Citation:

Exploring Regional Variation in Spatial Language Using Spatially-Stratified Web-Sampled Route Direction Documents

Sen Xu, Alexander Klippel, Alan MacEachren
GeoVISTA Center, Department of Geography, Pennsylvania State University, PA, USA

Prasenjit Mitra
College of Information Science and Technology, Pennsylvania State University, PA, USA
Department of Computer Science and Engineering, Pennsylvania State University, PA, USA
Abstract

Spatial language, such as route directions, can be analyzed to shed light on how humans communicate and conceptualize spatial knowledge. This article details a computational linguistic approach using route directions to study regional variations in spatial language. We developed a web-sourcing approach to collect human-generated route direction documents on a geographical scale. Specifically, we built the Spatially-strAtified Route Direction (SARD) Corpus through automated scraping, classifying, and georeferencing of route directions. Based on semantic categories of cardinal and relative direction terms, the analysis of the SARD Corpus reveals significant differences and patterns on both national (U.S., U.K., and Australia) and regional (contiguous U.S. States) levels. Combining computational linguistics and georeferencing approaches offers the potential for extending classic spatial linguistic studies.

Keywords: spatial language analysis, corpus linguistics, georeferenced web sampling, regional linguistic difference, cardinal vs. relative direction
1. Introduction

Route directions have become an important medium through which the representation, perception, and communication of spatial information is studied. For instance, analyzing route directions with respect to their linguistic characteristics and structure provides insight into how spatial knowledge is represented and communicated (Allen, 1997; Allen, 2000; Daniel & Denis, 1998; Daniel & Denis, 2004; Denis, 1997; Eschenbach et al., 2002; Gryl et al., 2002; Hölscher et al., 2011; Klein, 1983). With respect to differences in spatial language, it has long been noted that 1) different languages can represent the same spatial information differently (Boroditsky, 2001; Boroditsky et al., 2010; Burenhult & Levinson, 2008; Lamarre, 2008; Munnich et al., 2001); and 2) that these differences can occur even within the same language across different regions (Davies & Pederson, 2001; Ishikawa & Kiyomoto, 2008; Liu, 2008; Montello & Xiao, 2011; Zelinsky, 1955). Regional differences in language can manifest in accent, vocabulary, and spelling. Zelinsky (1955), for example, explored regional variations using generic terms in place-names and revealed examples of regionalization in spatial language usage. Different terms referring to waterbodies (e.g., brook, creek, branch, fork, stream) and agglomerated settlements (e.g., -ville, corners, center, village, -burg, town, city) were found to have distinctive spatial distributions. To interpret such distributions, Zelinsky pointed out the correlation between settlement cultures and spatial language usage (e.g., Anglicism and archaic terms were found predominantly in the New England region). Another study by Davies and Pederson (2001) compared Milton Keynes, U.K., and Eugene, OR, U.S., and found European irregular and non-aligned street grids can foster the use of landmarks in route directions.

In general, the physical environment of a route may influence human route descriptions. For example, it would be very difficult to use cardinal directions on curvy roads. Continue
driving north on X street would be confusing when X street is not northbound. Elevation changes have been employed in conveying route information: when there is substantial elevation change in a route, the route directions can contain go uphill or continue downhill to assist navigation (Couclelis, 1996; Schubert, 2006). Finally, cultural differences may also contribute to the way people talk about space—hence affecting route directions. Take the Guugu Yimithirr people (an Australian Aboriginal ethnic group) as an extreme example: there are no relative directions in their language, while cardinal directions are used even in indoor environments and with objects, such as on the southern edge of the western table (Deutscher, 2010; Haviland, 1998). Languages that primarily rely on cardinal reference frames can be found around the globe (Levinson, 1997). Montello and Xiao (2011) examined the Internet corpora and found the Mandarin Chinese language uses cardinal direction twice more frequently (over 2,000 per million) than English (less than 1,000 per million). Even in English-speaking regions, it is not unusual to find that people from certain regions prefer to use cardinal directions over relative ones. By analyzing spatial language usage systematically through a corpus with wide spatial coverage, the spatial distribution of language differences (or similarities) could help us reach a deeper understanding of where and why such phenomena exist, and to build the foundation for a better localized route direction generation system.

