The Pennsylvania State University

The Graduate School

Department of Geography

EXPLORING REGIONAL VARIATION IN SPATIAL LANGUAGE:

A CASE STUDY ON SPATIAL ORIENTATION BY USING VOLUNTEERED SPATIAL LANGUAGE DATA

A Thesis in

Geography

by

Sen Xu

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

August 2010

The Thesis of Sen Xu was reviewed and approved* by the following:

Alexander Klippel
Assistant Professor of Geography
Thesis Advisor

Alan MacEachren
Professor of Geography

Karl Zimmerer
Professor of Geography
Head of the Department of Geography

*Signatures are on file in the Graduate School

ABSTRACT

This thesis seeks to answer the question of how spatial language varies regionally within the same language on a geographic scale. Spatial language, such as route directions, is language pertaining to spatial situations and spatial relationships between objects. Spatial language is an important medium through which we study humans' representation, perception, and communication of spatial information. Existing spatial language studies mostly use data collected via time-consuming experiments, which are therefore limited to a small sample size—thus limiting the detection of how spatial language varies from one region to another. More recently, larger sample sizes have become possible due to the abundance of volunteered spatial language data on the World Wide Web (WWW), such as directions on hotels' websites. This data is a potential source for scaling up the analysis of spatial language data. Sourcing from the WWW, a spatial language data collection scheme has been developed. Automated web-crawling, spatial language text document classification based on computational linguistic methods, and geo-referencing of text documents are used to build a spatially-stratified corpus. Focusing on route directions on the WWW, the Spatially-strAtified Route Direction Corpus (the SARD Corpus) with more than 10,000 spatially distributed documents covering three countries (the United States, the United Kingdom, and Australia) is built. As a case study on the SARD corpus, a linguistic analysis scheme assisted by computational linguistic tools is designed based on the often raised question of cardinal versus relative direction term usage. Semantic usages of cardinal and relative

directions are identified as regional linguistic characteristics; a visual analytic toolkit is used to detect regional variations in the SARD corpus. Analysis results and possible indications of linguistic variations at the national and regional scale are presented and discussed, contributing to research on spatial language use. The analysis shows similarities and differences in directional term usages on the national level; regional level analysis shows that geographic patterns emerge on the linguistic term usage. The findings offer knowledge contributions to the field of spatial cognition; the design and implementation of building a geo-referenced large-scale corpus from documents crawled from the WWW offers a methodological contribution to corpus linguistics, spatial cognition, and the GISciences.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I am sincerely thankful to my advisor, Alexander Klippel, for his intellectual, emotional and moral support throughout the last two years. I am very grateful to work with Alan MacEachren, who provided intellectual guidance and encouragement for my research. This thesis would not have been possible without their help. I would also like to thank the GeoCAM project team. Ian Turton has given valuable help to me with regard to software development. Presenjit Mitra has provided many pieces of innovative ideas to solve technological challenges in the project. I am thankful to my colleagues, Scott Pezanowski, Anuj Jaiswal and Xiao Zhang, with whom I have spent hours discussing ideas about my research. Lastly, I would like to thank my family and friends for their endless love and support.

Terminology

Scheme: specifies a template, a framework for organizing and analyzing information. *Data collection scheme*, as in this thesis, refers to a detailed step-by-step workflow for carrying out data collection. *Data analysis scheme* refers to analysis procedures for interpreting the SARD Corpus.

Corpus: a large-scale, systematically-organized language text collection. In this thesis, the Spatially-strAtified Route Direction Corpus (SARD Corpus) is a spatial language corpus collected from the World Wide Web (WWW), focusing on route directions written in English from the U.S., the U.K., and Australia, organized by states (U.S.) or postal districts (U.K. and Australia).

Spatial language: "the encoding of objects, their motions, locations, and properties" [Landau, 1998, p. 14]. Another definition is "the terms in human language that people use to refer to spatial situations" [Mark and Frank, 1989, p. 540]. This study focuses on spatial language documents from the WWW, in which route directions take up a major proportion.

Postal code: a series of letters and digits in fixed format, used in postal addresses assigning to geographical areas.

Postal region: a geographical area which can be defined by a set of postal codes. In this study, the postal regions in the U.S. refer to U.S. states (such as “PA”, “NY”); the postal regions in the Australia refer to states and territories in Australia (such as “ACT”, “NSW”); the postal regions in the U.K. refer to U.K regions (such as “East Anglia”, “Midlands”). A detailed list of postal regions in the U.K. can be found in Appendix B.

Regional linguistic characteristics: derived from the spatial language of a region, the characteristics of the regional language corpus. In this study, the proportion of relative directions representing “change of direction” in Pennsylvania is an example of a regional linguistic characteristic.

1 Introduction

Spatial cognition, defined as “the study of knowledge and beliefs about spatial properties of objects and events in the world” [Montello, 2001, p.14771], is an important research topic in geography as it enriches understanding of how humans perceive, represent, and communicate spatial information. Analyzing linguistic phenomena has been a crucial approach for advancing spatial cognitive research [Bateman et al., 2007, Allen, 1997]. For example, using language as a window into spatial cognition, individual and group similarities and differences in using spatial language have been studied [Montello et al., 1999, Ward et al., 1986, Ishikawa and Kiyomoto, 2008]. With respect to spatial language differences, it has long been noted that 1) different languages can refer to the same spatial information differently [Munnich et al., 2001, Lamarre, 2008, Burenhult and Levinson, 2008]; and 2) these differences can occur even with the same language in different regions [Davies and Pederson, 2001, Ishikawa and Kiyomoto, 2008, LIU, 2008, Zelinsky, 1955]. This thesis focuses on an extensive analysis of the latter question—is there regional variation within the same language regarding the use of spatial language?

Exploring regional differences in spatial language requires extensive spatial language datasets on a geographic scale, that is, a scale that “cannot be apprehended directly through locomotion; rather, it must be learned via symbolic representations such as maps or models” [Montello, 1993, p.315]. A continent, or countries across the

globe, would be examples of a geographic scale. This poses challenges for both data collection and analysis. Given the limitations of the small data sample sizes that result from existing experiment-based methods (refer to section 1.1.2), new data sources are needed for detecting regional linguistic variations. One potential data source for this study is the World Wide Web (WWW). The WWW is already large and is also rapidly growing due to its popularity in information sharing and for its easy accessibility. The WWW has become a popular platform for sharing information about how to get from a location A to another location B—that is, *route directions*. Route directions are a culturally universal phenomenon [Allen, 1997], which represents how humans perceive and communicate spatial knowledge. Therefore route directions are an excellent candidate for exploring the possibilities of the WWW as a data source for spatial cognitive-linguistic studies.

Inspired by information retrieval methodologies, this thesis presents a web sampling method to collect large-scale geo-referenced spatial language documents from the WWW with a focus on route directions. The Spatially-strAtified Route Directions Corpus (SARD Corpus) is built using this data collection method. Focusing on an often-raised question with regard to the usage of cardinal (for example, north, south) versus relative (for example, left, right) directions (defined in more details in Section 3.2, Table 2), a semantic analysis scheme with assisting text processing tools and visual analytics tools was designed to carry out linguistic analysis on the SARD Corpus.

The remainder of this chapter will detail the motivation for this research, including the value of studying spatial language, the limitations of existing

methodologies, the inspiration from computational studies and the applied value of this study (Section 1.1). Section 1.2 formulates the research questions on the broad level as well as on the focused level. The approaches that are adopted from various disciplines are highlighted in Section 1.3. The scope of this study is discussed in Section 1.4. Section 1.5 provides a roadmap to the remainder of the thesis.

1.1 Motivation

The motivation of this study comes not only from curiosity in increasing our knowledge on regional language differences and thereby adding to research in spatial cognition, but also from research developments in related fields. This section illustrates the driving force behind this study from five perspectives:

First, spatial cognitive differences may come from various aspects, for example gender or wayfinding ability [Vanetti and Allen, 1988]. From a geographic perspective, the specific region that is being described in route directions is also a potential factor that relates to spatial cognitive differences. Investigating the regional variations in spatial language would enrich the spatial cognition knowledge base.

Second, language has been used together with other behavioral experiments to provide insights into spatial cognitive questions. However, there are certain limitations of experiment-based methodologies. Although experiments allow high control over the data collected, most of the data is not being reused or extended for other research purposes. The large scale data demands posed by the goal of studying regional linguistic

variation asks for an automated data collection method, which would benefit other spatial language studies as well. Therefore, a methodology specifically focusing on collecting, processing, and analyzing spatial language text on a geographic scale is called for.

Third, information retrieval techniques and text processing tools have been matured for collecting and analyzing large quantities of text data. Automatic retrieving and processing information from large sets of text documents becomes possible with technological advancements. There also have been algorithms and tools developed for information retrieval (for example, web crawler) and text processing (for example, corpus linguistic toolkits), which would be helpful for the data collection and analysis in this thesis.

Fourth, the WWW has become a hotbed for volunteered spatial data [Goodchild, 2007], where crowdsourcing has demonstrated to be effective in various research fields [McConchie, 2002]. Spatial language data available from the WWW has great potential for spatial cognitive studies because of its unrivaled coverage and easy accessibility. The value and feasibility of retrieving and organizing volunteered spatial language data should be considered for spatial cognition research.

Fifth, from an applied perspective, the result from regional spatial language variation may provide improvement possibilities for digital map service providers and GPS clients for a better localized route directions output—what region the route description is talking about should affect how the spatial language is formed.

1.1.1 The Reasons of Studying Spatial Language

Spatial language, defined as “the encoding of objects, their motions, locations, and properties” [Landau, 1998, p.14], is considered a window into human cognition. Spatial language research has addressed topics such as spatial reference systems and geographic concepts to reveal knowledge about space [Ward et al., 1986]. Through language, complex mental conceptualizations of spatial movement patterns become accessible to scientific research. Although there are different models available to explain the relationship between spatial language and spatial cognition [Landau, 1998, Jackendoff and Landau, 1991], the validity of studying spatial cognition through the representation of linguistic descriptions of spatial information is generally accepted. Compared to other spatial communicative media, such as sketch maps, the abundance and popularity of daily spatial language usage offers a more generous data source that is appealing to the goal of exploring regional variation on a geographic scale.

[Montello, 2001] pointed out that studying spatial cognition offers an improved understanding of many cognitive phenomena of interest, such as the idea that “where people choose to shop should depend in part on their beliefs about distances and road connections” [Montello, 2001, p.14772]. Similarly the value of studying regional variations in the usage of spatial language offers understanding of linguistic phenomena of interest such as language preference on reference frames (preference of using relative directions such as *left* and *right* over cardinal directions such as *east* and *north*, see examples of cardinal and relative reference frames below). Spatial language is an important form of how people communicate spatial information. For example, consider

the following two sentences that describe identical spatial movements: “go ahead on Atherton Street, turn *left* at the traffic light” and “go *north* on Atherton Street, turn *west* at the traffic light”—one may be preferred by different groups of people on the reference frame preference, [see Ishikawa and Kiyomoto, 2008]. Studying the regional variation in the usage of cardinal and relative direction terms provides insights into why people perform wayfinding tasks or give route directions differently using different reference frames.

Route directions, as a type of spatial language, have been studied among the spatial cognition community extensively for their close relationship to wayfinding problems. Despite the widespread use of maps (and the rapidly growing use of digital maps), route directions are still a popular and effective means that people use to communicate spatial information in a daily manner. Analyzing route directions regarding structure and linguistic characteristic helps to provide insight into how spatial knowledge is represented and communicated through language systems and how language systems affect such representation and communication. The same route or space, for example, can be conceptualized differently by different people (for various reasons, language difference being an important one). The analysis of route directions helps to reveal the differences in an environment’s representation and allows for the comparison of two different route directions for the same route with respect to their effectiveness in conveying spatial information.

As one of the most identifiable kinds of spatial language, route directions are abundant on the WWW—under headlines such as “Directions and Transportation” on

websites for hotels or churches, “Route Instructions” for trails, and so forth. Previous researchers have used route directions to gain insight into wayfinding performance and differences in spatial cognition [Ishikawa and Kiyomoto, 2008, Allen, 1997, Daniel and Denis, 1998, Lawton, 2001, Lovelace et al., 1999, MacMahon et al., 2006, Vanetti and Allen, 1988, Lewis, 1976]. This thesis focuses on using route directions on the WWW as a representation of people’s everyday usage of spatial language to explore regional linguistic variations.

Research in spatial cognition has posed questions that can be answered from different perspectives. For example, the difference in spatial cognition may be caused by various factors: age, gender, activity preference, culture, language, mode-of-transportation, scale of route, street grid layout, or physical environment of the route. This variety of research perspectives attracts researchers including psychologists, architects and planners, linguists, philosophers, and computer scientists to approach questions such as why and how spatial cognitive differences exist [Montello et al., 1999] and what is the effect of spatial language preference on wayfinding [Ishikawa and Kiyomoto, 2008] from various perspectives. Existing research in spatial cognition mostly follows psychological approaches with highly controlled human participant experiments. From a geographic point of view, geo-referencing and analyzing spatial language usage extracted from the WWW offers a lens to look at spatial language variation on a broader scale. This topic is the focus of this thesis.

