 alpha level, which was selected based on the Bonferroni correction described above. A significant positive z-score indicates that that location is part of a high value cluster and a significant negative z-score indicates that that location is part of a low value cluster.

Once computed, the Getis-Ord *Gi** z-scores were mapped for each location for each variable. This produces a map that displays the z-scores for each of the 200 cities, which indicates whether each city is a member of a statistically significant high value cluster, a statistically significant low value cluster, or a region of variability/transition. Mapping the Getis-Ord *Gi** z-values therefore allows for the locations of high and low value clusters to be statistically identified in the distribution of each variable. In essence, calculating the Getis-Ord *Gi** z-score values for each location in the distribution of a variable filters the data, identifying statistically significant regional patterns that were not obvious in the raw values. Mapping the Getis-Ord *Gi** z-score values therefore permits regional patterns to be easily identified.[13]

5. Results

This section presents the results of the statistical analysis of the values of the 7 adverb position variables. For each variable, the raw values were plotted across the 200 city sub-corpora.[14] The variable was then subjected to the two autocorrelation analyses. Global Moran's *I* was calculated in order to test if each variable exhibits a significant degree of high and/or low value clustering. Local Getis-Ord *Gi** z-scores were then calculated and mapped for each location for each variable in order to identify the locations of statistically significant high and low value clusters.

5.1 *Split adverbs*

Figure 2 maps the raw values for *infinitive splitting*. *Infinitive splitting* does not exhibit significant autocorrelation across the 200 city sub-corpora at the global level, as indicated by an insignificant negative value for global Moran's I

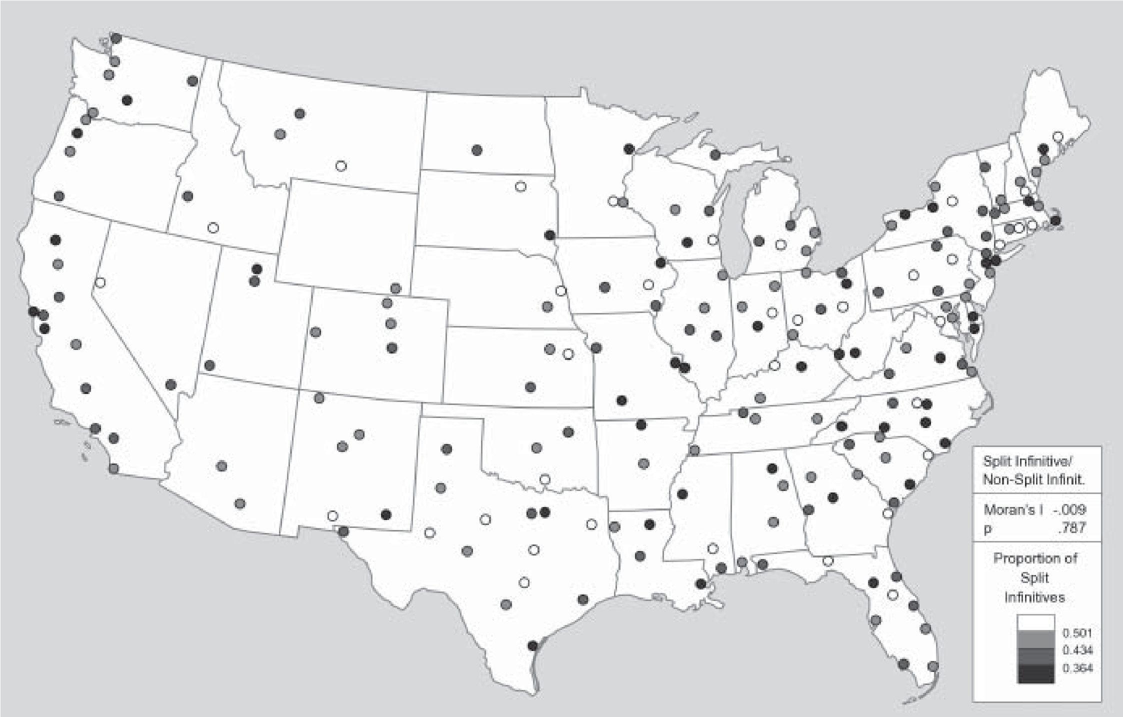

Figure 2. *Infinitive Splitting Raw Values*

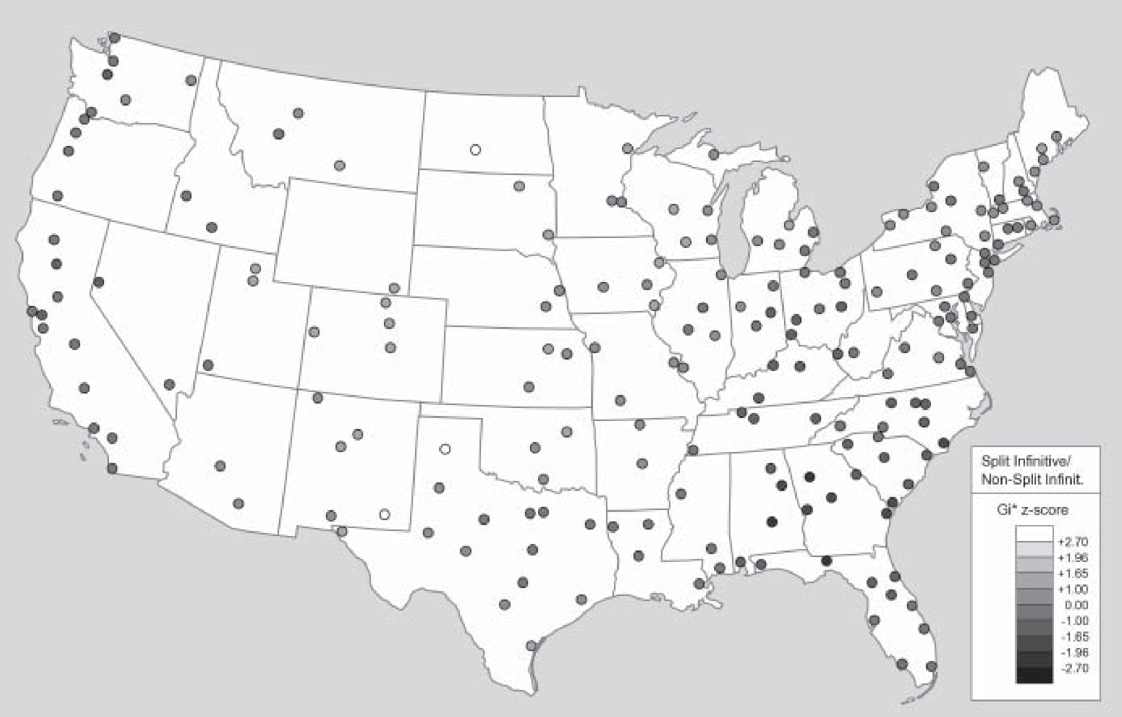

Figure 3. *Infinitive Splitting Getis-Ord Gi* z-scores*

(I = −0.009, p = 0.787). Figure 3 maps the local Getis-Ord Gi* z-score at each location. The Getis-Ord *Gi** analysis identified very limited clustering in the distribution of *infinitive splitting*.

Figure 4 maps the raw values for *non-modal auxiliary splitting*. *Auxiliary splitting* does not exhibit significant autocorrelation across the 200 city sub-corpora at the global level, as indicated by an insignificant positive value for global Moran's I (I = 0.004, p = 0.497). Figure 5 maps the local Getis-Ord Gi* z-score at each location. The Getis-Ord *Gi** analysis identified limited clustering in the distribution of *auxiliary splitting*. Low values were found to cluster in the western Midwest, indicating that *auxiliary splitting* is relatively infrequent in this region. However, this cluster is relatively small and the rest of the United States is characterized by variable *auxiliary splitting*.

Figure 6 maps the raw values for *modal splitting*. *Modal splitting* does exhibit significant autocorrelation across the 200 city sub-corpora at the global level, as indicated by a significant positive value for global Moran's I (I = 0.052, p = 0.000). Figure 7 maps the local Getis-Ord Gi* z-score at each location. The Getis-Ord *Gi** analysis identified considerable clustering in the distribution of *modal splitting*. Low values were found to cluster in the western Midwest and the Central States, indicating that *modal splitting* is relatively infrequent in these regions. High values were found to cluster in the Northeast, especially the Mid Atlantic, indicating that *modal splitting* is relatively frequent in this region.