One challenge in analyzing regional linguistic variation is the definition of the elusive term region (Guelke, 1977). For this article, we defined two levels of analysis: the national-level (i.e., comparing U.S., U.K., and Australia) and the regional-level (i.e., states in the contiguous

---

1 Human interpretation of cardinal directions with respect to roads is very tolerant. Instructions such as Following US-1 North do not require US-1 to head north consistently.
U.S. such as Pennsylvania, New York). A second challenge is collecting reliable data that represents multiple regions. So far, existing spatial language studies collect data mostly by recruiting human participants. These studies provide a valuable understanding of spatial cognition associated with spatial language (Davies & Pederson, 2001; Ishikawa & Kiyomoto, 2008; Montello et al., 1999; Tenbrink & Winter, 2009; Ward et al., 1986). While the results from these studies are insightful, we note that the data collection procedure is expensive, time consuming, and therefore often spatially constrained. Exploring regional spatial language variation would benefit from a methodology in which spatial language is collected and analyzed on a geographic scale (Montello, 1993). Scaling up the dataset to a language corpus that is spatially stratified and spans different nations is challenging, and calls for new data sources and collection methods. Analysis of large quantities of linguistic data intended to reveal regional patterns also requires automated text processing and visual analytics toolkits to assist in the interpretation.

Researchers have experimented with using the Web as an alternative source of data for empirical studies. Jones at al. (2008) used queries such as hotels in <toponym> to harvest georeferenceable information from documents on the Web. This trend of utilizing various kinds of volunteered or pseudo-volunteered spatial data from the Web shows the power of crowdsourcing (Rattenbury & Naaman, 2009), which provides a valuable alternative to the current data collection method in many research fields. Stock et al. (2013) developed a set of geospatial syntactic templates for building a spatial natural language corpus with web data; this corpus reached a precision of 0.66 (i.e., 66% of the final corpus is evaluated as geospatial phrases). In this article, we lay out a framework for semi-automatically collecting and analyzing spatially stratified route directions sampled from the Web; then we present results that show
statistically significant spatial patterns for the dominant use of relative over cardinal directions, and vice versa.

The rest of the article is structured as follows: Section 2 provides background on challenges and potential benefits of sampling route directions from the Web. Section 3 provides a step-by-step workflow detailing how we built our Spatially-strAtified Route Direction (SARD) corpus. Section 4 presents the text analysis tools, analysis schema and results reference frame use comparing regions and nations. Section 5 discusses results, possible extensions, and implications. Section 6 lays out future work.

2. Background

2.1 Scraping the Web for Route Directions

It is crucial to understand basic features of the target documents, that is, human-generated route directions on the web, to better design data collection and analysis schemata. Figure 1 shows a typical route direction webpage. Most route direction documents have a postal address of the destination on the same page with the route direction text (from a random set of 100 web route direction documents examined, 86 had postal codes). Route directions on the Web are usually written to assist web site visitors to navigate to a desired destination. In this article, the postal code in the destination address is used as a geo-stamp for the linguistic characteristic in route directions.
A major challenge in using the Web to collect route direction samples is that it is not straightforward to automatically identify the target document (route directions) from other, less relevant types of documents. To resolve this problem, text classification has been adopted, as it has been previously shown to perform well for classifying human-generated route directions from general web documents (Xu et al., 2009; Xu et al., 2010; Zhang et al., 2009).

Another challenge is georeferencing route directions: Linguistic phenomena need to be associated with regions for analyzing regional variation. Although there are geo-features in the form of street names and landmarks in route directions, the ambiguity of the available geo-features makes precise georeferencing a demanding challenge (Lee & Lee, 2007). For example, a search for Main Street on OpenStreetMap will return over 3,000 instances in the United States.
In this study, we adopt an existing text classification solution and build a data pipeline that utilizes the high occurrence of postal codes for efficient data collection.

2.2 Analyzing Route Directions Focusing on Direction Terms

Our analysis of cardinal vs. relative direction usage is inspired by existing studies from various communities (Davies & Pederson, 2001; Ishikawa & Kiyomoto, 2008; Lawton, 2001; Montello & Xiao, 2011). We scope the data analysis to be focusing on the direction term usage difference on a geographic/regional scale, which is made possible through data collection on the Web. The first step in our analysis was to develop a set of semantic categories of route direction elements (see Daniel & Denis, 1998). As shown in Table 1, we identified three semantic categories for relative directions, and four semantic categories for cardinal directions. Note that both cardinal and relative directions can represent change of direction or static spatial relationship (bold in Table 1).
Table 1