1.1.2 Limitation of Existing Research Methods

The experiment-based methodology for research on spatial language provides a first-hand understanding of aspects of spatial cognition [Montello et al., 1999, Davies and Pederson, 2001, Ishikawa and Kiyomoto, 2008, Ward et al., 1986]. Experiments offer researchers high control over the form and focus of the data collected, which provides the experimenter with more background information (such as gender, culture groups, educational background of the participants) that allows for better interpretations of the observed language phenomena. Although the results from the above studies are insightful, we note that this methodology has the following limitations when it comes to exploring regional variation.

First, the high cost of experiment-based data collection limits such studies to small sample size data, which is usually sufficient for exploratory purposes. And, the analysis in experiment-based spatial language research is often carried out by examining data word-by-word. While with proper participant selection and sampling methods, the results from such studies can be representative to an extent and provide insight into exploratory research questions. However, the small sample size limits experiment-based studies to offering statistical evidence on questions such as whether people prefer to use relative directions or cardinal directions when expressing *change of directions*. A large size spatial language corpus is needed to answer such questions. The size and scale of the target corpus in this study far exceeds the capacity of manual interpretation. Computational text processing toolkits are required to reveal linguistic patterns over large quantities of text documents.

Second, the sample size limit is also a bottleneck for expanding the data to a geographic scale. To conduct a spatial language study on regional linguistic patterns, extensive spatial coverage of linguistic data is required. To meet the demand for an extensive language study over a geographical region (for example, a continent), the high cost and time required make the experiment-based data collection method unaffordable.

In summary, it has been demonstrated to be effective to use experiment-based data collection methods to improve the understanding of small-scale phenomena through manual interpretation by analysts. However, exploring regional spatial language variation requires a methodology in which spatial language is collected and analyzed on a geographic scale. Scaling up the dataset to a language corpus that is spatially distributed is challenging and calls for new data sources and collection methods. Analysis of large quantities of linguistic data to reveal regional patterns also requires text processing and visual analytics toolkits to assist the interpretation.

1.1.3 Inspiration from Computational Approaches

With the development of computational linguistics and geographic information retrieval, the WWW is quickly becoming a resource that allows for systematically collecting a large corpus of text containing geo-referenceable route directions. Research from the fields of information retrieval and computational linguistics provide inspirations to solve the data collection challenges in this thesis. Focused crawling, defined as “seeks by seeds and links, acquire, index, and maintain pages on a specific set of topics that represent a relatively narrow segment of the web” [Chakrabarti et al.,

1999, p.1625], has been demonstrated to be effective in retrieving topic-specific documents with high precision. Inspired by the web crawling techniques, a data collection scheme aiming at retrieving spatial language documents from the WWW is developed. [Zhang et al., 2009] demonstrate the effectiveness of machine learning methods for route-related document classification and sentence classification. The crawling scheme introduced in this thesis is modified from Zhang et al., 2009's work by using postal codes list instead of keywords-of-interests (refer to Section 3.1.1).

On the data analysis aspect of this thesis, [Hearst, 1999] puts forth the hypothesis that a data mining approach could be effective for the linguistic analysis of very large sets of text documents. [Martins et al., 2006] offers methods to deal with the location information of documents from the WWW. The WWW, because of its free and abundant data, has already attracted researchers interested in, for example, linguistics [Beesley, 1988, Banko and Brill, 2001] or information retrieval research [Dumais et al., 2002]. The very high number of route direction documents on the WWW makes it an ideal candidate as the data source in this study.

To sum up, the study of spatial language can benefit from various disciplines—from a computational perspective, new data sources, collection and analysis methods are available to approach spatial language on a geographic scale.

1.1.4 The Power of Crowdsourcing

Crowdsourcing, a term coined by Jeff Howe [Howe, 2006], is used to describe outsourcing to a large group of people or community (often referred to as “a crowd”). For example, in the area of stock photography, individual works of photographers were gathered together to form a useful photography database [Howe, 2008]. Nowadays, crowdsourcing has expanded to a much wider range of topics including commercial projects [List_of_crowdsourcing_projects,]2010]. The notion of crowdsourcing has also expanded from simply gathering individually created work together to outsourcing contributions from a crowd to accomplish a task performed by an employee or contractor. Despite its success in business applications, crowdsourcing has also been demonstrated to be effective in various scientific research topics such as language translation [Ambati et al., 2010], usability studies [Heer and Bostock, 2010], and GIS [Hudson-Smith et al., 2009a, Goodchild and Glennon, 2010]. A related crowdsourcing success of studying regional difference is the famous Pop versus Soda Map [McConchie, 2002], where McConchie sets up a website taking inputs from people visiting the website about their preference on using the term “Pop” or “Soda” to refer to soft drinks. To georeference people’s preference, he also collected location information in form of postal codes. This volunteered data collection method is demonstrated to be effective as the regional preferences clearly stand out in their mapping results.

The demonstrated effectiveness of crowdsourcing offers a new inspiration for collecting spatial language text. Spatial language data has never been so accessible thanks to the rapid development of the WWW. For example, hotels may put route

directions on their websites for the convenience of their customer. This type of volunteered data gives the WWW great potential as a data source for various research purposes—among which spatial cognition takes an important place. The WWW could now be viewed as an experiment platform with the widest range of participants, who are uploading enormous amounts of spatial information (together with other non-spatial information) onto the platform every second. The popularity and abundance of texts on-line that communicate spatial knowledge ensure the abundance of sample data for analysis from this data source. Harnessing the volunteered route directions on-line is one major focus of this thesis.

1.1.5 Application Oriented Thoughts

One applied value from spatial cognition studies is to "improve the use and design of maps and other geographic information products" [Montello, 2001, p.14772]. Digital map service providers and GPS clients are becoming more and more personalized—you can even set what voice you want to hear when the satellite navigation system gives you instructions. However, no system has been developed that allows a spatial language preference setting (for example, one want to use cardinal directions when the instruction is about "change of directions"). The result of this study goes further for a better localized route directions output—the route directions generated should use a spatial language style that fits the regional characteristics of the route.

To sum up this section, studying spatial language is a valid and effective approach for gaining understanding in spatial cognition. The limitation of existing experiment-based methods, especially in data sample size, restrains the exploration of regional variation on a geographic scale. With computational methods such as information retrieval and crowdsourcing techniques, the WWW has become a potential data source to study regional variation in spatial language.

1.2 Research Question

1.2.1 The Broad Research Question

In this thesis, I demonstrate the applicability of advances in computational linguistics to address fundamental questions in spatial cognition research: how language differs regionally and how the tremendous resource of volunteered linguistic information can be used in a spatial analysis approach.

1.2.2 The Focused Research Question

As a case study using the SARD Corpus, tools, various self-developed computational research tools, and a sophisticated analysis scheme, the research question focuses on the usage of cardinal versus relative directions in route directions. Regarding their semantics, they can both be used to indicate actions as well as static spatial relations. But, is there any regional difference in the usage of cardinal versus relative direction terms in spatial language?

1.3 Approach

For the data collection, a spatially-stratified web sampling scheme is developed that produced the Spatially-strAtified Route Direction Corpus (the SARD Corpus). We seek innovation from crowdsourcing and computational linguistic techniques to develop a novel methodology to achieve the goal of building a geo-referenced, geographical-scale, spatial language corpus. A machine learning-based document classification approach is applied to crawl and identify route direction documents from the WWW. Postal code data is used to coarsely geo-reference each route direction document in the SARD Corpus across the three countries within our research range: the U.S., the U.K. and Australia. These three countries are chosen for their widespread internet usage and accessibility, and because they all use English as the dominant language. Corresponding regional linguistic analysis schema are developed focusing on cardinal versus relative direction term usage. Following the computational analysis, visual analytics tools are used for a more precise analysis of linguistic usage patterns. Figure 1 illustrates the framework of this methodology (details of the framework are provided in Chapter 3).

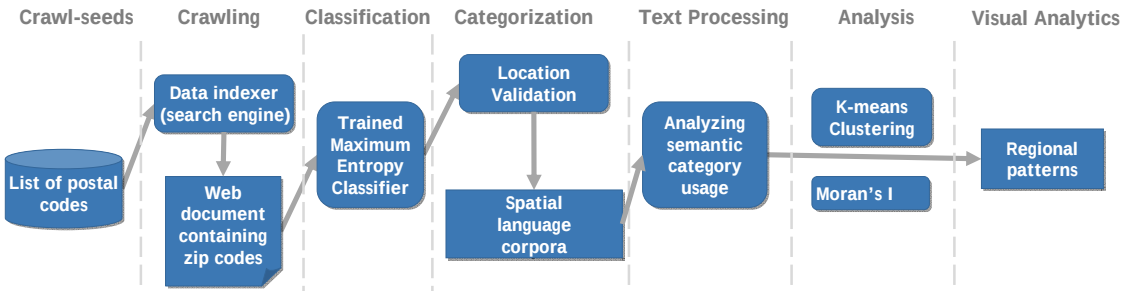

Figure 1. Overview of the methodology for building and analyzing the SARD Corpus

1.4 Scope

This study mainly focuses on three aspects. To start with, developing a data collection scheme for large quantities of spatial language (that is, route directions) is required to detect regional variation. To conduct linguistic analysis on the harnessed texts from the WWW, the methodology also covers the process of identifying semantic categories and an analysis scheme for spatial analysis of linguistic data. The geo-referencing of each document and the scale of geo-referencing is also a key point in this thesis. Obtaining world-wide coverage of documents is one of the major advantages this study has compared to existing studies. The target corpus is much larger than the ones in behavioral research (over 10,000 spatial language documents have been collected and processed). This study focuses on three English-speaking countries (the U.S., the U.K., and Australia) to detect the regional linguistic differences in route directions in English.

For studying spatial languages on a geographic scale, this thesis focuses on text data from the WWW. Admittedly, certain differences exist between volunteered spatial language text retrieved from the WWW and data collected from controlled experiments. Nevertheless, spatial language data from the WWW is a valuable representation of how people currently communicate spatial information via written language, and one which has not drawn enough research attention. The advantage in accessibility, spatial coverage, and abundance drives this study forward. This data source and collection scheme offers a different lens for looking at spatial language on a geographic scale, which is an enhancement for research in spatial cognition.

The usage of cardinal and relative directions is selected to be analyzed as an initial analysis using the SARD Corpus (refer to detailed analysis scheme in Section 3.2). The terms “left”, “right”, “north”, “south”, “east” and “west” are easy to identify in route directions which makes the analysis feasible. Their usages, such as “from the south” or “fork left” have multiple variations and could be used to identify regional preferences. However, cardinal and relative directional terms are not comprehensive to reflect all aspects of linguistic characteristics in route directions; cardinal and relative directions are selected representations of reference frames in human spatial cognition that is the focus of this thesis. This analysis serves as a case study to demonstrate the analysis approaches developed in this thesis applied to the SARD Corpus. It is also an extension to spatial linguistic studies from a larger geographical perspective.

From a computational perspective, advances in geo-referencing technologies [Li et al., 2003, Chrisment et al., 2004] draw an intriguing blueprint for spatial language research: the spatial strata of the SARD Corpus is chosen to be based on postal codes and organized by states (or postal code districts). Because the scale of each spatial language document is quite coarse (considering most route direction covers several postal codes), fine geo-referencing on the document level needs to be accumulated to a coarser scale. Finer geo-referencing for spatial language on a sentence level or phrase level could offer insights into a finer scale. However, those aspects are beyond the scope of this study.

1.5 Thesis Structure Overview

The rest of this thesis is organized as follows: Section 2 provides background information on existing topics in spatial language studies, computer linguistics, and solutions to the challenges raised by the research question. Section 3 provides a step-by-step workflow for building a spatially-stratified spatial language corpus from documents crawled from the WWW, a case study on cardinal versus relative direction usage and a corresponding semantic analysis scheme with tool descriptions. Section 4 presents results from the data collection and the case study; the regional differences of spatial orientation are analyzed and interpreted. Chapter 5 discusses the limitations, possible extensions and implications of the methodology. Section 6 summarizes this thesis and lays out opportunities for future work.

2 Background and Literature Review

In this chapter, findings from existing research regarding spatial language, corpus linguistics and useful computational methods are highlighted. The target route direction documents from the WWW is analyzed through examples and illustrated in greater detail to provide inspiration for designing the data collection and analysis scheme. The challenges in both data requirements and data analysis requirements are listed with potential solutions.

2.1 Spatial Cognitive Research Using Language

Spatial cognitive research using language as a medium has been carried out since the 1970s [Klein and Levelt, 1978]. Obtaining insight into the acquisition and communication of spatial knowledge by studying language that is spatially related has stimulated researchers to work across disciplines [Jackendoff, 1983, Jackendoff and Landau, 1991, Retz-Schmidt, 1987, Herskovits, 1986]. There are several important fields discussed in this section in the topic of spatial cognitive research focused on language: the generation and structure of route directions, the individual and group similarities and differences of spatial language usages in route directions, the regional differences of spatial language usages in route directions and a frequently raised question on the usage of cardinal directions versus relative directions.

2.1.1 Generation and Structure of Route Directions

Spatial language has been used for "relating various models of geographic space and concepts of fundamental spatial relations" [Mark and Frank, 1989, p.540]. Concepts of space (for example, spatial relationships and reference frames from small-scale space to large-scale space) and their relationship to spatial language are important topics not only for cognitive science, but also for the future of geographic information systems and analysis. In Mark and Frank's study, they aim at designing a model of spatial relations and properties by analyzing language towards "a geometry of language" [Mark and Frank, 1989, p.550]. They approach through studying properties of spatial reference in natural language to gain understanding of spatial concepts.