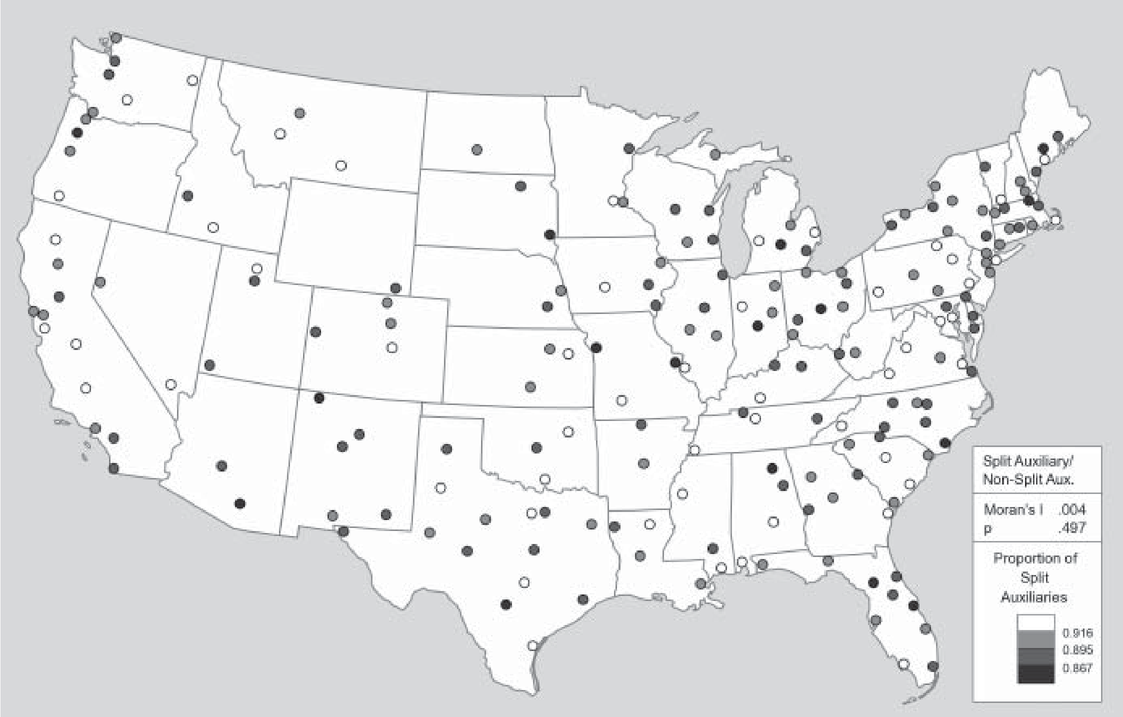

Figure 4. *Non-Modal Auxiliary Splitting Raw Values*

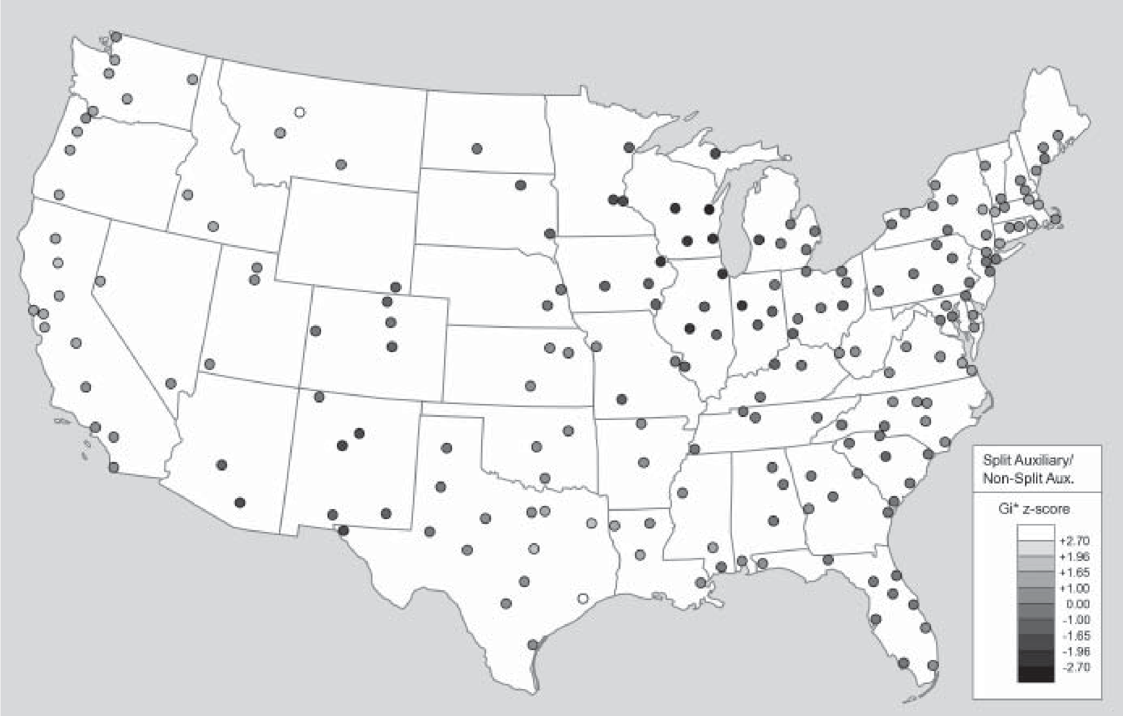

Figure 5. *Non-Modal Auxiliary Splitting Getis-Ord Gi* z-scores*

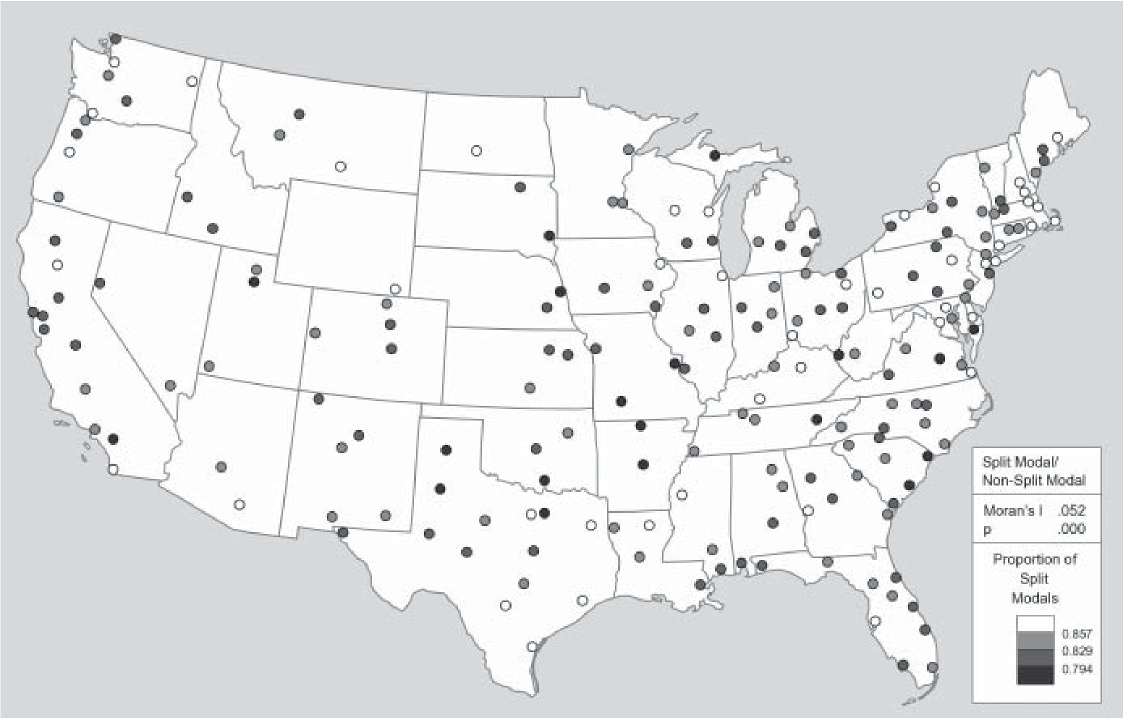

Figure 6. *Modal Splitting Raw Values*

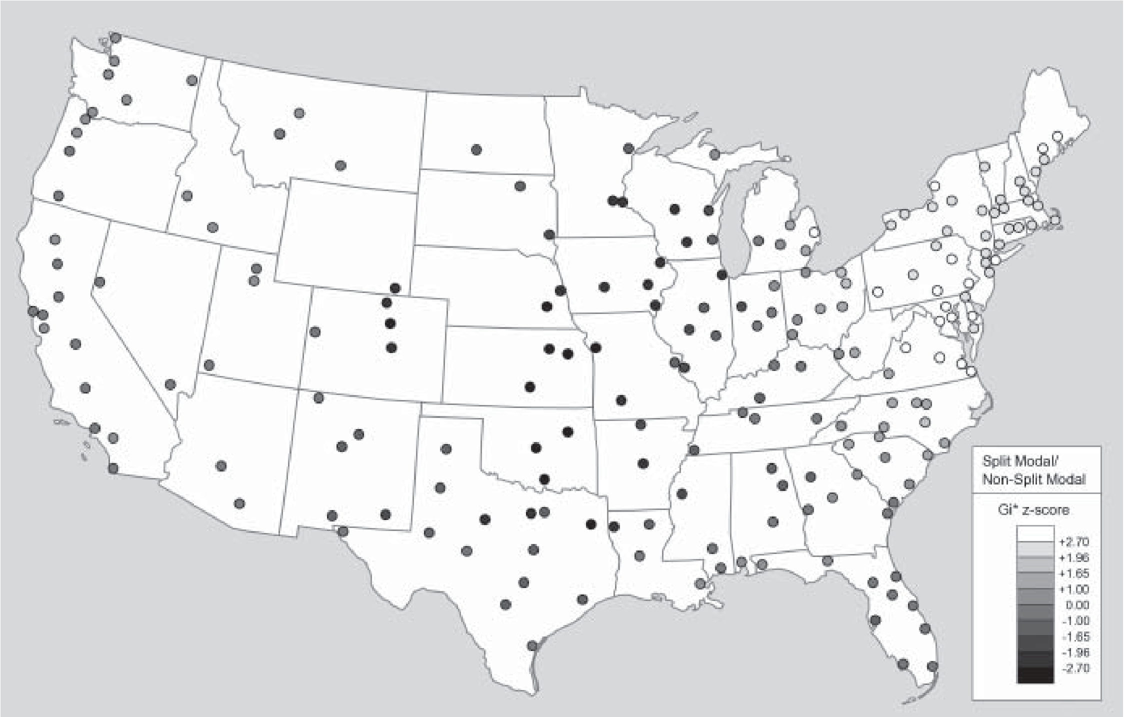

Figure 7. *Modal Splitting Getis-Ord Gi* z-scores*

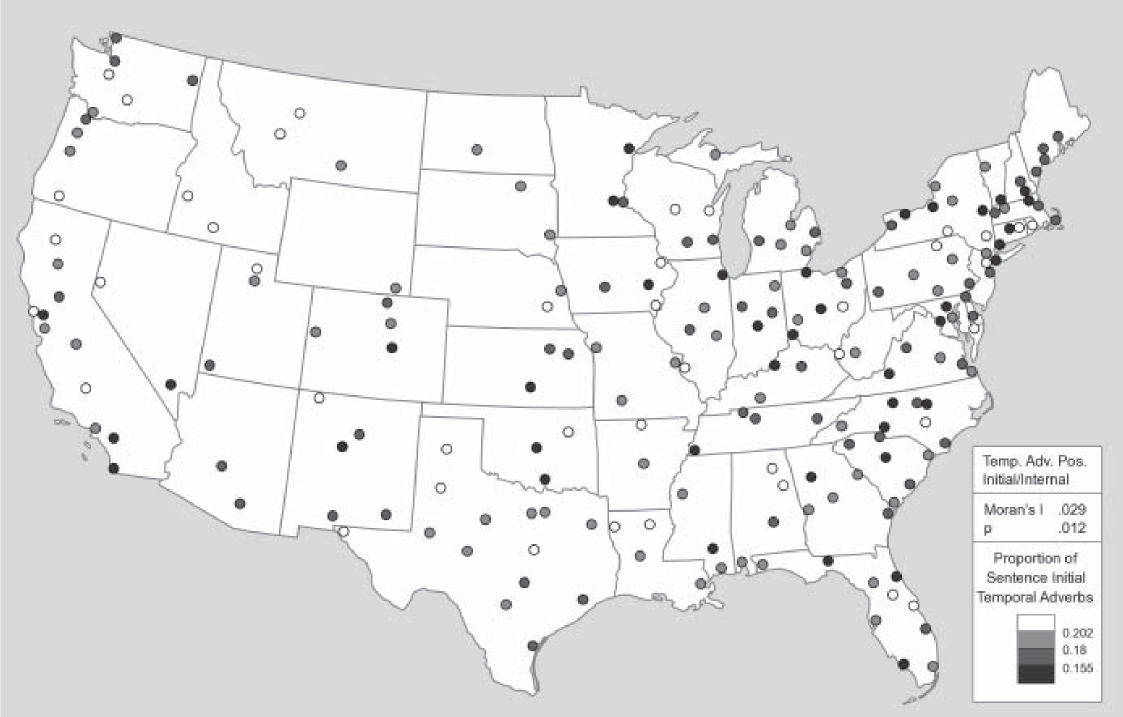

Figure 8. *Temporal Adverb Position Raw Values*

5.2 *Sentence-initial adverbs*

Figure 8 maps the raw values for sentence initial temporal adverbs. *Temporal adverb position* does not exhibit significant autocorrelation across the 200 city sub-corpora at the global level, as indicated by an insignificant positive value for global Moran's I ($I = 0.029$, $p = 0.012$). Figure 9 maps the local Getis-Ord Gi* z-score at each location. However, reflecting the nearly significant level of global spatial autocorrelation, the Getis-Ord *Gi** analysis did identify clustering in the distribution of *temporal adverb position*. Low values were found to cluster in the Northeast, eastern Ohio, West Virginia, Virginia, and the Carolinas, indicating that sentence internal temporal adverbs are relatively frequent in these regions. High values were found to cluster to a limited extent in the South Central States and the Northwest, indicating that sentence initial temporal adverbs are relatively frequent in these regions.

Figure 10 maps the raw values for sentence initial *however*. *However position* does exhibit significant autocorrelation across the 200 city sub-corpora at the global level, as indicated by a significant positive value for global Moran's I ($I = 0.043$, $p = 0.000$). Figure 11 maps the local Getis-Ord Gi* z-score at each location. The Getis-Ord *Gi** analysis identified considerable clustering in the distribution of *however position*. Low values were found to cluster in the Northeast, and to a lesser extent across the Midwest, indicating that sentence internal *however* is relatively frequent in these regions. High values were