*Semantic Categories for Cardinal and Relative Directions*

<table>
<thead>
<tr>
<th>Semantic categories</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative Direction</td>
<td></td>
</tr>
<tr>
<td>1. Change of direction</td>
<td><em>take a left, bear right</em></td>
</tr>
<tr>
<td>2. Static spatial relationship</td>
<td><em>the destination is left to #{landmark}</em></td>
</tr>
<tr>
<td>3. Driving aid</td>
<td><em>keep to the left lane, merge to the right lane</em></td>
</tr>
<tr>
<td>Cardinal Direction</td>
<td></td>
</tr>
<tr>
<td>1. Change of direction</td>
<td><em>veer southwest on U.S. Hwy 24, turn north</em></td>
</tr>
<tr>
<td>2. Static spatial relationship</td>
<td><em>2 blocks east of #{landmark}</em></td>
</tr>
<tr>
<td>3. Traveling direction</td>
<td><em>head north, traveling south</em></td>
</tr>
<tr>
<td>4. General origin</td>
<td><em>from North, coming from South of New York</em></td>
</tr>
</tbody>
</table>

In order to efficiently perform categorical analysis of direction usages, we used TermTree Tool (Turton & MacEachren, 2008), which supports determination of token occurrence count for phrases of interest. Additionally, the GeoViz Toolkit (Hardisty & Robinson, 2011) was used to translate the semantic analysis results into map visualizations that help evaluate regional patterns; GeoDa (Anselin et al., 2006) was used to examine the statistical significance of the spatial distribution.

In summary, we resolved the challenges of automatically classifying target documents by adopting an existing text classification method. Georeferencing is resolved by using the postal

---

2 Cardinal directions commonly occur as part of street names (e.g., West Main Street, I-101 North) and place names (e.g., East Village, North Star restaurant). We excluded these types of cardinal direction usages in this study.
code as a geo-stamp for each document, and a semantic categorical analysis schema has been
developed for analyzing cardinal vs. relative direction usage.

3. Building the SARD Corpus

We developed a comprehensive data pipeline to systematically collect and analyze route
direction documents scraped from the web. The major components are:

- Data collecting schema that scrapes a spatially-stratified sample of Web
documents;
- Text classifier that automatically and efficiently classifies route direction
documents from general web documents (Zhang et al., 2009; Zhang et al., 2012);
- Georeferencing that uses a postal code to locate the linguistic characteristics of a
  route direction document.

Figure 2 presents the data collection and processing workflow that will be discussed in
the following sections.
3.1 Data Collection Pipeline

The data pipeline is designed with text classification and georeferencing taken into consideration from the beginning. Due to the fact that postal codes often occur together with route directions on the Web, we adopted a postal-code-based data scraping schema together with a validation process.
3.1.1 Spatially-Stratified Sampling with Postal Codes. Postal codes offer low ambiguity and sufficient spatial resolution in this study, as our targeted smallest spatial unit is the state level. Postal codes also provide comprehensive spatial coverage, which is ideal for collecting a spatially stratified corpus.

A list of postal codes for the U.S. (41,119 ZIP codes selected for the contiguous U.S.3) was obtained for non-commercial usage (zipcodeworld.com, 2009, Jan). This list of postal codes was then used as seed data to build the SARD Corpus of the contiguous U.S. The most commonly used ZIP code format in the U.S. consists of state abbreviation followed by a space and five digits (e.g., PA 16802). For the U.K., postcodes were retrieved from www.freethepostcode.org;4 8860 postal codes were collected (freethepostcode.org, 2009, Apr). U.K. postcodes format consists of one or two letters followed by one or two digits, a space, then one or two digits and two letters (e.g., AB10 6BB). The list of postal codes for the U.K., although not officially complete, covers most areas of the U.K. For Australia, the official complete list containing 3,312 postal codes was obtained from the Australian Post (AustraliaPost, 2009, Apr).

3 The complete list of postal codes contains 42,293 ZIP codes in the Territories of the United States in 2009. The list of ZIP codes in the contiguous U.S. used in this study excludes locations with ZIP codes in AK (Alaska), HI (Hawaii), PR (Puerto Rico), VI (Virgin Islands), AS (American Samoa), GU (Guam), PW (Palau), FM (Federated States of Micronesia), MP (Northern Mariana Islands), MH (Marshall Islands) or Military District (AE, AA, AP).

4 The Ordnance Survey released the full postcode database under an open license on May 10th, 2010, after this study was conducted.
Australian postcode format consists of region abbreviation followed by three or four digits (e.g., ACT 2610).

To obtain documents from the Web with postal codes, we designed a web scraper that used custom queries on a search engine to collect raw web documents; the process is as follows:

1. Construct query: Read one entry from the list of postal codes. The keyword “directions” is added to the query in order to bring more targeted route direction documents to the top of the return hits.

2. Analyze the first N hits from the search result. Check the previously collected URL in order to avoid duplicate data collection.