[Allen, 1997] specifically looks into the generation and interpretation of route directions. The most important part of communicating route directions is route description, which is defined as "a set of communicative statements that provides sufficient information to the questioner to reach the designated destination."[Allen, 1997, p. 365]. Route description is where spatial information is mostly conveyed. Allen breaks route description down to two types of communicative statements: *directives* (verb phrases or sentences that are "reducible categorically to go or turn" [Allen, 1997, p. 365], for example, "turn left at the traffic light") and *descriptives* (verbs phrases or sentences that are "reducible categorically to some form of to be" [Allen, 1997, p. 365], for example, "there is a church at the intersection"). This categorization serves as basis for developing the analysis scheme for the case study in this thesis (refer to Table 2, "change of direction" and "static spatial relationship"). [Allen, 1997] raises the

important questions of the ecology of wayfinding: how do people normally and naturally use spatial language to convey spatial knowledge? Is there any difference among individuals or among groups? These questions lead to the next section.

2.1.2 Individual and Group Similarities and Differences of Spatial Language Usage in Route Directions

Among spatial cognitive research studies, individual [Vanetti and Allen, 1988] and group differences and similarities, for instance, sex-related spatial abilities [Ward et al., 1986, Montello et al., 1999] are established topics that have received widespread attentions. [Vanetti and Allen, 1988] develop a psychometric test of spatial and verbal abilities and examine the relationship between participants' ability to produce and following directions and their test score. The results show that there is interplay between verbal and spatial abilities: the group with a high spatial ability test score produces more accurate and efficient directions than the low score group; the group with high verbal score also outperforms the low score group in way-finding tasks. [Ishikawa and Kiyomoto, 2008] investigated the effect of using different reference frames in route direction on wayfinding performances. The result of their study showed that the preference of Japanese speakers is to use relative directions; using cardinal directions in route directions can render a negative effective in wayfinding performances.

In [Montello et al., 1999], experiments including psychometric tests, map-learning tests, and tests of object-location memory were carried out for detecting sex-

related differences and similarities in spatial abilities. Among the series of test conducted, verbal spatial description is one test that compares sex-related difference or similarities regarding spatial language usage. A noteworthy result is that male participants have a higher tendency of using “Metric-distance terms” and “Cardinal-direction terms” than female participants. This finding is supported by several studies [Ward et al., 1986, Dabbs et al., 1998, Lawton, 2001].

The difference among cultures or linguistic groups has also been noticed with regard to the usage of spatial language [Mark and Gould, 1992, Mark and Egenhofer, 1995, Mark and Frank, 1989]. For example, the different language systems of English and Spanish can result in similar descriptions for spatial relationships, suggesting the natural language system for the two languages “might be cross-linguistic robust” [Mark and Egenhofer, 1995, p.255]. Some Hawaiians and other island dwellers have been noted with a radial coordinate reference frame choice, different from the cardinal or relative reference frame. This reference frame preference also resulted in spatial language usage difference [Mark and Frank, 1989, Haugen, 1957].

The source of the differences of spatial language usage in route directions among individuals and among groups has been discussed and several explanations have been put forward. As mentioned above, the difference in spatial language usage among individuals of the same culture and language background might come from individual spatial or verbal ability differences [Vanetti and Allen, 1988]. Reference frame preference for cardinal directions or a preference for using metric information can come

either from sex-related [Montello et al., 1999] or culture causes [Mark and Frank, 1989, Haugen, 1957]. These preferences may also have an impact on spatial language usage.

2.1.3 Regional Differences of Spatial Language Usage in Route Directions

Regional spatial cognitive variation is another related research question raised from similar concerns as individual or group differences. The difference in culture, population, dialect, urban environment, landscape in different regions may yield spatial cognitive variations, even among people speaking the same language. [Davies and Pederson, 2001] investigated the similarity and difference in spatial abilities in a variety of identical experiments in two different cities: Milton Keynes in the U.K. and Eugene in the U.S. to compare the effect of grid pattern in the city. Both Milton Keynes and Eugene have grid pattern urban environment, but residents from Milton Keynes are expected to use wayfinding strategies “more appropriate to other UK cities” [Davies and Pederson, 2001, p.401], where grid patterns are very rare. This comparison can be considered as a small sampled regional comparison, as the goal is to compare how people from different regions, because of the culture and landscape differences, perform spatial tasks differently. The results, particularly on spatial language usage (experiment of giving driving directions), showed that subjects in Milton Keynes use landmarks more frequently (61% higher) than the subjects from Eugene. The variations found in spatial language usage from this research calls for further investigations on regional level analysis. [Lawton, 2001] also suggested that the street grid of the environment may affect the way people choose referencing frames in giving route directions.

The *Phonological Atlas of North America* [Labov et al., 2006] is another example of regional variation in language use. The atlas specifically demonstrates the phonological variation at a geographic scale. For example, the "o/oh merger map" demonstrates that people from the west pronounces *cot* and *caught* similarly or the same while people from the east pronounce the two words distinctively different. The atlas is a useful approach to study dialect diversity in sociolinguistic research. [Zelinsky, 1955] studied the regional variations in spatial terms used in place names, such as *fork* and *run* used in stream names, -ville and –burg used in town or city names and so forth. In Zelinsky's study, mapping the location of place names with the same spatial term reveals regional patterns, which is found to relate to cultural and linguistic characteristics of a region. Inspired by the idea of the *Phonological Atlas* and Zelinsky's study, this thesis seeks to investigate if mapping regional linguistic characteristics in route directions can reveal regional variations in spatial language usages.

2.1.4 A Frequently Raised Question: Cardinal Directions vs. Relative Directions

This thesis focuses specifically on the referencing frame choice question: when giving route directions, is there a regional preference in using relative directions or cardinal directions? The special emphasis on this topic has been a topic of interest in the spatial cognitive research community [Gladwin, 1970, Lewis, 1972, Gumperz and Levinson, 1996, Lewis, 1976, Davies and Pederson, 2001, Ishikawa and Kiyomoto, 2008, Richter, 2005, Richter et al., 2008, Richter, 2008]. In these studies, the components of route directions were broken down and analyzed, which shed light on the semantic functionalities of the linguistic components in route directions. These

studies not only provide theoretical background to the analysis developed in this thesis, but link the results of route direction studies and other spatial cognitive activities such as wayfinding.

2.1.5 Summary of Spatial Cognitive Research

From existing studies in spatial language, differences among different cultural and linguistic groups have been noted. Sex-related, individual spatial or verbal ability-related differences have been investigated, producing interesting results. Studying the difference between cardinal directions versus relative directions usages is a commonly used approach. With the insights from existing studies, this thesis investigates the regional aspect of this frequently raise question on referencing frame preference using route directions. Text processing techniques borrowed from corpus linguistics and computational linguistic researches are highlighted in the next sections.

2.2 Computational Linguistic Research

The developments in computational linguistic research, including information retrieval, text classification and corpus linguistic text processing made collecting route direction text on a geographic scale possible. The approaches in those fields are the fundamental techniques for the research in this thesis.

2.2.1 Web-based Information Retrieval

The field of information retrieval emerged in the 1950s [Luhn, 1957] for finding relevant information from a large collection of data. With the expansion of the WWW, scientific research and has a lot of applications uses web-based information retrieval for data collection. There are several distinctive characteristics of the WWW [Goker and Davies, 2009, p. 86] that made it especially appealing as well as challenging. First, documents on the WWW are referred to as “Web Pages”, each of which is identified with a unique Uniform Resource Locator (URL). The URL can be used as a marker to prevent redundant data collection (refer to Section 3.1.1.2.) Second, web documents have a remarkable diversity. There is a large amount of web documents of various types (news, advertisements, blogs, .etc) in various formats (text, pictures, videos, .etc) that exist on the WWW. This thesis focuses on route directions texts, which need to be extracted from the plethora of documents on the WWW. Third, search engines provide indexing of the massive quantity of data. This offers a very convenient way to retrieve web documents containing keywords of interest. Related research includes using web-based information retrieval for question answering and decision supporting systems [Yao and Yao, 2003, Dumais et al., 2002].

2.2.2 Development in Text Classification and Text Analytics

Text classification aims at identifying (electronic) text documents to one or more categories based on the content [Sebastiani, 2002]. Given the variety of text documents and the fact that research often focuses on one specific topic, text classification has

attracted increasing research interests and has been demonstrated effective in many different fields, for example, book recommendation [Mooney and Roy, 2000], view classification [Turney, 2002]. Using text classification to separate route direction documents out from the various types of documents on the WWW provides a feasible solution to the data collection challenges in this thesis.

Text classification can be divided into supervised document classification (where expert knowledge, such as rule sets or training documents is applied to build the classifier) and unsupervised document classification (where no expert knowledge is introduced). Machine-learning techniques have been demonstrated to be effective in text classification and information extraction [Lee and Lee, 2007]. [Zhang et al., 2009] applied supervised document classification using a rule set (frequently appeared language patterns, such as “turn left at”, “drive 2 miles”, .etc) and training document set (samples of hand-selected route direction documents). Zhang et al. evaluated several text classification models (naive Bayes model, decision tree model, and maximum entropy model) for a route direction document classifier. Among the several models evaluated, the highest precision is the maximum entropy model-based classifier with over 97%. In Section 3.1.2.2, details of [Zhang et al., 2009]’s route directions classifier and its role in the data collection scheme used in this thesis is presented.

2.2.3 Corpus Linguistic Research

A corpus is “a collection of texts when considered as an object of language or literary study” [Kilgarriff and Grefenstette, 2003, p.334]. Although corpora were

originally studied by hand, due to the large corpus size nowadays and thanks to computational linguistic techniques, modern corpus linguistics uses automatic annotation toolkits for the analysis. The methods of corpus linguistics include *annotation* (such as part-of-speech tagging, parsing), *abstraction* (such as rule extraction from a particular corpus), and *analysis* (such as statistical evaluations). Toolkits have been developed and demonstrated to be effective in assisting experts in analyzing corpora [Anthony, 2006]. The goal of corpus linguistics is to study 1) how words are related; 2) how words are used with each other; 3) how common they are in a given domain [Kilgarriff and Grefenstette, 2003]. The semantic usage of cardinal and relative directional words can be studied using methods in corpus linguistics.

Corpus linguistics is a large discipline in linguistic research, the methodology of which has not been explicitly or widely applied to spatial language studies. Existing linguistic studies are either based on existing corpus (such as the Brown, the British National Corpus, or the Penn Treebank) or based on construction of a specialized target corpus with pre-defined resources [Biber, 2006]. A spatial-associated corpus building schema is yet to be developed. In the mean time, the rapid expansion of WWW creates a lot of opportunities for corpus linguistic studies. The route direction corpus, as our target corpus, can be considered representative for language usages in a written, on-line context from users all over the globe.

From a data analysis perspective, data mining approaches have been introduced to linguistic analysis on massive texts (Hearst, 1999). These computational advances

provide a promising strategy to overcome the challenges we face in data analysis over a large quantity of text data to support extracting regional linguistic characteristics.

2.2.4 Emerging of Crowdsourcing in Spatial Cognitive Research

Crowdsourcing [Howe, 2008] in GIScience results in the phenomenon of "volunteered geographic information" [Goodchild, 2007], which opens a new era of data collection in GISciences and spatial cognitive research. Using Web 2.0 applications (such as facebook, youtube, vimeo, myspace and .etc), crowdsourcing with survey data has become inexpensive and easy to carry out. [McConchie, 2002] applied a webpage-based survey to collect survey results (answers to the question of which term web site visitors use to refer to soft drinks) and location (postal codes). Crowdsourcing volunteered geographic information also opens new possibilities such as real-time geographic surveys. Similarly, [Hudson-Smith et al., 2009b] conducted an online survey on "the most hurting factor about the credit crunch" in the U.K. and produced a near real-time mapping of the survey result for investigating regional variation: "48.8% of responses saying that fuel was most significant factor … more respondents within Greater London saying it was either mortgage or rent, or food" [Hudson-Smith et al., 2009b, p. 8].

Inspired by these creative crowdsourcing studies, harnessing route directions from the WWW is a potential research direction for spatial language studies. Rather than having active users conducting surveys, the route directions can be directly harnessed while location information is inherited in the text.

2.2.5 Summary of Computational Research

Web-based information retrieval enables researchers to harness the large quantity of text data from the WWW. Crowdsourcing, a recently emerging research topic, has working on utilizing the huge amount of volunteered data on-line to answer research or industrial questions. Compared to the conventional human-participant involved data collection method, crowdsourcing from the WWW is also inexpensive, fast, and most importantly, has easy accessibility of the target route direction documents from all over the globe. Although documents from the WWW have high diversity, documents of interest (in this thesis, route directions) can be extracted using text classification. Corpus linguistic provides text processing tools for handling the large quantity of text data.