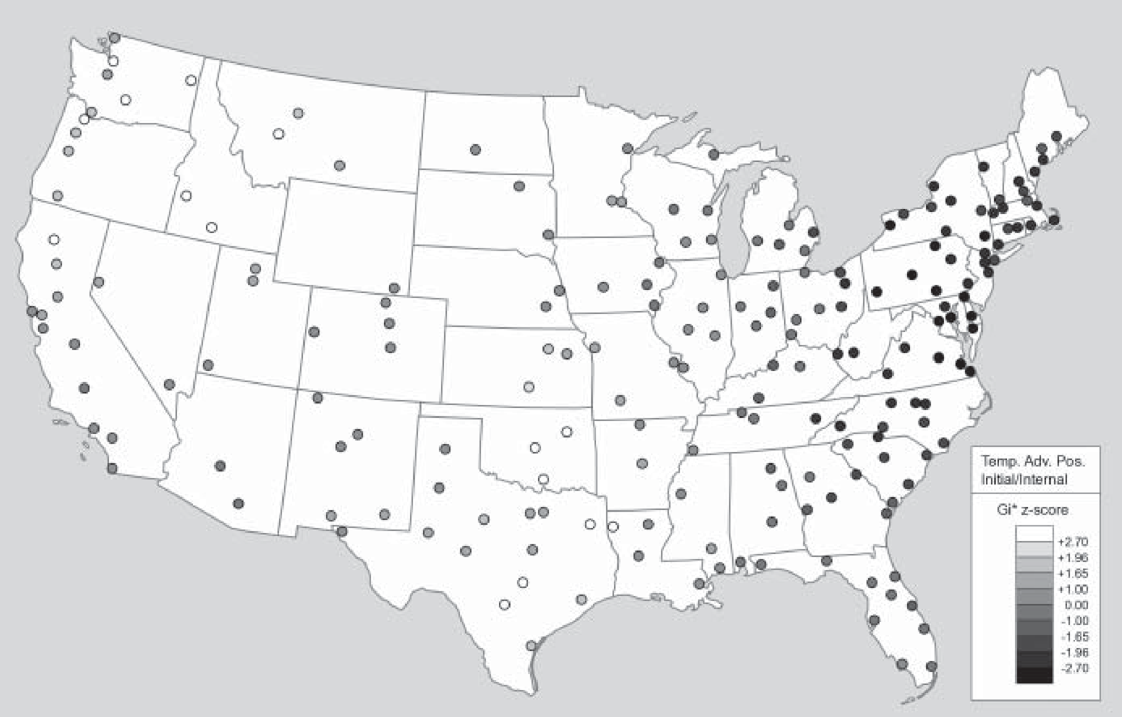

Figure 9. *Temporal Adverb Position Getis-Ord Gi* z-scores*

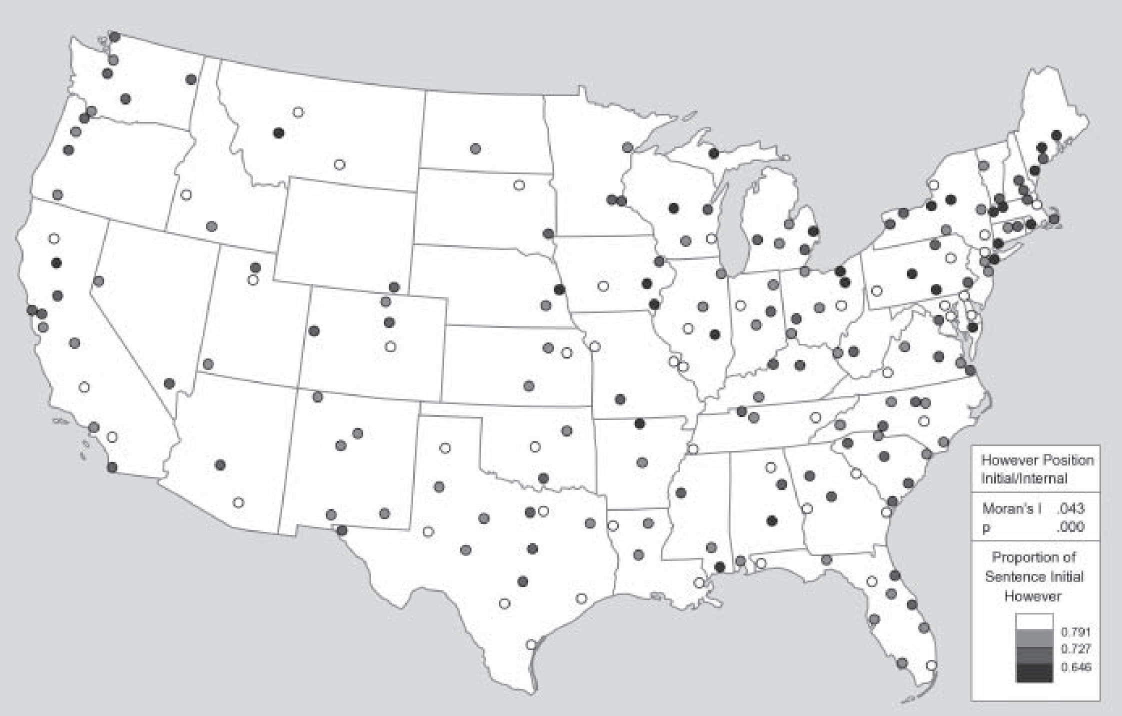

Figure 10. However *Position Raw Values*

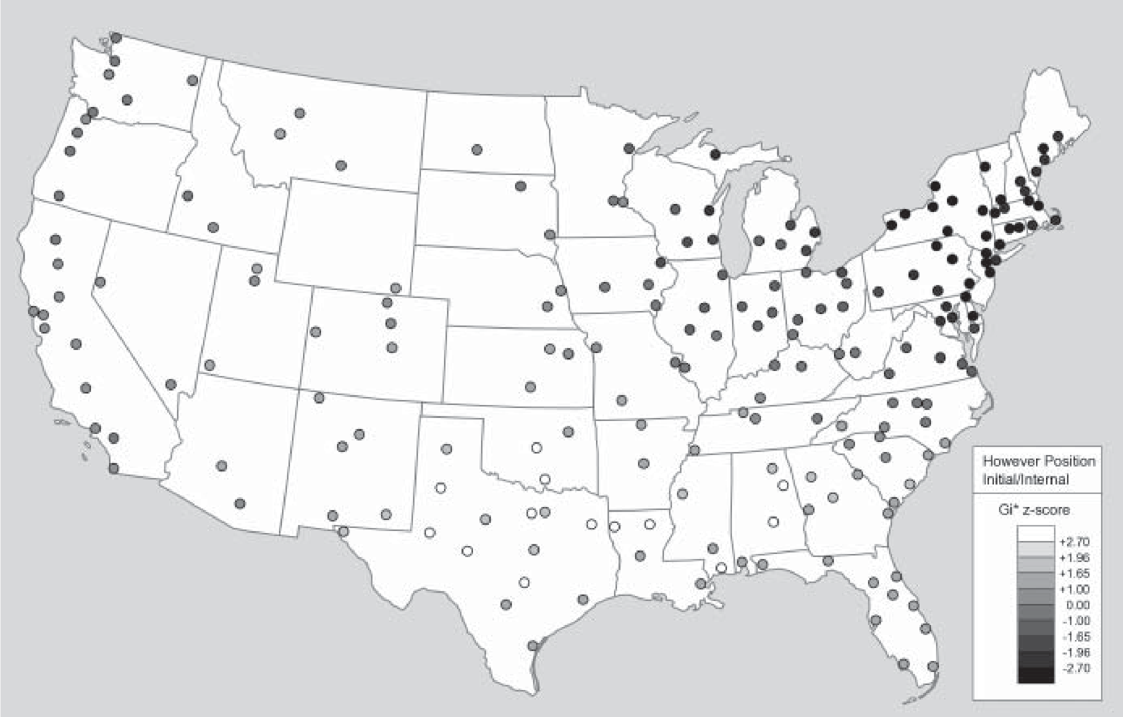

Figure 11. However *Position Getis-Ord Gi* z-scores*

found to cluster in the South Central States, and to a lesser extent across the Southeast, indicating that sentence initial *however* is relatively frequent in these regions.

Figure 12 maps the raw values for sentence initial *also*. *Also position* does exhibit significant autocorrelation across the 200 city sub-corpora at the global level, as indicated by a significant positive value for global Moran's I ($I = 0.057$, $p = 0.000$). Figure 13 maps the local Getis-Ord Gi* z-score at each location. The Getis-Ord *Gi** analysis identified considerable clustering in the distribution of *also position*. Low values were found to cluster in the Northeast, and to a lesser extent across the northern Midwest, indicating that sentence internal *also* is relatively frequent in these regions. High values were found to cluster in the Southeast and the South Central States, indicating that sentence initial *also* is relatively frequent in these regions.

Figure 14 maps the raw values for sentence initial *instead*. *Instead position* does not exhibit significant autocorrelation across the 200 city sub-corpora at the global level, as indicated by an insignificant positive value for global Moran's I ($I = 0.028$, $p = 0.016$). Figure 15 maps the local Getis-Ord Gi* z-score at each location. However, reflecting the nearly significant level of global spatial autocorrelation, the Getis-Ord *Gi** analysis did identify clustering in the distribution of *instead position*. Low values were found to cluster in the South Central States, the Central States, and in the southwestern Midwest, indicating that sentence internal *instead* is relatively frequent in these regions.