3. Repeat steps 1-2 until all postal codes have been used to query websites.

3.1.2 Document Classification and Location Validation. A rule set combined with machine-learning-based text classification (Zhang et al., 2009) was implemented to identify human generated route directions from scraped documents. In brief, over 1,000 route direction documents (positive training set) and non-route direction documents (negative training set) were fed to a Maximum Entropy classifier, which uses a variety of features to classify new documents. To achieve high precision that ensures the validity of the analysis, several iterations of training and classifying were carried out. False positives were put into the negative training set to re-train, in order to improve the precision of the classifier. Based on a hand evaluation of the classification result for over 900 documents, we found that the text classification performance converged after three iterations to a precision of 93% in correctly identifying human generated route direction documents from web documents.

The subsequent analyses addressed two granularities: national and regional (e.g., individual states in case of the contiguous U.S., see Table 2). We introduced a location validation
step to ensure that every region included documents containing postal codes from only the region in question. Documents that contain postal codes from more than one region were removed from the corpus to ensure the validity of the analyses.

Table 2

Two Scales for Analysis: National-level and Regional-level

<table>
<thead>
<tr>
<th>National</th>
<th>Regional (contiguous U.S. States)</th>
</tr>
</thead>
<tbody>
<tr>
<td>U.S. (contiguous), U.K., Australia</td>
<td>AL, AR, AZ, CA, CO, CT, DC, DE, FL, GA, IA, ID, IL, IN, KS, KY, LA, MA, MD, ME, MI, MN, MO, MS, MT, NC, ND, NE, NH, NJ, NM, NV, NY, OH, OK, OR, PA, RI, SC, SD, TN, TX, UT, VA, VT, WA, WI, WV, WY</td>
</tr>
</tbody>
</table>

3.2 Characteristics of the SARD Corpus

The resulting SARD Corpus 1) covers the contiguous U.S., the U.K., and Australia; 2) has region-level document organization (collection is conducted at the level of postal codes); 3) covers route directions in different environments (urban, rural, highway). Table 3 gives an overview of the SARD Corpus characteristics. Figure 3 provides information on its spatial coverage. Unsurprisingly, the figure shows that route direction documents are more frequently collected from populated (or urbanized) regions, and that route direction documents were not obtained from some ZIP codes (e.g., 2,275 of 10,240 ZIP codes in the U.S., most of which are areas of low population).
Table 3

Attributes of the SARD Corpus (Spatially-strAtified Route Direction Corpus)

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus topic</td>
<td>Route Directions</td>
</tr>
<tr>
<td>Document format</td>
<td>HTML</td>
</tr>
<tr>
<td>Language</td>
<td>English</td>
</tr>
<tr>
<td>Spatial coverage</td>
<td>The contiguous U.S., the U.K. and Australia</td>
</tr>
<tr>
<td>Size (total of documents)</td>
<td>11,439 documents (10,240 in the U.S., 710 in the U.K., and 489 in Australia)⁵</td>
</tr>
<tr>
<td>Purity (percentage of true route directions)</td>
<td>93%</td>
</tr>
<tr>
<td>Organization</td>
<td>National-level — Regional-level—Postal code-level</td>
</tr>
</tbody>
</table>

⁵ The difference in number of documents retrieved from the three nations could have come from differences in population, habits of posting route directions online or not, whether the postal code was included in the route direction, or other factors.
4. Case Study: Cardinal/Relative Direction Usage

In this section, we present our tools, schemata, and results for analyzing regional similarity and variation on both national and regional levels.

4.1 Analysis Schema

To explore regional variations of cardinal vs. relative direction usages in the SARD Corpus, we combined semantic analysis with geovisualization tools to develop a thorough analysis schema. First, the semantic analysis proposed in Table 1 was carried out using TermTree, which provides token occurrence counts of phrases containing cardinal and relative directions. The occurrence counts of every semantic category from different regions were then further analyzed using GeoViz and GeoDa.
TermTree (Turton, 2008; Turton & MacEachren, 2008, Figure 4) takes custom wild cards as a query and delivers tree structure visualizations of the contextual information of target phrases. This interface provides an analyst with both flexible context information (to determine the semantic category), and the convenience of counts for a particular phrase. It enables the analyst to load all route direction documents from the same regional directory and put the following queries in the query box:

\[ w \cdot (left|right) \cdot w \]
\[ w \cdot (north|south|east|west|northeast|southeast|northwest|southwest) \cdot w \]

The above query captures three word phrases with the center word being either a relative direction or cardinal direction term. Then the analyst can read from the tree-structured result and use the context to determine to which semantic category one phrase belongs, and record the number of occurrences of this phrase to the semantic category occurrence in this region.
We used *token occurrence* (how many times a token occurs in documents from one region) and *proportion of usage* (token occurrence divided by sum of token occurrence)—common measures in corpus linguistics—to compare directional term usages in different regions. For example, Pennsylvania (PA) has 796 route direction documents; these documents are fed into TermTree and analyzed using the two queries above. After examining phrases in TermTree and aggregating the token occurrence count, we found 5,113 relative directions (*left* and *right*) that are being used to represent *change of direction*. This token occurrence count is one of the seven regional linguistic characteristics (see Table 1) in the state of Pennsylvania. The proportion of usage is indicated by the normalized token occurrence within relative or cardinal directions.