2.3 A Closer Look at Route Directions on the WWW

Since this thesis focuses on route directions from the WWW, it is necessary to determine basic features of the target documents by manually analyzing examples. Figure 2 demonstrates a typical route direction document found on the WWW. There are three particularly important features of route direction documents on the WWW. First, most route direction documents have the postal address of the destination on the same page with the route directions text. Route directions on the WWW are usually written to assist web site visitors to find the organization (a church, a restaurant and so forth). The author of these route directions is usually familiar with the area of the

destination (which is crucial for providing accurate route directions) and uses spatial language that is most suited for the route being described. Hence the address (or the postal code) can be used as a geo-stamp for the linguistic characteristic in the route directions. Second, text within each route direction document from the WWW can be divided into *destination*, *origin*, *route descriptions* and *unrelated text* [Zhang et al., 2009]. To investigate cardinal and relative direction usages, the spatial language usages in the first three categories should be extracted for analysis. Third, the linguistic component of route directions usually involves motion verbs (for example, drive, go, turn), landmarks (for example, PA-26, Pennsylvania State University), prepositions for spatial relationships (for example, on your right, to the east). Extracting the cardinal and relative direction usage in natural language usages is a great challenge for the analysis.

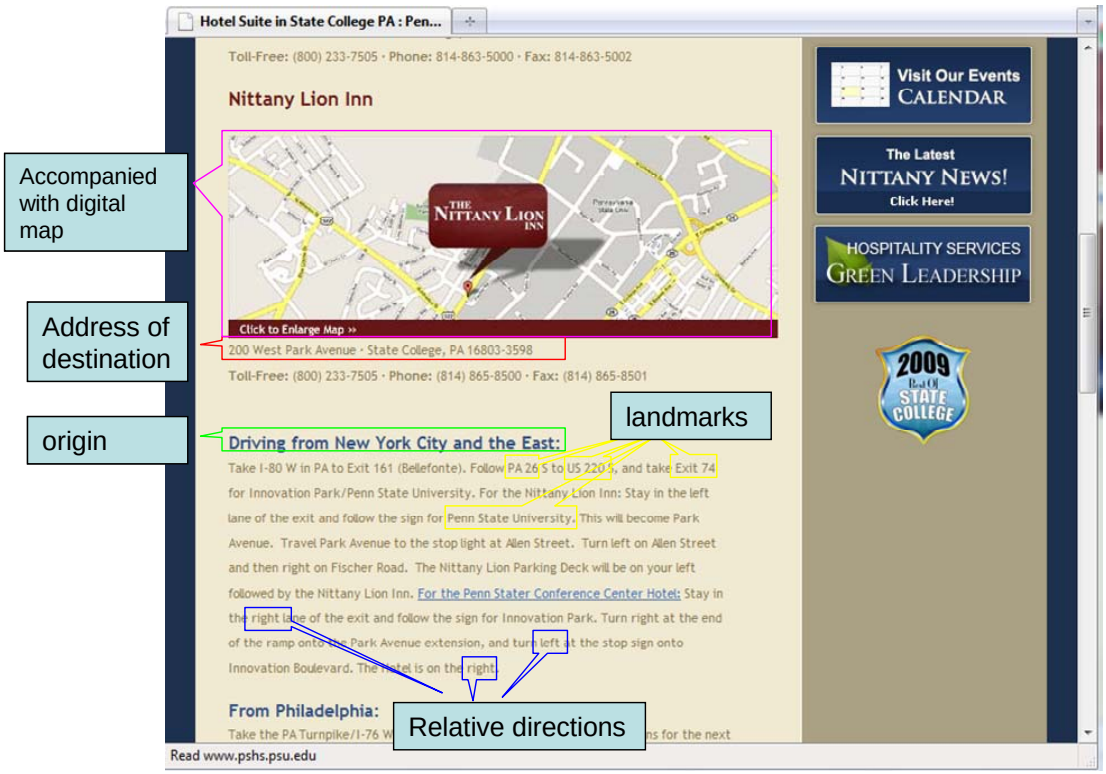

Figure 2. Screenshot of a route direction web page (from Nittany Inn website)

2.6 Challenges and Solutions

The research question of investigating regional variation in spatially language usage brings up challenges that are difficult to solve without computational tools. In this section, the challenges are detailed and potential computational solutions are discussed.

The idea of using documents from the WWW to meet the demand of analyzing route direction documents for regional linguistic variations poses two major challenges:

- First, develop a effective data collection schema for building a spatially-stratified motion-descriptive language corpora with documents from the WWW;

- Second, perform a linguistic analysis on a very large amount (more than 10,000 in our case) of text documents to reveal regional differences and identify usage patterns.

2.6.1 Data Requirements and Solutions

There are three important requirements (challenges) for the data collection in this thesis. First, distinguishing between the target route directions documents from other non-spatial language text is not an easy task. The linguistic complexity and flexibility of the target documents is quite high. Route directions are human generated linguistic documents (in contrast to stylized computer generated route directions) with all the computational challenges that the automatic analysis of natural language poses. Route directions do not have an apparent, distinct format that enables machine methods

to easily extract them from other unrelated text. This challenge can be addressed by applying the document classifier introduced by a maximum entropy model-based document classifier [Zhang et al., 2009]. The route direction document classifier can achieve 97% precision (refer to Section 3.3.1).

Second, documents from the WWW consist of not only plain text of one topic, but also html tags, advertisements, links, etc. These unrelated text parts may also carry words that are relevant to goals of this thesis (for example, html tags on picture positions, semantically different “right” in advertisements). Tools in sentence level tagging have been developed [Zhang et al., 2009], which provide a solution to this challenge. In this thesis, because each word of interest will be examined with its context semantically, the distraction from unrelated text parts can be ruled out in the semantic categorization process (refer to Section 3.3.1).

Lastly, although route directions are generated in a region, geo-referencing the route direction documents require further processing. The address of the destination often co-occurs with route directions, providing a potential geo-referencing marker. However, the scale of route directions (that is, the range of the route directions from origin to destination) cannot be covered by the address. This challenge can be solved by geo-referencing each route direction document on a postal code level, then organizing the whole corpus by postal regions (refer to Section 3.1.3).

2.6.2 Analytics Demands and Solutions

Analyzing linguistic phenomenon on a regional scale brings up another challenge. First, the spatial cognitive focus of this study requires researchers to look at the semantics of spatial language, which requires much more expert knowledge than the single word-based linguistic analysis method (such as part-of-speech classification). Designing a semantic category for target terms is a requirement for this research. Because the size of the SARD Corpus is quite large (considers more than 10,000 documents), text processing and analyzing tools [Turton, 2008, Turton and MacEachren, 2008] play an essential role in obtaining the regional linguistic characteristics for such a spatially-stratified corpus.

Second, the dual spatial and linguistic nature of this research requires an analysis that combines both. The combination of spatial analysis with linguistic data has been carried out since 1993 [Lee and Kretzschmar, 1993]. Lee focused on survey results. [Wise et al., 1995] also applied spatial analysis into spatial related text analysis. The regional linguistic characteristics from the SARD corpus can be obtained using semantic categorization and text processing tools. Then spatial analysis methods can be applied to investigate the regional variations with regard to the linguistic characteristics.

Third, the spatial nature of the linguistic analysis in this study makes geovisualization an appealing idea to assist analyst interpreting results. [Chen et al., 2007] provides a geovisualization toolkit for multi-visualization comparison (including

thematic maps, parallel coordinate plot, table view, scatter plot) that will be used in this thesis (see Section 3.3.3).

2.7 Summary of Background and Literature Review

Existing studies using spatial language provide insight into the spatial language usage differences in route directions. These studies offer a theoretical background for this thesis and guidelines for designing the analysis scheme. Web-based information retrieval, crowdsourcing, and corpus linguistics shed light on solutions to the challenges of data collection and data analysis. Having investigated possible solutions from data collection to analysis, the following development of methodology is proposed for regional analysis targeted at detecting regional linguistic variations.

3 Methods

This chapter addresses in greater detail how challenges raised in previous chapters are overcome and how the proposed approach is carried out. An interdisciplinary methodology is developed, integrating techniques and ideas from web-based information retrieval, computer linguistics, crowdsourcing, spatial database design, statistical model analysis and visual analytics techniques. Such a multi-perspective methodology is essential for solving complex analytical tasks addressing real world behavioral phenomenon (such as regional variation of spatial language). The components involved in this methodology are:

- Spatially-stratified sampling of documents from the WWW;
- Classifying route direction documents from the WWW using a combination of machine learning and rule-based algorithms [Zhang et al., 2009];
- Geo-referencing document locations on a postal code level (location validation);
- Developing semantic categories for terms of interest—in this study, cardinal versus relative direction terms.
- Using a text visualization tool, that is, the TermTree tool [Turton, 2008], to assist in processing the corpus for regional linguistic attributes.

- Using visual analytics tools that allow for analyzing language usage patterns.

- Detection, identification, and interpretation of regional patterns.

The topics above are organized into three steps: data collection phase (Section 3.1), case study design (Section 3.2) and data analysis (Section 3.3).

3.1 Data Collection

The first challenge of build a corpus from WWW documents that allows for analyzing regional linguistic variation in spatial language usage is distinguishing web pages that contain spatial language from web pages that do not. More specifically in the context of this thesis: Distinguishing web pages that contain route directions from those that do not. The linguistic complexity and flexibility of the target documents makes this task relatively difficult. To this end, an automatic text classification scheme for spatial language document is applied, which consists of a document classification algorithm [Zhang et al., 2009] and iterative training and testing (training set and evaluation of the document classifier is carried out by me).

The second challenge is geo-referencing route direction documents once they have been correctly classified. The postal codes that often occur together with route directions are used to geo-reference route direction documents from the WWW on a postal code level. Compared to fine scale location identification approaches, such as the geographic name entity recognition problem: [Lee and Lee, 2007], geo-referencing

documents from the WWW on a postal code level is much easier and more economical. After a postal code level geo-referencing process and a location validation process, I organize the SARD Corpus in postal regions.

3.1.1 Sourcing Route Directions from the World Wide Web

Web crawlers are a common tool for collecting documents from the WWW [Kobayashi and Takeda, 2000]. The web crawling process can starts from a few keywords of interests and a search engine—after querying the keywords of interest from the search engine, the result web pages can be used as seeds to continue web crawling. Similar strategy is applied for crawling route direction documents from the WWW in this study.

3.1.1.1 Seed Data for Web Crawling Route Directions

The first step of the crawling process is to determine the seed data. As several linguistic characteristics of route directions on the WWW are noted in Section 2.3 (directional verbs, distance nouns, addresses, .etc), there are potentially two ways of creating the query sets for search engines in this study: On the one hand, keywords commonly seen in route directions could be used. Examples are: "directions, turn, mile, go, follow, take, exit". On the other hand, postal codes could be used as seeds because they commonly co-occur with route directions, for example, "PA 16802". Keywords can be used to create a simple crawling schema as the returned hits from web

search engines usually contain a high volume of target documents. However, there are several drawbacks as listed below:

1) Keywords-based crawling will introduce linguistic bias. It is possible that there are spatial language documents of interest to this study that do not contain the keywords queried.

2) The returned result from keywords-based crawling cannot be guaranteed to contain route directions. It is possible that there are web pages of other topics that contain the predefined keywords. The ratio of route direction documents in the returned documents can be increased by using more keywords. The drawback of increasing the number of keywords is that the more keywords are used, the more it will limit the linguistic variety in the returned documents.

3) Keyword-based crawling cannot guarantee that the resulting corpus is spatially stratified.

In contrast, crawling using a postal code list is a convenient way to get geo-referenceable documents on a postal code level. The rationale for using postal codes as seed data (that is, query for search engines) to crawl from the WWW and build the target SARD Corpus is as follows: First, postal codes, as part of addresses of destinations, appear frequently together with route direction documents from the WWW. Postal codes provide an easy yet effective way to spatially locate the target document on a postal code level. Second, postal codes have a rigorous format (for example, in the US

the format is two capitalized letters followed by 5 digits, as in “PA 16802” or 9 digits, with a hyphen after the fifth digit as in “PA 16802-0001”, similar formats apply to postal codes in the U.K. and Australia). The low ambiguity of the postal code format makes it an ideal candidate for creating a spatially-stratified corpus with documents from the WWW. The benefits and limitations will be discussed in more detail in Chapter 5. Third, compared to crawling using keywords, postal codes introduce less linguistic bias as they only serve as a “geo-stamp” for addresses (which commonly appears in route directions). Fourth, because the postal code system has an extensive spatial coverage, it provides a way to ensure that the spatial language documents in the final corpus are spatially stratified. However, this crawling scheme requires a precise route direction document classifier, as there are many types of documents containing postal codes on the WWW besides route direction documents. However, the advantages of spatial coverage and non-linguistic bias are crucial for detecting regional linguistic variations. With a complete postal code list as seed data, crawling with postal codes offers sufficient spatial coverage that the WWW can provide on a postal code level. The comparison of advantages and disadvantages of the two proposed crawling schemes is shown in table 1.

Table 1. Comparison between two crawling schemes for retrieving spatial language documents

Crawling with:	Keywords	Geo-referencable Token (Postal code)
Example:	"Turn, left, right, mile, go"	"PA 16802"
Pros	Simple; Get result with related terms directly; With the proper keywords, result may have high related document ratio.	Extensive spatial coverage; No linguistic bias; Serve as a vague geo-referencing symbol (geo-stamp)
Cons	Cannot guarantee the topic of the document. Introduce linguistic bias; Cannot guarantee spatial coverage	Cannot guarantee the topic of the document. Result may have less related documents ratio. May drop relevant documents that does not contain postal codes

To summarize, postal codes are *spatially distributed, identifiable, and linguistically-unbiased* metadata for building the Spatially-strAtified Route Direction Corpus (the SARD Corpus).