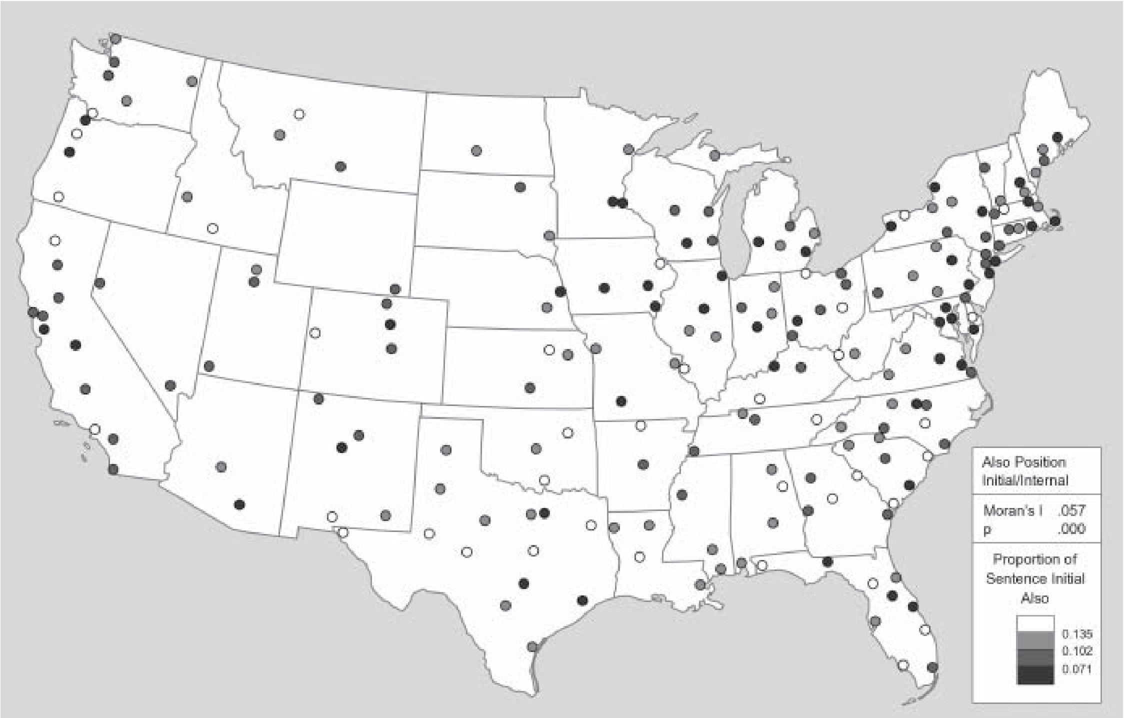

Figure 12. Also *Position Raw Values*

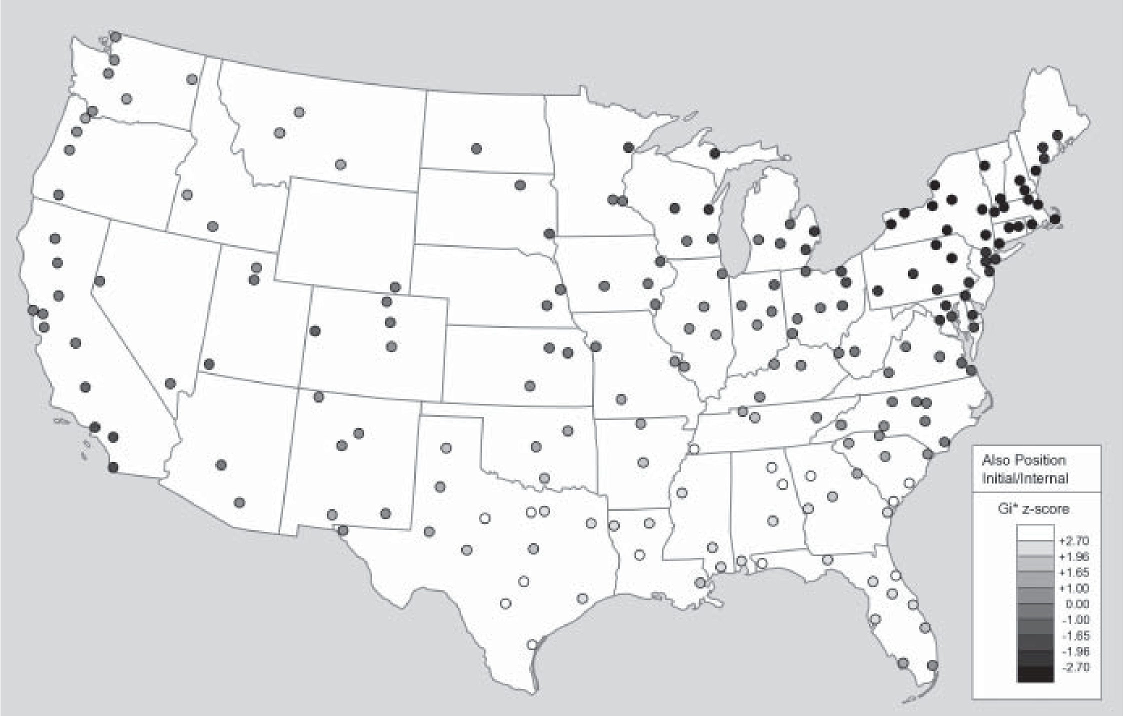

Figure 13. Also *Position Getis-Ord Gi* z-scores*

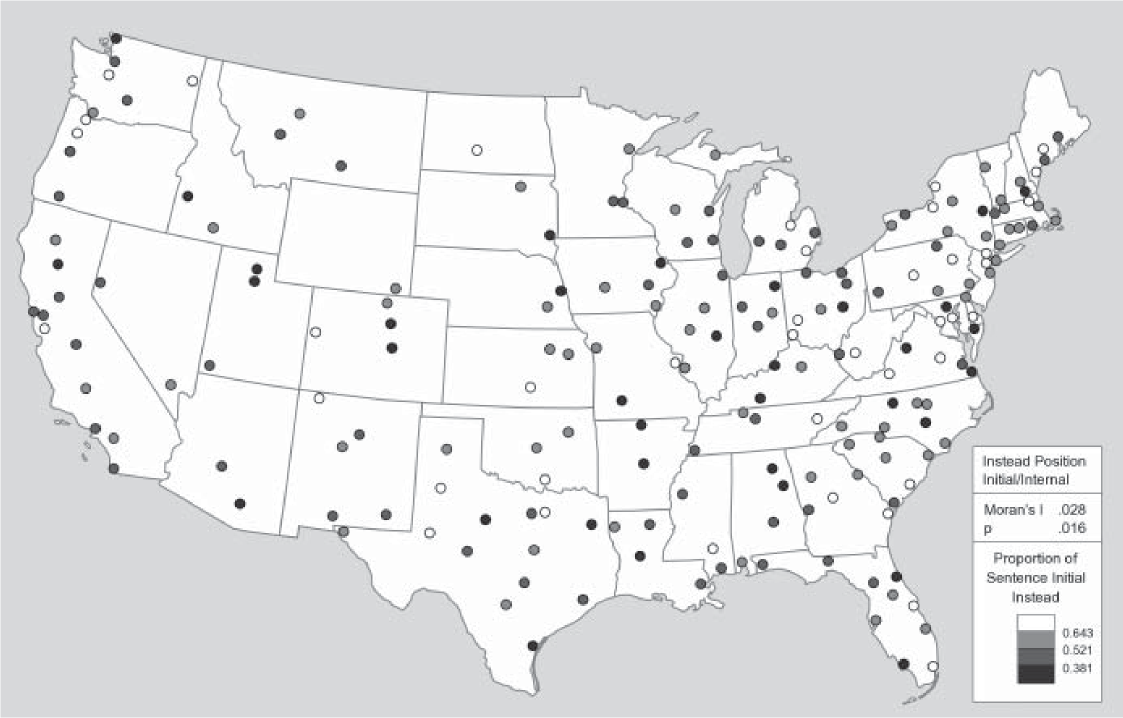

Figure 14. Instead *Position Raw Values*

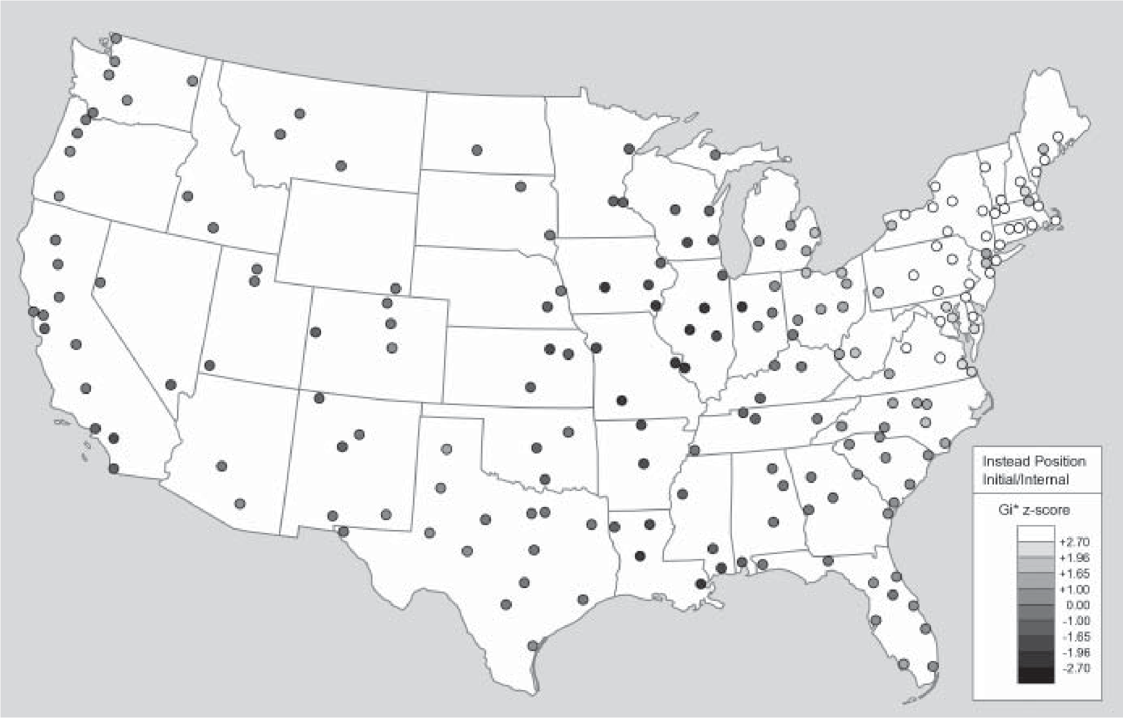

Figure 15. Instead *Position Getis-Ord Gi* z-scores*

High values were found to cluster in the Northeast and Virginia indicating that sentence initial *instead* is relatively frequent in these regions.

6. Discussion

Based on the analysis of global spatial autocorrelation, three of the seven adverb position variables were identified as exhibiting statistically significant regional patterns: *modal splitting, however position,* and *also position.* In addition, two other adverb position variables were found to exhibit nearly statistically significant regional patterns: *temporal adverb position* and *instead position*. These results demonstrate that adverb position is regionally patterned in written Standard American English and, by extension, that continuous regional grammatical variation exists in American English and that regional linguistic variation exists in written Standard American English.