4.2 Data Analysis Setup

Our analysis compares the different semantic usages within cardinal and relative reference frames (see Table 1 for definitions of semantic categories); of particular interest are the two semantic categories shared by both reference frames (*change of direction* and *static spatial relationship*). The semantic categorization was coded by one coder (whose first language is Chinese but fluent in English) and validated by a second coder (first language English). The validation was carried out on a subset of the corpus (10 states out of the contiguous 48). The intercoder reliability, Krippendorff’s alpha (Krippendorff, 1970), is over 0.99, which is high enough to ensure valid result as “content analysis researchers generally think of K > .8 as good reliability” (Carletta, 1996, p.252). Results on both national and regional levels are shown below.

4.3 Cardinal vs. Relative Direction: National Level

Figure 5 and 6 illustrate direction term usages across the contiguous U.S., the U.K., and Australia. The token occurrences show that relative direction terms are predominantly used to indicate *change of direction* (Figure 5). In contrast, cardinal directions are mostly used to
indicate *travelling direction* (Figure 6). Comparing the three countries, the lowest proportion of cardinal directions used for *static spatial relationship* is in the U.K., while the U.K. has the highest proportion of cardinal directions used for *general origin*. The proportions of the semantic category usages from the U.S. and Australia seem to be more similar to each other, compared to those from the U.K.

*Figure 5.* National-level relative direction usages (bar chart: token occurrence, pie chart: proportion).
To compare cardinal and relative directions directly, we analyzed the token occurrences of change of direction (Figure 7) and static spatial relationship (Figure 8). As shown in Figure 7, the preference for expressing change of direction with relative direction terms (e.g., turn left instead of turn north) is present in all three nations. Change of direction is specified using cardinal direction terms for less than 1% of the time for the U.K and for around 2% for the U.S. and Australia (dark thin slice in Figure 7).

*Figure 6.* National-level cardinal direction usages (bar chart: token occurrence, pie chart: proportion).
Both cardinal and relative directions can also be used to describe static spatial relationships (e.g., the destination will be on your left, the hotel is to the north of the church). In this semantic category, relative directions are dominant. Relative directions take up nearly 90% of static spatial relationship usages in the U.K., while in the U.S. and Australia the proportion is around 75% (white slice in Figure 8).
Figure 8. National-level direction term composition when the describer wants to express static spatial relationship (bar chart: token occurrence count, pie chart: proportion).

To summarize, the three nations share some linguistic characteristics, but also show variations. Despite the difference in corpus size, route direction documents from the U.S. and Australia appear to be more similar while those from the U.K. vary somewhat. From the semantic category comparison, relative directions are more dominant in the U.K. than in the other two countries.

4.4 Cardinal vs. Relative Direction: Regional Level

Linguistic characteristics within every state in the contiguous U.S. were analyzed and compared in this section, providing insights into regional variation.

4.4.1 Validation of Number of Documents by State. The number of documents per state unsurprisingly correlates positively with population (see Figure 9). More populated states have more businesses (e.g., restaurants) that expect visitors—hence more route direction documents.

However, there are a few states with very few documents in the SARD Corpus (New Mexico, North Dakota, South Dakota have less than 30 documents in the final corpus after the
iterative scraping process). We first checked if the small numbers of documents in these states are outliers, using z-scores above/below 3.29 as suggested by Tabachnick & Fidell (2007). We found two positive outliers, California (z-score=4.19) and New York (z-score=3.67). Both states are popular destinations for tourism and business. Under the hypothesis that population correlates with the number of documents collected in a state, we checked the z-score of the DocPerState/Population ratio and found no outliers. This statistical examination shows that the small number of documents in certain states, in proportion to state population, is not statistically different than expected.

Figure 9. Correlation between document number per state and population per state (from U.S. Census 2010).