3.1.1.2 Spatially-strAtified Sampling of Route Directions

Following the discussion of the advantages of using postal codes in the first step of the crawling scheme, complete lists of postal codes are needed as seed data to be put into the search engine. A list of postal codes for the U.S. (41,119 zip codes selected for the continental U.S., excluding Alaska, Hawaii, and all off-shore U.S. territories[1]) was

[1] The original zip code list of postal codes contains 42,293 all possible zip codes in the Territories of the United States. List of postal codes in the continental U.S. used in this study excludes locations where zip

therefore obtained free of charge from www.zipcodeworld.com [zipcodeworld.com, 2009], Jan.]. This list of postal codes is used as seed data to build the SARD Corpus of the Continental U.S. The format of the free version of the list of postal codes consists of State abbreviation, space, followed by 5 digits, for example, "PA 16802", which is the most common postal code format found in route direction documents. For the U.K., the Royal Mail (the post office in the U.K.) does not provide a list of postal codes to the public for free at the time when this study started; a substitute resource was retrieved from www.freethepostcode.org [2]. A total of 8860 postal codes are available [freethepostcode.org, 2009], Apr.]. Postcodes from this list appear in the format of 1 or 2 letters with 1 or 2 digits, space, followed by 1 or 2 digits and 2 letters, for example, "AB10 6BB". The list of postal codes for the U.K., although not officially complete, covers most of the U.K. For Australia, the official complete list of postal codes with 3312 postcodes was obtained from the Australian Post [AustraliaPost, 2009], Apr.] The format in the Australia postcode list is a region abbreviation followed by 3 or 4 numbers, for example, "ACT 2610". The above mentioned list of postal codes forms the seed data for route direction crawling.

---

codes starting with AK, HI, PR, VI, AS, GU, PW, FM, MP, MH or Military District (AE, AA, AP.) These state or district abbreviations stand for Alaska, Hawaii, Puerto Rico, Virgin Islands, American Samoa, Guam, Palau, Federated States of Micronesia, Northern Mariana Islands, Marshall Islands. For Military District, AE is Short for Armed Forces Africa, Armed Forces Canada, Armed Forces Europe, Armed Forces Middle East; AA is short for Armed Forces Americas (except Canada), AP is short for Armed Forces Pacific.

[2] The Ordnance Survey has released the full postcode database under an open license on May 10th, 2010. Postal codes data in the U.K. are now open to the public. http://www.freepostcodes.org.uk/

To obtain documents from the WWW with postal codes, a popular public search engine—Yahoo! is selected. Yahoo! is selected for the fact that it has one of the largest and most comprehensive databases—over 38,300,000,000 web pages and growing [Yahoo, 2009]. More importantly, from test experience, it is comparatively generous to web crawlers. Google, for example, restricts crawlers to 1,000 hits per day per IP, which slows down the crawling process greatly. Although Yahoo! also limits continuous crawling, the restriction on hits is much looser.

The web crawler program in this study was designed as follows:

1. Form Query: Read one line from the list of postal codes. To increase the rank of route direction documents in the returned documents, the additional keyword "Directions" is used to bring more target spatial language documents to the top of the list of hits.

2. Get the hits of the query in step 1 from yahoo!. Set N as the number of documents to retrieve—in this study N=20. By default Yahoo! returns 10 hits per page, hence the first N/10 pages of Yahoo!'s result page needs to be obtained.

3. Retrieve the returned web pages by collecting hyperlinks in Yahoo!'s result page and record the URL to a local text file "URL_list". Each new URL is compared with previous stored URLs to prevent redundant crawling: if the URL is new, go to step 4. If the URL matches with previously stored URLs, continue to Step 5.

4. Store the web page. Download the document associated with the URL and stored it. In this study, 20 web pages from yahoo!'s two result pages were obtained and stored under a unified naming convention: postal code followed by "result" and index of hit, for example: "*PA_16802_directions_result1.html*".

5. Repeat step 1-4 until all the postal code has been queried.

By setting N=20, the total number of crawled documents would be close to a million for the U.S. zip codes (20*41,119). Hence it is very important to have a generous web search engine that does not limit the time for crawling. In reality, there are some postal codes that return less than 10 hits so the actual number of crawled web pages was less than expected.

3.1.2 Text Classification for Spatial Language Documents

Because the crawling process doesn't guarantee the resulting documents are all route direction documents, it is crucial to develop a route direction document classifier to select documents of interest from the various documents retrieved from the WWW that contains postal codes. This task is accomplished by using a rule set and machine learning based text classification. To achieve high precision that ensures the validity of the analysis, iteration of training and classifying was carried out and evaluated.

3.1.2.1 Demand for Text Classification

Although postal codes in addresses are a clear feature in spatial language documents, there are many web pages where postal codes appear without route directions:

Type 1: Web pages that contain address of related organization, such as contact information;

Type 2: Address lists;

Type 3: Machine generated route direction text from map service provider's websites.

The type 3 web pages usually come from commercial map service providers, such as Google Map and MapQuest. Adding a URL checking step to avoid retrieving web pages from map service provider's websites in the crawling process can reduce this type of error greatly. Detailed pseudo code for the crawling process is included in Appendix I. Text classification is applied to deal with the other type 1 and type 2 errors.

3.1.2.2 Choosing Models for the Classifier

The classifier used in this thesis is built by [Zhang et al., 2009], which I participated in developing rule sets and evaluation of the classifier. The task of the text classifier is to separate route direction documents from non-spatial ones. A rule set is

created by analysts of typical linguistic patterns in route direction, such as “turn left at”, “the hotel will be on your left”. This rule set is used together with machine learning algorithms for binary text classification to accomplish this task [Zhang et al., 2009]. The classifier was built using MALLET 2—Machine Learning for LanguagE Toolkit, [McCallum, 2002], using machine learning algorithms to pick up linguistic features of training documents and differentiate these features from the negative ones. The workflow (Figure 3) consists of a hand-selected training set, a Maximum Entropy model-based classifier, and an iteration process to improve the accuracy of the classifier.

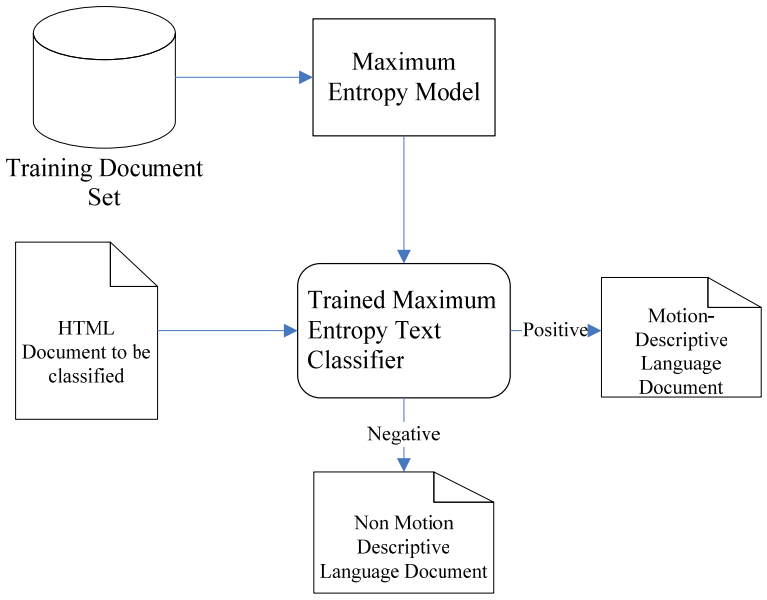

Figure 3. Maximum entropy classifier architecture

As illustrated in Figure 3, the binary classifier was built by feeding training document sets into the Maximum Entropy model [Nigam et al., 1999]. In this study, the training sets were prepared in two parts: 1000 route direction documents as *positive* documents and 1000 non-spatial language documents as *negative* documents. The idea

of binary text classification is to project linguistic features from both positive and negative documents into a feature space and calculate a boundary that separates the two classes. When a new document comes in, it can also be projected into the feature space and be determined by the trained boundary which class it should belong to [Nigam et al., 1999]. The 1000 spatial language documents were hand-selected human-generated documents from the WWW. Details about the design and coding of the classifier can be found in [Jaiswal, 2010].

3.1.2.3 Training and Evaluating the Classifier

To train the classifier, we start by feeding 500 route direction documents (selected from the 1000 positive set) and 500 non-spatial language documents (randomly selected from the 8906 negative set). After the classifier is trained, we evaluate its performance on classifying new documents from the crawling process. From 97 documents classified as positives, 75 were real spatial language documents while 22 were errors. It should be noted that 20 of the 22 errors belonged to Type 2 (Address List) errors. The classifier, although performs well on eliminating Type 1 errors (only 2 false positive in 97 classification), needs modification to perform better against Type 2 errors.

To increase the classifier's performance in differentiating Type 2 documents (address lists), we feed the 22 false positives with more Type 2 documents to the negative training set. The classifier was re-trained to calculate an updated class boundary in the linguistic feature space. After the second iteration, evaluation of 438

classified results gave over 93% precision (407 true positives). This precision was the limit of the classifier; further iterations did not improve the performance. The 7% error would not harm the analysis as the analysis scheme used in this thesis involves computational tool-assisted hand-tagging, which allows analysts to rule out the language usages in unrelated texts (refer to Section 3.2 & 3.3). The classification process greatly improved the precision of the SARD Corpus which shortened the analysis process and enable the SARD Corpus to serve other research purposes (refer to Section 5.3).

3.1.3 Location Validation

After the above two steps, the target corpus is gradually forming. To meet the demand of analyzing linguistic patterns regionally, the corpus should be organized in regions where each dataset (documents within a region) should contain documents of the corresponding region exclusively. In this study, documents are collected on a postal code level; the analysis, however, is performed on the higher regional level. Because the crawling method does not guarantee that a crawled document only contains postal codes of one region, we also needed to clear out documents that contain multiple postal codes from more than one state. The location validation process in this study is a postal code check applied using regular expressions:

- Matching the regular expression of postal codes format (according to which nation the document belongs to) for content in each route direction document.

- Extract the postal region abbreviation (state abbreviation in the U.S., state and territory abbreviation in Australia, postal region abbreviations in the U.K.) found in the matches. Compare the postal region abbreviation with the one that the file is assigned to (found in the file name). Exclude the document if a postal code from another postal region is found.

The rigorous format of the postal codes made the validation relatively easy without using complex geographic name entity recognition. For the U.S. and Australia, postal regions are defined according to the first 2 or 3 letters in the postal code (abbreviation for state or territory). For the U.K., postal regions are defined by a list of postal codes (refer to Appendix B for definition of postal regions in the U.K.). This is the last step to ensure the linguistic features from a document in a region are related exclusively to this region. Taking these prerequisites into account, the strata of the corpus are formed.

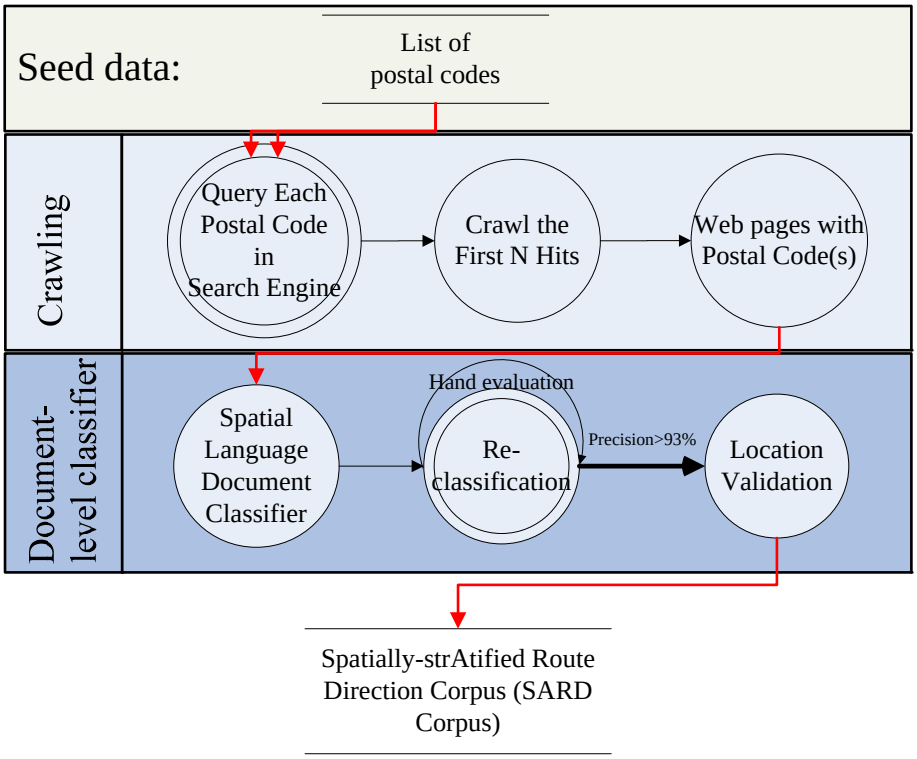

Figure 4. Detailed data collection scheme for building the SARD Corpus

After the above three steps, the data collection method for building the Spatially-strAtified Route Direction Corpus (SARD Corpus) is complete (see Figure 4). The SARD Corpus is built with the following characteristics: the region range covers the Continental U.S., the U.K., and Australia; the stratification is based on the postal region level while collection is conducted on postal code granularity; the certainty that the documents are true route direction documents is above 93%; the route directions collected in the SARD Corpus covers route directions in different environments (urban, rural, highway), different modes-of-transportation (mass transit, automobile, walking) and for different purposes ("how to get here" on a church website, or "hiking instructions" for a trail). The variety in the corpus provides representativeness of the

linguistic characteristics in them; the corpus contains a total of 10,055 documents for the continental U.S., 710 documents for the U.K., and 489 documents for Australia.