In addition, all five significant or nearly significant measures of adverb position were found to exhibit similar regional patterns through an analysis of local spatial autocorrelation. To generalize, these variables were patterned in manner that contrasts the language of the Northeast, as well as the Midwest to a lesser degree, with language of the rest of the United States, especially the South Central States, as well as the Southeast and, to a lesser degree, the Central States. This pattern is most clearly expressed in the regional distributions for *however position* and *also position.* Sentence initial *however* and *also* are relatively frequent in the Southeast and South Central States and sentence internal *however* and *also* are relatively frequent in the Northeast and the Midwest. Although it does not exhibit a statistically significant level of global spatial autocorrelation, *temporal adverb position* also exhibits a similar pattern. Sentence initial temporal adverbs are relatively frequent in the South Central States, as well as the Northwest, and sentence internal temporal adverbs are relatively frequent in the Northeast. Similarly, although it does not exhibit a statistically significant level of global spatial autocorrelation, *instead position* follows a similar pattern as well. Sentence initial *instead* is relatively frequent in the Northeast and sentence internal *instead* is relatively frequent in the Central States, although the Southeast is identified as a region of variability. Finally, the regional pattern is also similar for *split modals* – the third variable exhibiting a statistically significant level of global spatial autocorrelation. Split modals are relatively frequent in the Northeast and relatively infrequent in the Central States, including the South Central States; however, similar to *instead position*, the Southeast is identified as a region of variability.

What is most striking about these results is the consistent identification of a Northeast dialect region, which always includes New England, New York, Pennsylvania, New Jersey, Delaware, Maryland, and Washington, and which often extends into Virginia, Ohio, Michigan, and sometimes into West Virginia and North Carolina and across the Midwest. This consistent identification of a Northeast dialect region offers strong support for the claim that measures of adverb position are regionally patterned in American English. The region that contrasts with the Northeast is not identified as consistently, shifting primarily between the Southeast and the Central States, although this region does seem to be focused on the South Central States: Louisiana, Mississippi, Alabama, Arkansas, Oklahoma, and eastern Texas. It should also be noted that both global and local spatial autocorrelation were calculated based on other cutoff distances as well as a reciprocal weighting function. These alternative analyses produced very similar results to the findings reported here, offering additional evidence of the robustness of these patterns. Future research, however, should investigate the consequences of using different spatial weighting functions for regional dialect studies in a more systematic manner.

Overall, the regional patterns uncovered by this study are similar to the findings of previous American dialect surveys, which identified clear lexical and phonological differences between the language of the North and the South in the eastern United States (e.g., Kurath 1949; Atwood 1953; Kurath and McDavid 1961; Carver 1987; Labov et al. 2006). The results of this study, however, do not directly support the tripartite division on the American East Coast endorsed by the majority of these surveys (e.g. Kurath 1949; Atwood 1953; Kurath and McDavid 1961; Labov et al. 2006), which identify a Midland dialect region between the Northeast and Southeast dialect regions. The results of this study are more similar to the findings of Carver (1987), who identified only two major dialect regions in the eastern United States. However, the northeast region identified by this study is even larger as it stretches into Virginia and North Carolina – states whose eastern halves are always classified as being part of the southeast dialect region. Furthermore, the focal point of the southern region identified by this study seems to be the South Central States, not the Southeastern States, which are traditionally associated with Southern English. Despite these differences, the results of this study are still largely in line with the findings of previous American dialect surveys, which always identify linguistic difference between the North and the South in the eastern United States.

Because of the similarities between the results of this study and the results of previous American dialect surveys, it would appear that the same basic explanation invoked by dialectologists to explain regional linguistic variation in American English based on historical settlement patterns can also explain the basic regional linguistic patterns observed here. This explanation, however, cannot explain why the Northeast region has been found to extend into Virginia and North Carolina, as northerners did not originally settle this region. Rather, this southern expansion of the Northeast dialect region seems to be a result of modern migration patterns, specifically the recent influx of northerners, especially from New York and Pennsylvania, into Virginia and North Carolina (Perry 2003; U.S. Census Bureau 2005).[15] This migration from the Northeast dialect region has undoubtedly affected the language of Virginia and North Carolina, and is thus most likely responsible for the southward expansion of the Northeast dialect region observed here. It is also important to note that traditional methods of data collection, which focus on the language of long-term residents, cannot identify recent changes in the location of dialect regions. This is not necessarily a problem, as long as it is acknowledged that the traditional method of data collection only allows for the identification of historical dialect regions, but it is also important to acknowledge that historical settlement patterns are not the only determinant of regional linguistic variation, as the results of this study demonstrate.

Aside from the purely regional linguistic patterns identified by this study, an important regional functional linguistic pattern has also been identified: it appears that the language of the northeastern United States is more formal than the language of the southeastern and central United States, in terms of adverb position. This formality pattern is particularly evident in the case of the two most significant sentence initial adverb variables. Placing *however* and *also* sentence internally rather than sentence initially is a characteristic of formal English. This is also probably true of placing temporal adverbs sentence internally. While following the opposite pattern structurally, the regional distribution of *instead position,* although not statistically significant, also follows the same functional pattern: placing *instead* sentence initially as opposed to sentence internally appears to be the more formal choice. Similarly, *modal splitting,* which did exhibit significant regional patterning, seems to follow this same basic formality pattern, for while adverb splitting is often identified as an ungrammatical construction in traditional prescriptive grammar, it is very common in modern English and is actually associated with a dense informational style, as opposed to a more conversational and thus informal style (Biber, 1988). It is unclear, however, why the other two adverb splitting variables analyzed in this study, which do not exhibit significant regional patterns, do not follow a similar pattern. This regional formality pattern also seems to make intuitive sense, as the language of the Northeast is generally seen as being more proper than the language of the rest of the United States, especially the South.

Finally, the positive results of this study support the claim that corpus-based dialectology is a viable approach to the analysis of regional linguistic variation. Furthermore, it is clear that a corpus-based approach allows for certain issues to be addressed that could not be easily addressed using traditional methods of data collection. In particular, the corpus-based approach has allowed for continuous grammatical variation to be analyzed, which is difficult to accomplish when data is collected through the linguistic interview. In addition, the corpus-based approach has allowed for regional linguistic variation (at the sub-national level) to be observed in a variety of the English language where regional variation has not been observed in the past: written Standard English.[16] The finding that regional variation exists in written Standard English is particularly important because it has been assumed that regional variation and sociolinguistic variation in general cannot exist in such formal registers (Schneider 2002), especially written language. More generally, these findings imply that the extent of regional linguistic variation is much greater than has been previously assumed. Indeed, because we know that regional linguistic variation exists, based on a long history of regional dialect studies, it seems that the most important task in modern dialectology is to identify the limits of regional linguistic variation – to determine just how pervasive regional linguistic variation is in natural language. As this study has demonstrated, a corpus-based approach to data collection, coupled with a statistical approach to data analysis, constitutes an appropriate method for testing the limits of regional linguistic variation.

Acknowledgements

I would like to thank Doug Biber, Bill Crawford, Eniko Csomay, Dirk Geeraerts, Ray Huang, Randi Reppen, Tom Ruette, Benedikt Szmrecsanyi, Dirk Speelman, Joeri Theelen, Emily Waibel, and an anonymous reviewer for their comments on this paper and their comments on the approach to the analysis of regional linguistic variation presented in this paper.

Bionotes

Jack Grieve is a lecturer in forensic linguistics at the Centre for Forensic Linguistics in School of Languages and Social Sciences at Aston University in Birmingham, UK. His research focuses on the quantitative analysis of language variation and change. Email: Jack.Grieve@arts.kuleuven.be

Notes

1. The use of the linguistic interviews is especially common in American dialectology, where all major studies have been based on data gathered by interviewing informants, whether in person, by phone or by mail (Hempl 1896; Kurath et al. 1939; Davis 1948; Kurath 1949; Atwood 1953; Marckwardt 1957; Kurath and McDavid 1961; Atwood 1962; Allen 1973;

McDavid and O'Cain 1979; Cassidy 1985; Pederson 1986; Carver 1987; Kretzschmar et al. 1993; Labov et al. 2006).