4.4.2 Mapping Cardinal vs. Relative Direction Usage in the U.S. GeoViz (Hardisty & Robinson, 2011) was used to produce the maps to show the proportion of cardinal and relative directions within the same semantic category. Figure 10 shows the breakdown of change of direction. The five quantiles are based on normalized values; lighter greys indicate states with a
higher proportion of cardinal directions (and lower proportion of relative directions), darker greys indicate states with a lower proportion of cardinal directions (and a higher proportion of relative directions). The middle gray, thus depicts states near the median for use of cardinal directions in the contiguous U.S. A clear distinction between eastern states and inland states can be observed: relative directions dominate in states east of the Mississippi River (most of these states are in the fourth and fifth quantile), compared to lower numbers west of the Mississippi (with Oregon the only western state above the median).

*Figure 10. Proportion of relative direction vs cardinal direction usage for expressing change of direction in the U.S. (Dark: more relative direction usage; light: more cardinal direction usage)*

Figure 11 shows the breakdown of static spatial relationship. The difference between coastal states and inland states is still visible, but not as pronounced as in Figure 10. We can still see that the states which use cardinal directions more frequently are inland (brightest quantile: AR, AL, CO, ID, IL, LA, SD, TX and UT).
In sum, the dominance of relative directions can be observed in all states for change of direction and in most states for static spatial relationship. For regional differences, inland states use cardinal directions for both change of direction and static spatial relationship more than eastern states, while the distinction is much more pronounced in change of direction.

4.4.3 Spatial Autocorrelation. We applied a Moran’s I (Moran, 1950) analysis to the contiguous U.S. corpus to assess the significance of the spatial pattern discussed in Section 4.4.2. From Table 4 we can see the Moran’s I values for the two shared semantic categories (change of direction and static spatial relationship). The Moran’s I value in the semantic category change of direction is 0.50, with a high z-score of 5.60. This positive spatial autocorrelation is statistically significant, which means that when people are expressing change of direction, certain parts of the U.S. are doing so more significantly than others. For the static spatial relationship category, the z-score 6.96 also indicates a strong regional clustering. From this statistical analysis, we
conclude that when expressing change of direction or static spatial relationship, there is a significant regional difference between using relative directions or cardinal directions within the contiguous U.S.

Table 4. Spatial Autocorrelation (Moran’s I) Result in the U.S. Using Four Regional Linguistic Characteristics (proportion): RD: Relative Directions, CD: Cardinal Directions

<table>
<thead>
<tr>
<th>Regional Linguistic Characteristics</th>
<th>Change of direction</th>
<th>Static spatial relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RD/CD</td>
<td>RD/CD</td>
</tr>
<tr>
<td>Moran’s I Value</td>
<td>0.50</td>
<td>0.64</td>
</tr>
<tr>
<td>Z-score</td>
<td>5.60</td>
<td>6.96</td>
</tr>
</tbody>
</table>

To further explore this pattern at the local level, we use *local indicators of spatial association* (LISA, Anselin, 1995) to evaluate which states are significantly different from their neighbors. GeoDA (Anselin et al., 2006) was used to produce the following LISA maps (see Figure 12). We can see that the central parts of the U.S. from MT to NM form a significant cluster of high numbers for cardinal directions, and low numbers of relative directions. Additionally, parts of the New England region show significance clusters of exactly the opposite.

Figure 12. LISA map of *change of direction* (left) and *static spatial relationship* (right). High-High indicates a state with a high value of relative direction usage that is significantly similar to...
its neighbors; Low-Low indicates a state with a low value of relative direction usage that is significantly similar to its neighbors; Low-High and High-Low indicates a state with that is dissimilar to its neighbors (with low/high relative direction usage respectively); all significance levels above are of 0.05.

5. Discussion

This study presented a new methodology for addressing pertinent questions of spatial language use. In this section, we discuss findings and seek explanations for our observations. We also discuss potential refinements in the methods we adopted.

5.1 Interpretation of Regional Patterns

The analyses demonstrate that statistically significant regional differences exist at the state level across the U.S.; eastern states use relative direction more often, and inland states use cardinal direction more often. Some possible contributing factors are the characteristics of roads and highways as they relate to the physical environment (e.g., how much roads curve, the alignment of the street grid to absolute NSEW directions, or the flatness of physical environment), and differences between urban versus rural environments.

Figure 13. Extent of the U.S. Public Land Survey (left, Thompson, 1988) and the township grid partitioning applied (right).
First, the coasts and land that borders an ocean (with their mountainous hinterland) potentially have more curvy roads compared to central parts of the country (with fewer or no mountains). Using cardinal directions on curvy roads, although feasible, may cause confusion. This may be one of the reasons behind the dominance of relative directions in coastal states. The difference between eastern and inland states is also possibly related to a second aspect of the physical environment that has its origin in the United States Public Lands Survey System (Township and Range, see Figure 13). It has been observed that the grid patterns applied in the PLSS have influenced the built environment in the Midwest. The grid pattern may also influence the spatial language usage of cardinal directions as most states with higher-than-median usage of cardinal directions are the states covered by the PLSS (inland and western states shown in Figure 10 and Figure 11.)