3.2 Case Study on Cardinal vs. Relative Direction Usage

To apply the proposed methodology and investigate regional linguistic variation by analyzing the SARC Corpus, cardinal/relative direction usage is chosen to be the case study. The topic itself has certain advantages for detecting variation in spatial language usage: First, as a linguistic phenomenon for spatial orientation, which is a popular topic in spatial cognition [Gladwin, 1970, Lewis, 1972, Gumperz and Levinson, 1996, Lewis, 1976, Davies and Pederson, 2001, Ishikawa and Kiyomoto, 2008], existing research on relative versus cardinal direction usages provides a theoretical background and guideline for analyzing them. Second, the difference in linguistic behavior of *cardinal/relative* direction usage, as a representation of human spatial reference systems, has been studied by experimental means (Ishikawa & Kiyomoto, 2008). Studying the difference on a regional scale (although by a different methodology) can be related to previous studies to determine whether people in different regions speaking the same language have different preferences about using cardinal or relative directions in giving route directions. Finally, from a practical point of view, cardinal/relative directions have a finite list of words (see Table 2). It is feasible for analyst to obtain a comprehensive usage of the directional terms.

Table 2. Cardinal direction and relative direction word list

Relative Direction Terms	**LEFT, RIGHT**
Cardinal Direction Terms	NORTH, SOUTH, EAST, WEST, NORTHEAST, SOUTHEAST, NORTHWEST, SOUTHWEST

Some may argue that UP, DOWN should be counted in the directional terms, example of usages: “traveling up from South Carolina to Maine”, where “up” is representing the abstract “North” as on a map. These usages are not commonly seen in route directions. Hence in this study they are not considered.

Based on the list in Table 2, usages of direction words can be extracted and analyzed using the SARD Corpus. For semantically enriched processing, we classified the usage of cardinal and relative direction terms as indicated in Table 3. These semantic categories have been established on the basis of our expertise in analyzing route directions and existing classifications proposed in the literature [Daniel and Denis, 1998]. We use *Token Occurrence and Proportion of usage*—common measures in corpus linguistics—to make comparison of directional term usages in different regions.

Table 3. Semantic categories for cardinal and relative directions

	Semantic categories	examples
Relative Direction	1. Change of direction	take a left, bear right
	2. Static spatial relationship	see a landmark on your right, the destination is left to a landmark
	3. Driving aid	keep to the left lane, merge to the right lane
Cardinal Direction	1. Traveling direction	head north, traveling south
	2. Change of direction	veer southwest on US Hwy 24, turn north
	3. Static spatial relationship	2 blocks east of landmark
	4. General origin	from North, coming from South of New York
	*used in POI names	North Atherton Street, West Street.

Using the semantic categories defined above, each document set (documents within the one region) can be analyzed with the result of token occurrences for each semantic category. Hence the regional linguistic characteristics can be obtained. For example, there are 796 route direction documents under the state PA in the continental U.S., from which every appearance of directional terms is categorized according to the semantic categories in Table 3. There are, for example, 5113 occurrences of relative directions (that is, “left”, “right”) being used to represent change of direction. This token occurrence count is one regional linguistic characteristic for the state of Pennsylvania. There are 7 regional linguistic characteristics for each region in this case study according to Table 3. Note that relative directions and cardinal directions can both be used to represent “Change of direction” or “Static spatial relationship” (underlined in Table 3), the preference within each semantic category (for example, whether cardinal direction or relative direction is preferred when used for “change of directions”) can

also be compared across regions. The process of categorizing each directional term usage from the SARD Corpus must be assisted by computational tools due to the corpus size. These tools are discussed in the next section.

3.3 Data Analysis

This section introduces two types of tools that make investigating regional linguistic characteristics on a geographic scale feasible: the text processing tool and the visual analytics tool. The function and procedure of each tool is described in details, as each plays a different role in the analysis process. Statistical analysis and spatial analysis is also introduced to provide geovisualization and validation of regional patterns detected in regional linguistic characteristics.

3.3.1 The Text Processing Tool: TermTree Tool

For the case study on cardinal versus relative direction usage in route directions, there are several challenges to analyze the regional linguistic characteristics and to perform a regional analysis. Corpus linguistics offered inspiration for analyzing a large amount of text documents. For instance, AntConc [Anthony, 2006] is a commonly used corpus linguistic analyzing tool with capacity to handle large quantities of texts. However because the target analysis goal of spatial language focuses on semantic categorization, the corpus linguistic tools that are primarily designed for lemmatization and concordance (two major procedures in corpus linguistic analysis) will not contribute to our analysis goal. Semantic categorization is a complicated task for which we have

yet to find a reliable computational method. To obtain semantic categorical data, expert examination by hand is still required. Fortunately, assisting tools are available to speed up the examination, which is crucial to accomplish the analysis within a reasonable time considering the number of documents (over 10,000) and size (over 5 million words) of the SARD Corpus. The TermTree tool [Turton, 2008, Turton and MacEachren, 2008] is a text processing and visualization tool used for assisting the expert hand examination by providing phrase occurrence counts. It takes regular expression as query and is capable of delivering tree structure visualizations of contextual information of target phrases. Figure 5 illustrates the interface of the TermTree tool and an example analysis scenario. On the bottom left panel, the "Search" text box takes queries with regular expressions and allows wide card tokens (that is, "\w" for a word). This capacity allows the analyst to specify a target word set, that is, "(left|right)", "(north|south|east|west|northeast|sou-theast|northwest|southwest)", and their preceding and following word to determine the semantic usage based on the context. The matched result phrases are shown on the top panel in a tree display, each phrase followed by an occurrence count of the phrase. Each phrase is also linked with many branches to the left (the preceding words) and to the right (the following words) for the analyst to investigate the context if the semantic category cannot be determined by the returned phrase. The original text is also shown in the bottom right panel, with selected returned results highlighted in red. The TermTree tool is capable of handling multiple text documents at the same time. This interface provides an analyst with both flexible context information—to determine the semantic category—and the convenience of getting the count of the same usage.

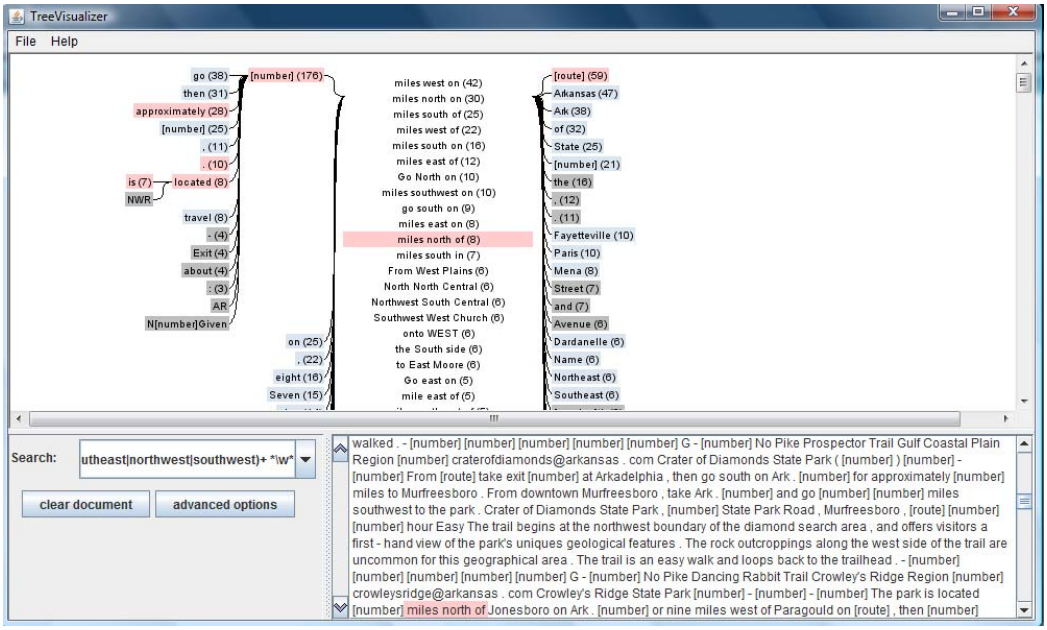

Figure 5. Screenshot of the TermTree tool during the analyzing procedure

3.3.2 Processing Regional Linguistic Characteristics

The detailed analysis procedure using the TermTree tool is listed below:

1. (Optional) For a document set (documents within one region), extract a set length of W (W=10) words in the adjacent context for every target direction terms occurred in each document. Store the extracted result of all documents within one region into a single text file. This procedure can prevent unrelated content in web documents from loading into the TermTree tool, such as html tags. This procedure speeds up the loading time of the TermTree tool.

2. Import the extracted text file (or original document sets) into the TermTree tool. Use the following two queries to capture three word phrases with the target word in the middle.

/w* *(left|right) /w*

/w* *(north|south|east|west|northeast|southeast|northwest|southwest) /w*

3. Manually categorize each of the phrases in the tree display (demonstrated in Figure 4) and record the token occurrence of each returned phrase. Each of the branches can be expanded to reveal adjacent words, in case the semantic category of the phrase is unclear or ambiguous.

4. Calculate the sum of token occurrences in the same semantic category. Get the regional linguistic characteristics of relative semantic usages (3 categorical counts) and cardinal semantic usages (4 categorical counts). The analysis should result in a data table similar to Table 4.

Table 4. Sample of processed regional linguistic characteristics data in Token Occurrence (RD: Relative Directions, CD: Cardinal Directions)

Postal Regions / RD's Semantic Category	AL	AR	AZ	CA	…
Represent change of direction	754	315	269	5727	
Indicating static spatial relationship	59	85	88	901	
Indicating driving aid	14	1	3	126	
Postal Regions / CD's Semantic Category	AL	AR	AZ	CA	
Representing traveling direction	112	193	58	1477	
Representing change of direction	16	14	31	115	
Indicating static spatial relationship	43	131	46	348	
Indicating General origin	8	8	12	228	
*.used in POI(doesn't count as direction usage)	87	94	82	495	

The above procedures greatly improved the efficiency of the hand-labeling process. Work time for semantic categorization of the SARD Corpus is shortened to less than 100 hours of one expert's work. How to visualize and interpret the results of the regional linguistic characteristics and detect regional variations are discussed in the next section.

3.3.3 The Visual Analytics Tool: Visual Inquiry Toolkit (VIT)

After analyzing regional linguistic characteristics, the information that the SARD Corpus carries becomes quantitative: numbers associated with regions that

represent the occurrence counts of target terms divided in semantic categories linked to the geographical region as the linguistic characteristics. The Visual Inquiry Toolkit (VIT) [Chen et al., 2007] allows for convenient map visualizations on the basis of regional attributes (counts for each semantic category) and corresponding shapefiles. It allows for multiple visualizations at the same time, including map displays, table views, and Parallel Coordinate Plots (PCP). The VIT is capable of highlighting regions and linking corresponding displays in different windows. It provides a convenient multi-view interface for an analyst to get a visual understanding of the data, which is demonstrated to be effective in interpreting regional linguistic characteristics. Figure 6 demonstrates a multi-visualization set up of the 7 regional linguistic characteristics of the U.S. The window on the top left is the map visualization of one regional linguistic characteristic. The scatter plot on the top right window shows the projection of the U.S. states onto a two dimensional space where X axis and Y axis are selected linguistic characteristics. The parallel coordinate plot in bottom right window shows the 7 regional linguistic characteristics (each vertical bar represents one regional linguistic characteristic). The three visualizations are linked when a state is highlighted (by mouse-over). Figure 6 shows that Illinois (the highlighted state) has a high proportion of cardinal directions used for “traveling directions” (the forth bar in PCP) and a low proportion of cardinal directions used for “change of directions” (the fifth bar in PCP).