2. Often dialectologists refer to collections of transcribed and/or recorded linguistic interviews as "corpora" but these datasets do not constitute true corpora because these datasets do not represent an actual variety of natural language. These datasets are also subject to the Observer's Paradox (Labov 1972), as the informants were aware that their language was being recorded for a dialect survey. This is not the case in true corpus-based linguistics.
3. The corpus was cleaned and organized using computer programs written by the author using the programming language Perl. Feature counting and the statistical analyses described below were also conducted using programs written by the author.
4. The common adverbs that were considered are *alike, together, aside, sooner, seldom, regardless, irrespective, altogether, thereof, alas, only, really, ere, awhile, downright, somewhere, else, etc, thereby, albeit, nary, anymore, elsewhere, afterwards, furthermore, moreover, afterward, somehow, quite, beforehand, tomorrow, today, yesterday, nowhere, bloody, quite, earlier, everywhere, now, almost, then, next, ever, ago, perhaps, forever, so, besides, meanwhile, sometime, sometimes, often, instead, anytime, anywhere, anyhow, anyway, anyways, indeed, already, once, twice, rather, otherwise, yet, someday, maybe, also, as well, as such, too, very, even, now, later, soon, never, just, always, then, here, therefore, thus, however, regardless, nonetheless, again,* and *someday* (Biber et al. 1999).
5. The common infinitive verbs that were considered are *be, have, do, say, get, make, go, know, take, see, come, think, look, want, give, use, find, tell, ask, work, seem, feel, try, leave, call, serve, reduce, address, thank, check, share, fight, identify, hear, express, show, condemn, return, mention, keep, spend, watch, inform, deal, recognize, set, bring, vote, appreciate, pass, support, add, complete, agree, expand, speak, enter, save, die, talk, fix, suggest, drive, abuse, buy, throw, provide, stand, care, ignore, endorse, affect, investigate, engage, increase, understand, turn, impact, examine, run, communicate, move, attempt, end, reach, monitor, implement, force, become, follow, lead, respond, encourage, protect, eliminate, work, sell, promote, develop, start, assess, execute, limit, educate, prepare, cross, let, report, remind, demonstrate, stop, cover, disagree, handle, meet, defend, hold, allow, enforce, separate, begin, improve, put, discuss, attack, disregard, remove, kill, display, consider, pursue, request, sit, write, decide, oppose, honor, criticize, acknowledge, believe, accept, debate, change, walk, manage, realize, influence, step, point, pull, pay, ask, receive, fund, cut, read, apply, maintain, evaluate, review, participate, enjoy, represent, raise, continue, carry, learn, listen, live, dispose, forget, help,* and *destroy* (Biber et al. 1999).
6. The common irregular verbs were considered are *arisen, arose, ate, awoke, bade, beat, beaten, became, become, began, begun, bent, bet, bid, bide, bit, bitten, bled, blown, bore, borne, bought, bound, bred, broke, broken, brought, built, burnt, burst, came, cast, caught, chose, chosen, clung, come, cost, crept, cut, dealt, drank, drawn, dreamt, drew, driven, drove, drunk, dug, dwelt, eaten, fallen, fed, fell, felt, fled, flew, flown, flung, forbade, forbidden, forebade, forebidden, forecast, foresaw, foreseen, foretold, forgave, forgiven, forgot, forgotten, fought, found, froze, frozen, gave, given, gone, got, gotten, grew, grown, heard, held, hid, hidden, hit, hung, hurt, kept, knealt, knew, knit, known, laid, lain, lay, leant, leant, leapt, learnt, led, left, let, lit, lost, made, meant, met, misled, mistaken, mistook, misunderstood, overcame, overcome, overdid, overdone, overran, overridden, overrode, overrun, oversaw, overseen, overtaken, overthrew, overthrown, overtook, paid, partaken, partook, proven, put, quit, ran, rang, read, rid, ridden, risen, rode, rose, run, rung, said, sang, sank, sat, saw, sawn, seen, sent, set, shaken, shed, shod, shook, shot, shown, shrank, shrunk, shut, slept, slid, slit, slung, slunk, smelt, sold, sought, sown, span, spat, sped, spelt, spent, spilt, spit, split, spoilt, spoke, spoken, sprang, spread, sprung, spun, stink, stole, stolen, stood, stridden, striven, strode, strove, struck, strung, stuck, stunk, sung, sunk, swam, swept, swore, sworn, swum,*

swung, taken, taught, thought, threw, thrown, thrust, told, took, tore, torn, trod, undergone, understood, undertaken, undertook, underwent, undid, undone, upheld, upset, warn, wed, went, wept, withdrawn, withdrew, withheld, withstood, woke, woken, won, wound, wove, woven, written, wrote, and *wrung* (Biber et al. 1999).

7. The temporal adverbs that were analyzed are *currently, often, lately, afterwards, again, earlier, eventually, later, initially, instantly, momentarily, now, nowadays, presently, previously, recently, shortly, simultaneously,* and *soon* (Biber et al. 1999).
8. Although continuous measures of spatial autocorrelation have not been used in American dialect studies, Lee and Kretzschmar (1993) used the joint count statistic, which is a measure of spatial autocorrelation for categorical data, to analyze the LAMSAS data.
9. In order to implement a spatial weighting function, the distance between every pair of cities in the corpus was determined using the longitudes and latitudes for the cities provided by the U.S. Census Bureau and the great circle distance formula (Sinnott 1984).
10. The formula for calculating global Moran's *I* is provided in Equation 2.

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2} \tag{2}$$

Where N is the total number of locations, x_i is value of the variable at location i, x_j is value of the variable at location j, $\bar{x}$ is the mean for the variable across all locations, and w_{ij} is the value of the spatial weighting function for the comparison of location x_i and x_j ($w_{ij} = 1$ if $distance_{ij} \leq 500$ miles, $w_{ij} = 0$ if $distance_{ij} > 500$ miles or if $i = j$).

11. In order to interpret the significance of Moran's *I*, a standardized *z*-score is obtained (under the assumption of randomization) using Equations 3 through Equation 10 (Odland 1988).

$$z = \frac{I - E(I)}{\sqrt{Var(I)}} \tag{3}$$

$$E(I) = \frac{-1}{N-1} \tag{4}$$

$$Var(I) = \frac{NS_4 - S_3S_5}{(N-1)(N-2)(N-3)\left(\sum_i \sum_j w_{ij}\right)^2} \tag{5}$$

$$S_1 = \frac{\sum_i \sum_j (w_{ij} + w_{ji})^2}{2} \tag{6}$$

$$S_2 = \sum_i \left(\sum_j w_{ij} + \sum_j w_{ji}\right)^2 \tag{7}$$

$$S_3 = \frac{1/N \sum_i (x_i - \bar{x})^4}{\left(1/N \sum_i (x_i - \bar{x})^2\right)^2} \tag{8}$$

$$S_4 = (N^2 - 3N + 3)S_1 - NS_2 + 3\left(\sum_i \sum_j w_{ij}\right)^2 \tag{9}$$

$$S_5 = S_1 - 2NS_1 + 6\left(\sum_i \sum_j w_{ij}\right)^2 \tag{10}$$

Where $E(I)$ is the expected value for Moran's I, $Var(I)$ is the variance for Moran's I, and w_{ji} is the value of the spatial weighting function for the comparison of location x_j and x_i, which is equal to w_{ij} ($w_{ij} = 1$ if $distance_{ij} \leq 500$ miles, $w_{ij} = 0$ if $distance_{ij} > 500$ miles or if $i = j$).

12. The formulae for calculating local Getis-Ord Gi* z-scores are provided in Equation 11 and Equation 12 (Ord and Getis 1995).

(11)
$$G_i^* = \frac{\sum_j w_{ij} x_j - \bar{x} \sum_j w_{ij}}{S \sqrt{\frac{N \sum_j w_{ij}^2 - \left(\sum_j w_{ij}\right)^2}{N-1}}}$$

(12)
$$S = \sqrt{\frac{\sum_j x_j^2}{N} - \bar{x}^2}$$

Where N is the total number of locations, x_i is value of the variable at location i, x_j is value of the variable at location j, $\bar{x}$ is the mean for the variable across all locations, and w_{ij} is the value of the spatial weighting function for the comparison of location x_i and x_j ($w_{ij} = 1$ if $d_{ij} \leq 500$ miles, $w_{ij} = 0$ if $d_{ij} > 500$ miles or $i = j$).

13. The Getis-Ord *Gi** maps associate each location with one of seven levels of z-scores: z-scores larger than or equal to ±2.69, corresponding to the adjusted .007 alpha level; z-scores larger than or equal to ±1.96, corresponding to the standard .05 alpha level; z-scores larger than or equal to ±1.65, corresponding to a .10 alpha level; z-scores larger than or equal to ±1.00; and z-scores larger or smaller than 0. In these maps, high positive z-scores (i.e. light shades) indicate that the primary variant of the variable occurs relatively frequently in that area (i.e. adverb splitting or sentence initial adverbs occur relatively frequently); alternatively, high negative z-scores (i.e. dark shades) indicate that the primary variant of the variable occurs relatively infrequently in that area (i.e. non-adverb splitting or sentence internal adverbs occur relatively frequently). By identifying clusters of locations with highly positive or negative Getis-Ord *Gi** z-scores (especially z-scores larger than or equal to ±2.69 and, to a lesser extent, larger than or equal to ±1.96), regional patterns can be identified in the distribution of each variable.
14. In the raw value maps, the cutoffs for the different levels of the variable were selected based on the overall mean value of the variable (second cutoff) and the mean values of the variable for values that are lower than (first cutoff) and higher than (third cutoff) the overall mean value of the variable.
15. According to the 2000 Census (U.S. Census Bureau 2005), the top five sources of migration into Virginia from other U.S. states are (with percentage of total American born residents who were born out of state in parentheses) New York (11.1%), North Carolina (8.5%), Pennsylvania (8.4%), Washington D.C. (6.6%), and Maryland (5.9%), all of which, except for its southern neighbor North Carolina are north of Virginia. Similarly, although the trend is not quite as pronounced, the top five sources of migration into North Carolina from other U.S. states are New York (12.1%), Virginia (9.9%), South Carolina (8.3%), Pennsylvania (6%), and Florida (5.3%). On the other hand, a more southern pattern of migration is found in the states identified in other southern states. For example, the top five sources of migration into South Carolina are North Carolina (15.5%), Georgia (11.3%), New York (10.1%), Pennsylvania (5.6%), and Florida (5.2%). This southern pattern is even stronger in a state like Alabama, whose top five sources of migration are Georgia (13.6%), Florida (9.7%), Mississippi (9.2%), Tennessee (7.8%), and Texas (5%).
16. Strictly speaking, studies of national variation in Standard English do not qualify as regional dialect studies (i.e. they have not identified regional variation) as the independent variables in such studies are not geographically defined (e.g. longitude, latitude, distance) but rather are politically defined.