Street grids in the flat, inland regions may be more likely to have straight roads that align to NSEW compared to their coastal region counterparts. Under these circumstances, it is reasonable that cardinal directions are more common in the inland regions as an environmentally aligned form of communication (this is supported by the analysis in Section 4.2.2). It is reasonable to assume that when the street grids are curvy or not strictly aligned to NSEW (commonly seen on the east coastal states), people use cardinal directions less often.
Second, the difference between urban and rural environments could be another explanation. Compared to people living in rural areas, those living in urban areas have less access to cardinal direction markers (e.g., the sun) and potentially spend more time inside large buildings that require indoor navigation. Without access to the sun, using cardinal directions to navigate is much more difficult than using relative directions. On the other hand, people who live or work in the rural area have access to open space and the sun, which makes using cardinal directions feasible. Comparing the map of urban areas and urban clusters (Figure 14) with Figure 10 and 11, the use of relative directions seems to correspond with urbanization.

Figure 14. Urbanized areas and urban clusters (United States Census Bureau, 2010).
These hypothesized interpretations are challenging to prove and more research is required. Nevertheless, they provide potential explanations for the distinct differences between eastern states and inland states (see Figure 10 and 11). The proposed explanation agrees with existing studies on the possible influence of physical environment and cultural factors on direction usage (Davies & Pederson, 2001; Ishikawa & Kiyomoto, 2008; Lawton, 2001; Montello & Xiao, 2011).

5.2 Advantages and Limitations of the SARD Corpus and Analysis Scheme

The three step data collection scheme (scrape, classify, georeference) using postal code databases has advantages for collecting spatially distributed data with low spatial ambiguity. Three aspects of our data collection schemes have a potential impact on results and are important to consider in follow up research: regional density, machine-generated versus human-generated route directions, and the comprehensiveness of data collection.

First, the number of route documents per region within the SARD Corpus is proportional to the population of that region (see Figure 3, Figure 9 and Figure 14). Thus, the frequency of route documents in the corpus represents the geographic distribution of such documents well (more populated places have more businesses and other organizations that post directions). This geographic density bias in the corpus toward populated places results in a relatively small sample of spatial language for less-populated parts of the country.

A key question to pose is: What should be favored, representativeness of regions or representativeness of usage? If it was possible to collect the same number of direction usage from each postal code (assuming all postal codes had multiple relevant documents), the postal code regions would be represented equally. However, the sample would be less representative of people and organizations that generate directions within the more highly population regions (as
route directions are provided more frequently in these regions). In comparison to either extreme, our choice provides a relatively balanced way to obtain representation of overall route direction usages.

Second, machine-generated route directions by Natural Language Generation (NLG) systems and by human-generated route directions can both be found on the Web. We are only interested in the latter. The machine-generated route directions have a rigid style, in contrast to human-generated ones. For example, machine-generated route directions typically contain very precise travel distances (e.g., *drive for 10.2 miles*), while such distance measurements are not frequently found in human-generated route directions. In human-generated route directions, vague distances (e.g., *drive for approximately 2 miles*) are used instead of exact distance measurements. Most map service providers, such as MapQuest, Google Maps, Yahoo! Maps, do not apply landmarks in their machine-generated route directions. In contrast, landmarks are used frequently in human-generated route directions (e.g., *after the traffic light you will see a church on your right*). To be credible for use in understanding human spatial language, it is important to ensure that the corpus is built with only human-generated text. To prevent the machine-generated language from causing a major effect on the analysis, the route direction classifier includes machine-generated route directions as a negative training set. Examination of the final SARD Corpus shows that less than 3% of route directions conforming to the style of machine-generated route directions exist in the corpus, which should not affect the validity of the analysis results. From another perspective, it is difficult to prove that human-generated route directions did not used machine-generated directions. Some route directions, although generated by a human (with landmarks), clearly borrow some machine-generated distance measures and combine them together with the describer’s own text. On the other hand, the online map service providers are
improving their technology in generating route directions in a more human-like way as well. Bing Maps,\(^6\) for example, has started to provide landmark information in the machine-generated route directions, in the form of *Turn left onto Colonnade Blvd -- 0.3 miles. OUTBACK STEAKHOUSE on the corner. The last intersection is Waddle Rd. If you reach Theatre Dr., you’ve gone too far.* The last sentence learn from the style of human-generated route directions. It is evident that the spatial databases of online map service providers are evolving together with the NLG systems for route directions.