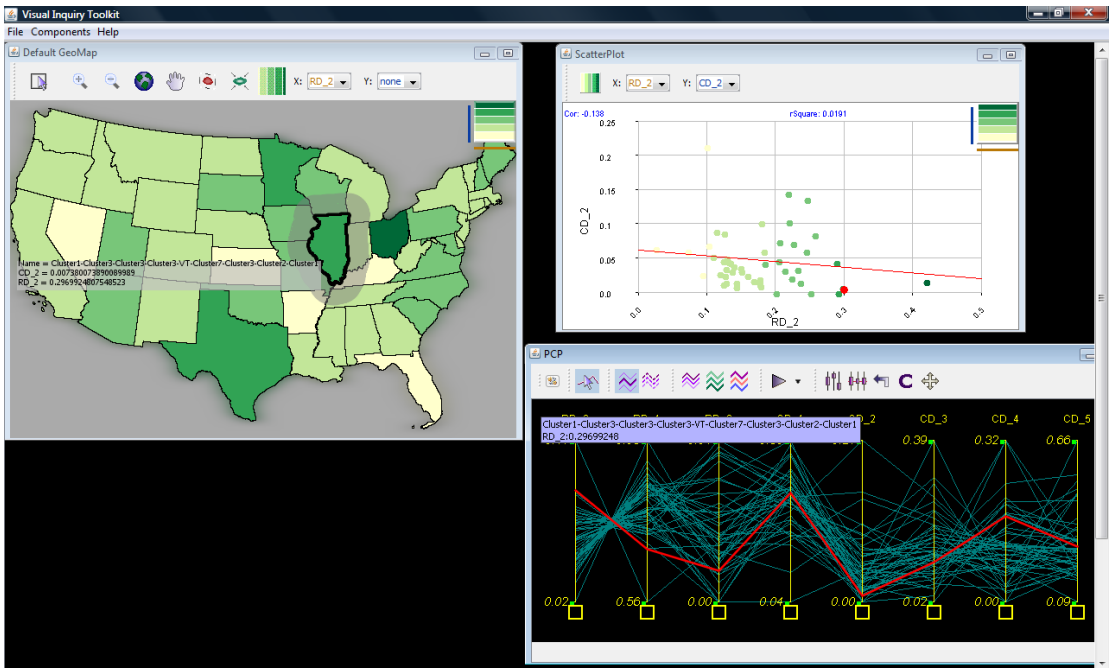

Figure 6. Screenshot of the VIT during visualized analysis procedure

3.3.4 Statistical and Spatial Analysis

The analysis result in the form of geographical region names and their multiple attributes can be used for cluster analysis of these regions to see if their linguistic attribute similarity is in concord with their spatial relationships. In this study, K-means clustering [MacQueen, 1967] is applied and evaluated based on the data retrieved from the TermTree tool. Moran's I [Moran, 1950] is a spatial autocorrelation measure that can be used in this study to evaluate whether the variation in regional linguistic characteristics is spatially autocorrelated, revealing whether Tobler's First Law of Geography [Tobler, 1970] that near things are more related than distant things holds. The various analysis methods can offer insights from various interpreting perspectives.

3.4 Methodological Architecture Overview

To sum up, this chapter illustrates the proposed methodology from data collection to data analysis, and for investigating regional variation in spatial language. At a high level, the general steps are: create (or obtain) seed data for crawling, develop a crawling scheme, applying a text classifier and evaluate it for high precision, geo-reference documents from the WWW, develop analysis schemes, and use visual analytics to interpret the results. The results following this workflow are interpreted and presented in Chapter 4 while the implications and extensions of this workflow are discussed in Chapter 5.

4 Results

In this chapter, statistics of the SARD Corpus are presented to illustrate the location of each route direction document in the corpus on a postal code level. Using the semantic categorization analysis scheme introduced in Section 3.3, the regional linguistic characteristics at both the national level and the regional level are presented and mapped. From a statistical analysis perspective, results from K-means cluster analysis [MacQueen, 1967] are also mapped to assist interpreting regional variation and detecting geographic patterns. From a spatial analysis perspective, global Moran's I [Moran, 1950] is calculated to evaluate the spatial autocorrelation structure in regional linguistic characteristics. This chapter presents the output from the proposed analysis scheme using histograms and map visualizations from the visual analytics tools.

4.1 The Statistics of the SARD Corpus

The SARD Corpus has the following statistics. It includes mostly direction documents with the expectation based on a sample evaluation that 93% of the documents included are actually direction documents. It covers documents from three English-speaking countries where the WWW is highly popular, that is, the U.S., the U.K., and Australia; each nation contributes 10,055, 710 and 489 route direction documents respectively. The number of route direction documents differs in the three nations, which could be caused by population or area differences (refer to Section 5.2).

The SARD Corpus is organized in a hierarchy: at the first level, the corpus is organized in nations; at the second level, documents within each nation are organized in postal regions (states for the U.S. and Australia, postal districts for the U.K.); each document is assigned to one postal code as its seed data used for crawling. Route direction documents with postal codes from multiple regions are filtered out from the SARD Corpus, leaving only documents with postal codes from one postal region. Table 4 lists the important attributes of the SARD Corpus.

Table 5. Attributes of the SARD Corpus (Spatially-strAtified Route Direction Corpus)

Attributes	Value
Corpus topic	Route Directions
Document format	HTML
Language	English
Data source	WWW
Spatial coverage	The U.S., the U.K. and Australia
Size (total of documents)	11,254 documents (10,055 in the U.S., 710 in the U.K., and 489 in Australia)
Size (mega-byte)	203 MB
How pure is the corpus (percentage of true route directions)	93%
Organization	Nation — Postal Region — Postal code
Other	Redundancy checked; does not contain documents with postal codes from different postal regions.

Important information that can be obtained from the postal code-based crawling scheme is the source of the corpus—where do all those documents come from? By examining classified route direction documents crawled from the WWW for each postal

code, the number of classified route direction documents per postal code (refer to the data collection scheme in Section 3.1.1) across the three countries varies from zero to five—while most are fewer than two route direction documents per postal code. Out of the 41,105 postal codes used for crawling route direction documents in the U.S., 10,055 route direction documents are collected. The route direction document/postal code ratio is 0.25—that is, one route direction document is collected out of 4 crawled postal codes on average. The data source of the SARD Corpus in the continental U.S. is mapped in Figure 7. It is noteworthy that route direction documents often come from certain regions more than others. The high frequency regions (for example, The East Coast of the U.S. and big cities) contain higher population, which might be the cause of this phenomenon. First, a high population within the same country is usually related to higher urbanization. As a result, hotels, churches, restaurant and other organizations also have a higher density in the regions with higher populations. These organizations are more likely to put route directions on the WWW, as they are the major route direction document sources on the WWW. Second, the area each postal code is assigned to is much smaller in regions with higher population than the ones with lower populations. Hence given the same amount of area, the region with the higher population will be crawled more because there are more postal codes in this region. The effect of such density differences within the SARD Corpus is discussed in Section 5.

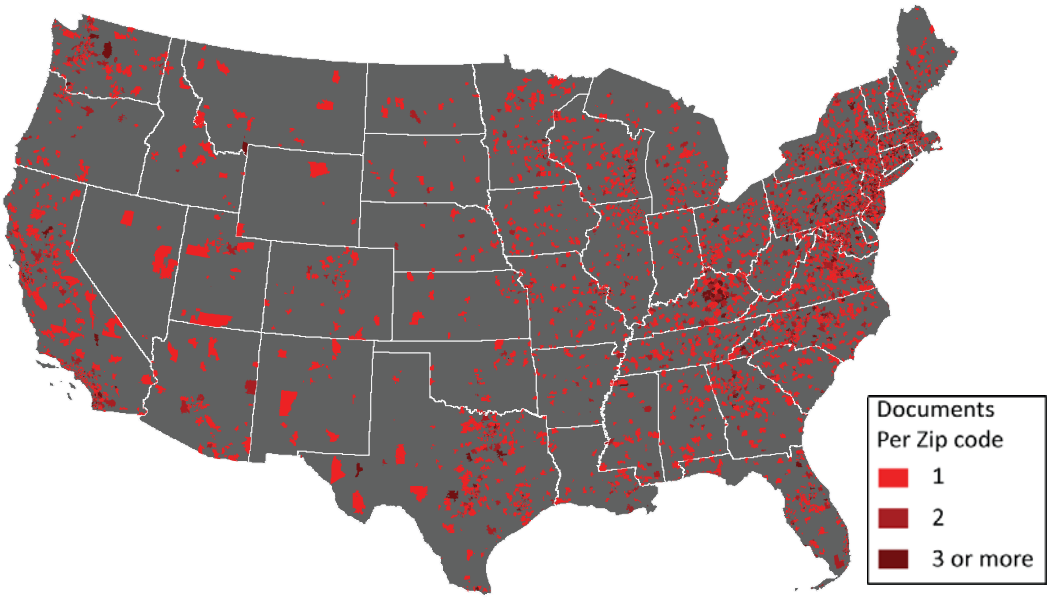

Figure 7. Data source of the SARD Corpus in the continental U.S. – by postal codes

For the U.K., Table 6 shows the statistics of the data. The overview statistics show that route direction documents come more frequently from certain regions (e.g., Midlands, North East England). However the difference among regions is rather small—the route direction document/postal codes ratio for most regions ranges from around 0.06 to 0.12. The overall route direction document/postal codes ratio is 0.08 for the U.K., which is lower than the 0.25 ratio for the U.S. This phenomenon has two potential indications. It shows that the people in the U.K. do not publish route direction on-line as frequently as the people in the U.S. do. It is also possible that the ratio difference results from the area or population difference between postal code regions in the U.K. and in the U.S.

Table 6. Statistics of the SARD Corpus in the U.K. – by postal regions

Postal region names	Total number of route direction documents	Number of postal codes per region	Ratio of route direction documents per region
East Anglia	56	838	0.07
London	68	1275	0.05
Midlands	93	862	0.11
North East England	108	925	0.12
Northern Ireland	1	55	0.02
North West England	56	793	0.07
Other Regions	2	23	0.09
Scotland	61	739	0.08
South Central England	102	1150	0.09
South East England	95	1254	0.08
South West England	49	669	0.07
EXETER	6	97	0.06
Wales	13	180	0.07
U.K.	710	8860	0.08

Table 6 shows the statistics of the route direction document from Australia in the SARD Corpus. Overall there are 3311 postal codes in the Australia list of postal codes, and the overall ratio of route direction documents/postal codes is 0.15, that is, between the ratios of the U.S. and the U.K. Similar to the U.S., there is a clear density difference in route direction documents in Australia. Route direction documents appear to come more frequently from NSW (New South Wales) and ACT (Australian Capital Territory, which is enclaved within NSW), both of which have higher population densities than other states in Australia. Compared to population and area data, the total number of route direction documents in a region is correlated with population rather than area. For

example, the area in the state VIC (227,416 km^2) is less than 1/7 of the area in the state QLD (1,730,648 km^2). However, because the population in both states are close in number (5.3 million in VIC and 4.3 million in QLD), the total number of route direction documents in the two regions are also close.

Table 7. Statistics of the SARD Corpus in Australia – by postal codes

Postal region names	Total number of route direction documents	Number of postal codes per region	Ratio of route direction documents per region
ACT (Australian Capital Territory)	4	12	0.33
NT (Northern Territory)	3	45	0.07
NSW (New South Wales)	213	976	0.22
VIC (Victoria)	94	755	0.12
QLD (Queensland)	82	471	0.17
SA (South Australia)	36	390	0.08
WA (Western Australia)	38	498	0.08
TAS (Tasmania)	19	164	0.12
Australia	489	3311	0.15

In summary, by examining the data source of the SARD Corpus, the route direction documents are collected from populated (or urbanized) regions more frequently. This phenomenon is quite obvious among documents in the U.S. (Figure 7) and Australia (Table 7), while the difference is not as distinctive among regions in the U.K. The statistics of the corpus provides insight into the data source. Understanding the source of the data is the foundation for further analysis and interpretation.

4.2 Result for Comparing Cardinal vs. Relative Directions

This section presents results of investigating regional variations of cardinal versus relative direction usages using the analysis scheme introduced in Section 3.2. The regional linguistic characteristics are presented and compared at both the national level and regional level. The semantic categorical usages are compared from two perspectives: within each direction type (for example, the token occurrence count and proportion of cardinal directions used for representing *change of directions* versus representing *traveling directions*) and between shared semantic categories (for example, when representing *change of directions,* the proportion of cardinal directions versus the proportion of relative directions). As cardinal directions and relative directions have different semantic categories (Table 3), the "within each direction type" analysis looks at the regional variations within cardinal or relative directions—that is, when people are using cardinal direction (or relative directions), is there a tendency to use it for one semantic category more frequently than others? On the other hand, both cardinal and relative directions can be used to refer to "change of direction" and "static spatial relationships." The "between shared semantic categories" specifically looks at the regional variations in expressing the shared semantic meanings—that is, when people want to express "change of direction" (or "static spatial relationship"), are there regional variations in using relative directions or cardinal directions? Histograms and maps are the two primary visualization means for interpreting the regional linguistic characteristics. Results from K-means cluster analysis and Moran's I are also presented to reveal regional patterns and their significance.

4.2.1 National Level Histograms

Because the SARD Corpus is organized in postal regions, the regional linguistic characteristics, which consist of token occurrence of cardinal and relative directions in different semantic categories, are organized in postal regions as well (refer to Table 4.) To make comparison at the national level, regional linguistic characteristics are summed up. The table below (Table 8) shows the token occurrence results on the national level (refer to Table 3 for example usages of the semantic categories).

Table 8. National level token occurrence in the following semantic categories: For relative directions, RD_1: "representing change of directions", RD_2: "representing static spatial relationship", RD_3: "representing driving aid". For cardinal directions, CD_1: "representing traveling directions", CD_2: "representing change of directions", CD_3: "representing static spatial relationship", CD_4: "representing general origin", CD_5: used in POI names.