References

Allen, Harold B. 1973–1976. *The linguistic atlas of the Upper Midwest.* Minneapolis: University of Minnesota Press.

Atwood, E. B. 1953. *A survey of verb forms in the Eastern United States*. Ann Arbor: University of Michigan Press.

Atwood, E. B. 1962. *The regional vocabulary of Texas*. Austin: University of Texas Press.

Biber, Douglas. 1988. *Variation across speech and writing.* Cambridge: Cambridge University Press.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman grammar of spoken and written English.* Harlow: Longman.

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

Cassidy, Frederic G. 1985. *Dictionary of American regional English.* Cambridge, MA: Harvard University Press.

Cliff, Andrew D. & John K. Ord. 1973. *Spatial autocorrelation*. London: Pion.

Cliff, Andrew D. & John K. Ord. 1981. *Spatial processes: Models and applications*. London: Pion.

Davis, Alva L. 1948. *A Word atlas of the Great Lakes region.* Ph.D. Dissertation. Ann Arbor, MI: University of Michigan dissertation.

Hempl, George. 1896. Grease and greasy. *Dialect Notes* 1. 438–444.

Herrmann, Tanja. 2005. Relative clauses in English dialects of the British Isles. In Bernd Kortmann (ed.), *Dialectology meets typology: Dialect grammar from a cross-linguistic perspective,* 479–496. Berlin: Mouton de Gruyter.

Hernandez, Nuria. 2006. User's guide to FRED: Freiburg Corpus of English Dialects. English Dialect Research Group. Albert-Ludwigs-Universitat Freiburg.

Ingham, Richard. 2006. On two negative concord dialects in early English. *Language Variation and Change* 18. 241–266.

Ihalainen, Ossi. 1976. Periphrastic DO in affirmative sentences in the dialect of east Somerset. *Neuphilologische Mitteilungen* 77. 608–622.

Ihalainen, Ossi. 1980. Relative clauses in the dialect of Somerset. *Neuphilologische Mitteilungen* 81. 187–196.

Ihalainen, Ossi. 1985. He took the bottle and put 'n in his pocket: the object pronoun IT in present-day Somerset. In Wolfgang Viereck (ed.), *Focus on: England and Wales,* 153–161. Amsterdam: John Benjamins.

Ihalainen, Ossi. 1990. A source of data for the study of English dialect syntax: the Helsinki Corpus. In Jan Aarts J & Willen Meijs (eds.), *Theory and practice in corpus linguistics,* 83–103. Amsterdam: Rodopi.

Ihalainen, Ossi. 1991a. On grammatical diffusion in Somerset folk speech. In Peter Trudgill & Jack K. Chambers (eds.), *Dialects of English: Studies in grammatical variation,* 104–119. London: Longman.

Ihalainen, Ossi. 1991b. A point of verb syntax in south-western British English: an analysis of a dialect continuum. In Karin Aijmer & Bengt Altenberg (eds.). *English corpus linguistics: Studies in honour of Jan Svartvik,* 290–302. London: Longman.

Ihalainen Ossi, Merja Kytö & Matti Rissanen. 1987. The Helsinki corpus of English texts: Diachronic and dialectal. Report on work in progress. In Willem Meijs (ed.), *Corpus linguistics and beyond*, 21–32. Amsterdam: Rodopi.

Kortmann, Bernd (ed.). 2004. *Dialectology meets typology: Dialect grammar from a cross-linguistic perspective*. Berlin: Mouton de Gruyter.

Kortmann, Bernd, Tanja Herrmann, Lukas Pietsch & Susanne Wagner. 2005. *A Comparative Grammar of British English Dialects*. Berlin: Mouton de Gruyter.

Kretzschmar William A., Virginia G. McDavid, Theodore K. Lerud & Ellen Johnson. 1993. *Handbook of the linguistic atlas of the Middle and South Atlantic States*. Chicago, IL: University of Chicago Press.

Kurath, Hans. 1949. *Word geography of the eastern United States.* Ann Arbor, MI: University of Michigan Press.

Kurath, Hans, Marcus L. Hansen, Bernard Bloch & Julia Bloch. 1939–1943. *Linguistic atlas of New England*. Providence, RI: Brown University Press.

Kurath, Hans & Raven I. McDavid. 1961. *The pronunciation of English in the Atlantic States.* Ann Arbor, MI: University of Michigan Press.

Labov, William. 1966a. *The social stratification of English in New York City.* Washington, DC: Center for Applied Linguistics.

Labov, William. 1966b. The linguistic variable as a structural unit. *Washington Linguistics Review* 3. 4–22.

Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia, PA: University of Pennsylvania Press.

Labov, William, Sharon Ash & Charles Boberg. 2006. *Atlas of North American English: Phonetics, phonology, and sound change.* New York: Mouton de Gruyter.

Lee, Jay & William A. Kretzschmar. 1993. Spatial analysis of linguistic data with GIS functions. *International Journal of Geographical Information Systems* 7. 541–560.

Marckwardt, Albert H. 1957. Principal and Subsidiary Dialect Areas in the North Central States. *PADS* 27. 3–15.

McCafferty, Kevin. 2003. The northern subject rule in Ulster: How Scots, how English? *Language Variation and Change* 15. 105–139.

McDavid, Raven I. & Raymond K. O'Cain. 1979. *Linguistic atlas of the middle and south Atlantic states*. Chicago, IL: University of Chicago Press.

Moran, P. A. P. 1948. The interpretation of statistical maps. *Journal of the Royal Statistical Society, Series B* 37. 243–251.

Odland, John D. 1988. *Spatial Autocorrelation*. Beverly Hills, CA: Sage Publications.

Ojanen, Anna-Liisa. 1985. Use and non-use of prepositions in spatial expressions in the dialect of Cambridgeshire. In Wolfgang Viereck (ed.), *Focus on: England and Wales*, 179–212. Amsterdam: John Benjamins.

Ord, J. K. & A. Getis. 1995. Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis* 27. 286–306.

Pederson, Lee. 1986–93. *Linguistic Atlas of the Gulf States* (7 Volumes). Athens, GA: University of Georgia Press.

Peitsara, Kristi. 1988. On existential sentences in the dialect of Suffolk. *Neuphilologische Mitteilungen* 1. 72–99.

Perry, Marc J. 2003. State to state migration flows: 1995 to 2000. Census 2000 Special Reports. CENSR-8. Available at http://www.census.gov/prod/2003pubs/censr-8.pdf

Schneider, Edgar W. 2002. Investigating variation and change in written documents. In Jack K. Chambers, Peter Trudgill & Natalie Schilling-Estes (eds.), *The handbook of language variation and change,* 67–96. London: Blackwell.

Schneider, Edgar W., Michael B. Montgomery. 2001. On the trail of early nonstandard grammar: An electronic corpus of southern U.S. antebellum overseers' letters. *American Speech* 76. 388–412.

Sinnott, R. W. 1984. Virtues of the Haversine. *Sky and Telescope* 68. 159.

Szmrecsanyi, Benedikt. 2010. Corpus-based dialectometry: aggregate morphosyntactic variability in British English dialects. *Corpora* 6.

U.S. Census Bureau. 2005. State of Residence in 2000 by State of Birth. PHC-T-38. Available at http://www.census.gov/population/www/socdemo/migrate/2000pob.html

Van Herk, Gerard & James A. Walker. 2005. S marks the spot? Regional variation and early African American correspondence. *Language Variation and Change* 17. 113–131.

Wagner, Susanne. 2004. ‘Gendered’ Pronouns in English Dialects – a Typological Perspective. In Bernd Kortmann (ed.), *Dialectology meets typology: Dialect hrammar from a cross-linguistic perspective,* 479–496. Berlin: Mouton de Gruyter.

Wolfram, Walt. 1969. A sociolinguistic description of Detroit negro speech. Washington, DC: Center for Applied Linguistics.

Wolfram, Walt. 1991. The linguistic variable: fact and fantasy. *American Speech* 66. 22–32.

Zelinsky, Wilbur. 1973. *Cultural geography of the United States*. Englewood Cliffs, NJ: Prentice-Hall.