Another limitation of the SARD Corpus is the lack of certain meta-data. For example, the origin, age, gender, native language of the person who wrote the route direction online cannot be easily collected in a passive web sourcing setup. With an experiment-based setup (such as an Amazon Mechanical Turk HIT\(^7\)), these meta-data could be collected for further exploration of certain language phenomena.

With regards to the analysis scheme, there are several important factors to consider:

- For analyzing the SARD Corpus, token occurrence rather than token frequency (for example, occurrence per million words) is chosen as a linguistic measure due to the variable size and length of web documents. All documents in the corpus are original web pages where the route directions were found. As many share a common webpage design, they may include headers, advertisements, and other body text that are not related to route directions. These unrelated text elements may greatly affect token frequency if calculated directly. Token occurrence is counted by hand examination and categorized in the context of the embedding text, which prevents counting tokens from the unrelated text. In


\(^7\) [http://www.mturk.com](http://www.mturk.com)
envisioning a sentence-level route direction classifier, we think that token frequency could become another linguistic measure.

- The regional linguistic analysis presented here focuses on U.S. states as the region of analysis. However, political boundaries represent a somewhat arbitrary aggregation scheme for analyzing regional linguistic characteristics. While some political divisions do correspond to cultural differences (e.g., where immigrants originates from), others are arbitrary (straight lines of latitude or longitude). The state boundaries in the U.S. sometimes ignore rivers and mountains, which have been found to have significant effect on spatial language usage (Zelinsky, 1955). An alternative to using political regions as the unit of analysis is to focus on cultural or explicitly linguistic regions. Attempts to define linguistic regions can be found in existing research (Zelinsky, 1955). An interesting follow-up to the research presented here would be to carry out the same analysis using previously identified linguistic regions rather than states to determine whether the broad range of linguistic differences that underlie the regionalization matches with regional differences in explicitly spatial language.

5.3 Related Study on the SARD Corpus

The SARD Corpus collected and used in this study has been used to analyze how activity (e.g., whether the route directions that describe a trip to a hospital are for visiting a friend or due to emergent sickness) influences spatial language use in route directions (Hirtle et al., 2011). This study, from another perspective, demonstrates the power of corpus linguistics and the potential wider use of the SARD corpus.
6. Future Work

Although the analysis in this paper focuses on relative versus cardinal directions, several other potential analysis opportunities have been noted during the process of conducting linguistic analysis of route directions.

Landmarks are a set of georeferenceable features in route direction documents and appear frequently in human-generated route directions. The type of landmarks used may vary in different route environments. For example, buildings may be the most effective type of landmarks at the local scale (e.g., in metropolitan environments such as New York City), but they may be less useful when describing regional-level routes (e.g., traveling on interstate highways). Landmark classification has been studied thoroughly (e.g., Hansen et al., 2006). To study landmark usages, route documents can be organized with regard to scale of route. Additionally, landmarks can be divided into the following classes: point-like (traffic lights and buildings), line-like (roads and streets), area-like (parks and campuses) and others. The occurrence and frequency of landmark categories in local-scale versus region-scale route direction document sets can be compared with each to provide an understanding of how humans use landmarks across various environments.

Explicit mentions of travel distance and time are included in both human- and machine-generated directions, but vary in the level of precision as noted above (with human references being typically less precise). In addition, references to distance and time will vary by mode of transportation. For example, the use of travel distance (e.g., drive for approximately 6 miles on U.S. 322) is appropriate for automobile travel, but it is less appropriate than reference to travel time when referring to mass transit (e.g., The travel time is 20-30 minutes longer during peak hours). Analysis of the frequency of references to travel distance and time in each mode of
transportation on a large spatial language corpus will provide a more concrete idea about how humans use these references in different contexts. Additionally, an analysis of regional differences in propensity to reference geographic distance versus travel time in directions would complement previous research that has demonstrated that travel time can form the basis of cognitive distance (MacEachren, 1980).

Acknowledgement

Research for this paper was funded by the National Geospatial-Intelligence Agency/NGA through the NGA University Research Initiative Program/NURI program. The views, opinions, and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the National Geospatial-Intelligence Agency or the U.S. Government.

Reference


Hansen, S., Richter, K.-F., & Klippel, A. (2006). Landmarks in OpenLS - a data structure for cognitive ergonomic route directions. In M. Raubal, H. Miller, A. Frank, & M. Goodchild...


Nittany Lion Inn Website (2012).


United States Census Bureau (2010). Thematic Map of Urbanized Areas and Urban Clusters from 2010 Census.


zipcodeworld.com (2009). Postal codes for the U.S.