Nation (No. of route docs)	RD_1	RD_2	RD_3	CD_1	CD_2	CD_3	CD_4	CD_5
US(10,055)	56425	10326	1670	11164	1023	3728	2887	7915
UK(710)	5786	2969	222	1325	41	378	705	2596
Australia(489)	2423	791	57	926	65	268	200	445

4.2.1.1 Comparison within Each Direction Type

On a national-level, Figure 8 illustrates that across the three nations relative directions are mostly used to indicate "change of direction" (the blue bar on the left). Similarly, cardinal directions are mostly used for indicating "travelling direction" (The crimson bar on the right). On the other hand, the preference for relative direction when

representing “change of direction” is much more common in the U.K. than in the U.S. and Australia. Correspondingly we find that cardinal directions are used more often in the U.S. and Australia than in the U.K. (the light blue bars on the right). Cardinal directions used for “static spatial relationship” is of lower proportion while cardinal directions used for “general origin” is of higher proportion in the U.K comparing to the other two nations.

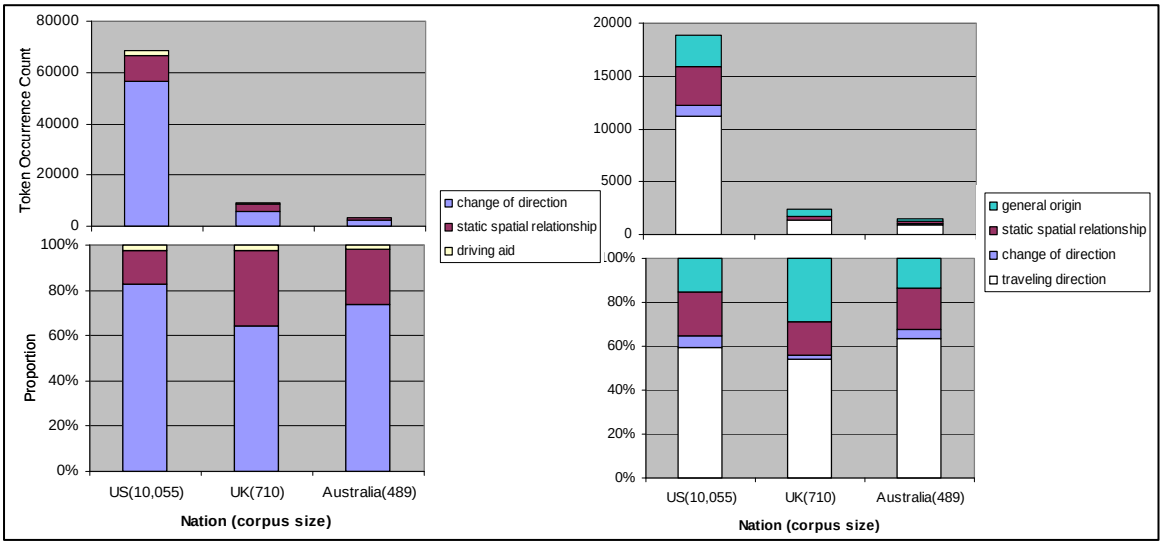

Figure 8. National level histogram of relative direction (RD) (left) and cardinal direction (CD) (right) usages (Top: token occurrence count, Bottom: Proportion)

4.2.1.2 Comparison between Shared Semantic Categories

The preference for using relative directions (RD) for “representing change of directions” is present in all three countries (shown in Figure 9). The cardinal directions representing “change of directions” are used less than 1% for the U.K and around 2% for the U.S. and Australia (orange bar in the right magnified proportion histogram).

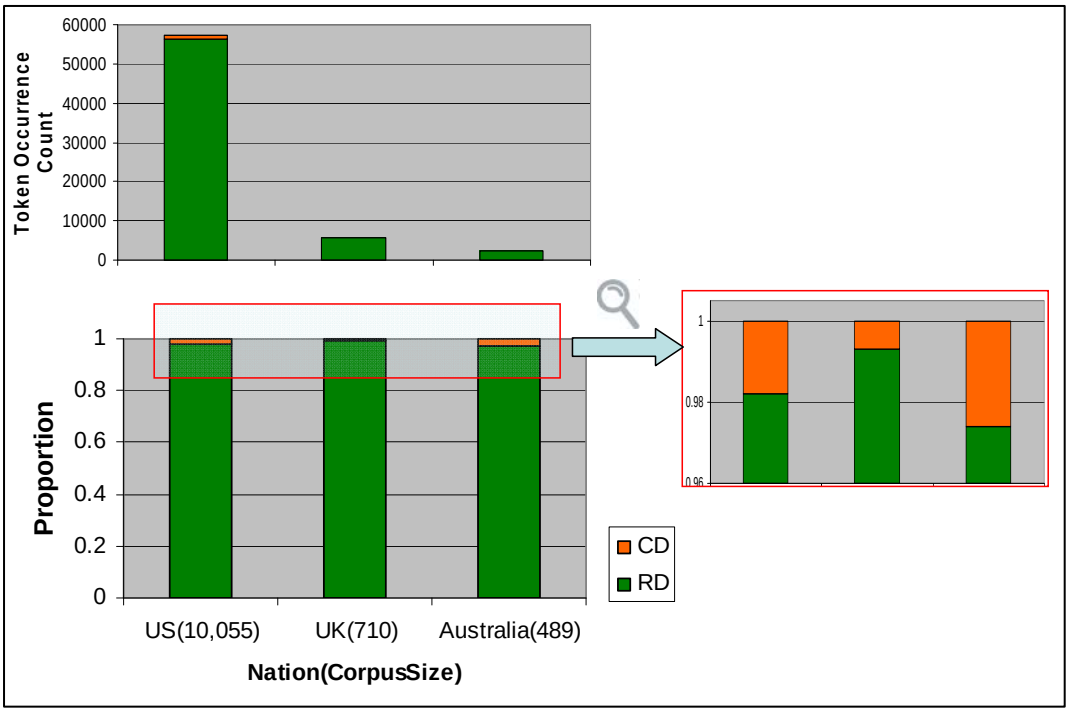

Figure 9. National level histogram of CD and RD usage in "change of direction" (Top: token occurrence; Bottom: proportion)

The other shared semantic category is "static spatial relationship", as both cardinal and relative directions can be used to describe spatial relationships (for example, "the destination will be on your left", "the hotel is to the north of the church"). Figure 10 shows a dominant preference for using relative directions in this semantic category. 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%.

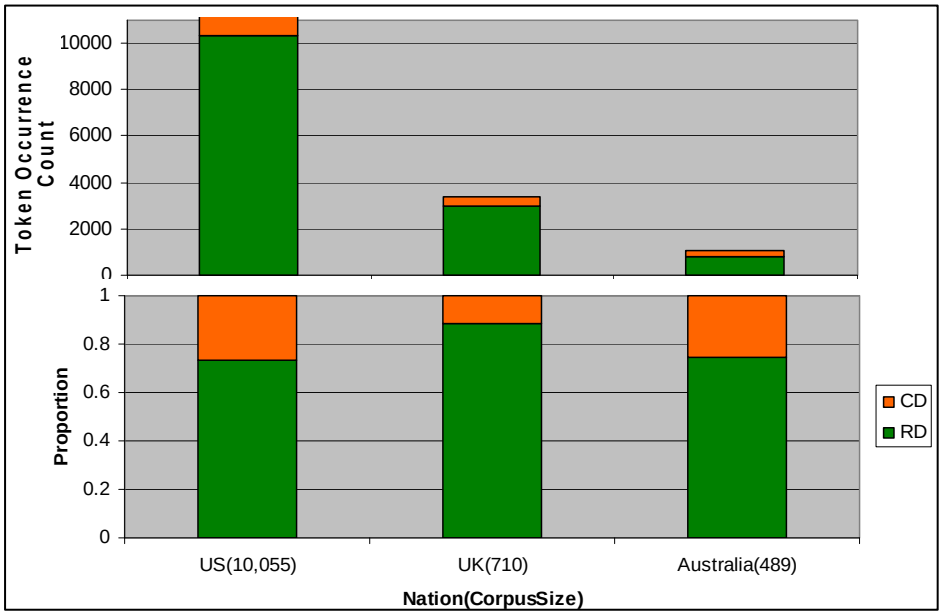

Figure 10. National level histogram of CD and RD usage in "static spatial relationship" (Top: token occurrence; Bottom: proportion)

To sum up the analysis at the national level, the three countries share similar linguistic characteristics in route directions but variations do exist. Despite the difference in corpus size, route direction documents from the U.S. and Australia appear to be similar while the ones from U.K. show some differences. From the semantic category comparison, the preference of relative directions is more common in the U.K. than in the other two countries. National level analysis offers a view of similarity and difference in the linguistic characteristics at a large scale.

4.2.2 Regional Level Histograms

Given the similarities and differences of linguistic characteristics at the national level, investigation of the linguistic variations at the regional level is carried out. Linguistic characteristics within every postal region in the U.S., the U.K. and Australia are analyzed and compared in this section[3], providing a deeper understanding of regional variation.

4.2.2.1 Comparison within Each Direction Type

In this section, the token occurrence counts and proportion of cardinal directions and relative directions are analyzed in postal regions. Postal regions are the 49 states selected in the continental U.S., 13 predefined regions (refer to Appendix B) in the U.K., and eight states and territories in Australia.

Figure 11 below shows the relative direction usages in token occurrence counts and proportions in the U.S. From the token occurrence histogram, there is a clear difference in the total token occurrence count in each state. Some states (such as CA, PA and NY) return more relative directions than other states. This is because there are more route direction documents in these states (CA, PA and NY have 955, 796, 852

[3] Because the data table for regional linguistic characteristics is too long to display (a sample of data table is shown in Table 4), the data in table format in included in Appendix C.

route direction documents respectively). The difference in document size in each state bin caused the difference in token occurrence counts. This difference is normalized in the bottom proportion histogram. The variation of relative direction usages exists but is not distinctive between States from the histogram. The majority of relative directions are used to represent *change of directions* (the white striped bar, more than 70% of relative directions are used this way in 46 out of 49 states). Using relative directions to indicate s*tatic spatial relationships* constitutes 10%-30% of relative direction usages in 44 out of 49 states; *Driving aid*, although it appears from time to time, is very scarcely used. The proportion of *driving aid* in all states is lower than 5%, while in more than half (27 out of 49 states) it is lower than 2%.

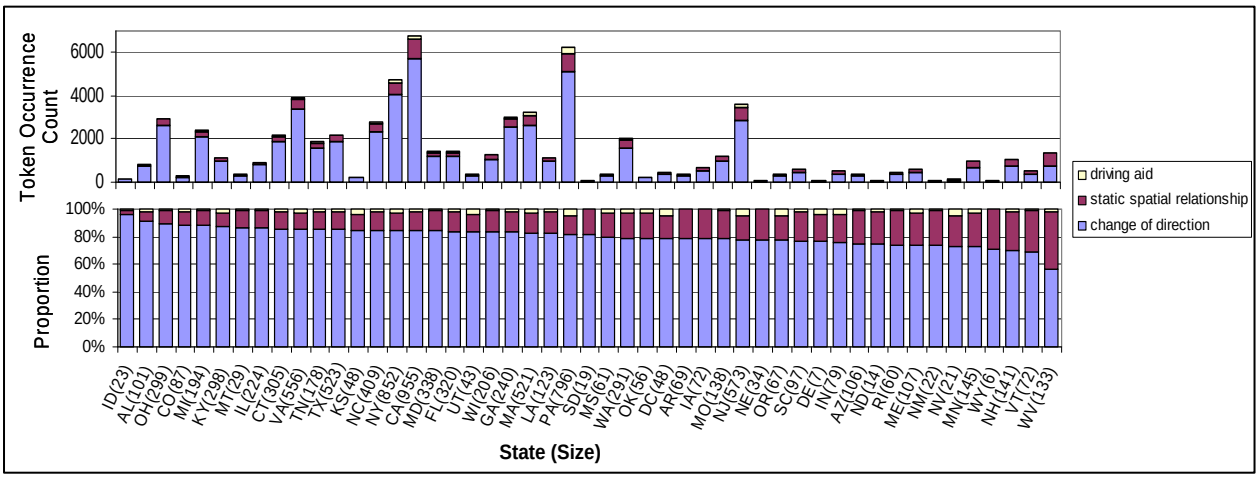

Figure 11. Regional level histogram of RD usage in the U.S. (Top: token occurrence; Bottom: proportion)

In contrast, cardinal direction usages showed more variations across the U.S. In Figure 12, from the top token occurrence count histogram, the states with the most route direction documents still return the highest number of token occurrence counts. However, it is noteworthy that PA outnumbered NY in relative direction usage while in

cardinal direction usage NY outnumbered PA. This phenomenon might be caused by the following circumstances: there are more cities (urbanized areas) in NY than in PA—62 cities in NY and 57 cities in PA. Cities, especially metropolitans, often have more street grids while there are more winding roads in rural areas. The direction on a straight road in a street grid can be represented by cardinal directions while winding roads do not have a consistent direction. Hence using cardinal direction is more sensible in the cities than the rural areas. The higher occurrence of cardinal directions in NY than PA could be influenced by the street grid patterns. The study in [Lawton, 2001] also supported this explanation.

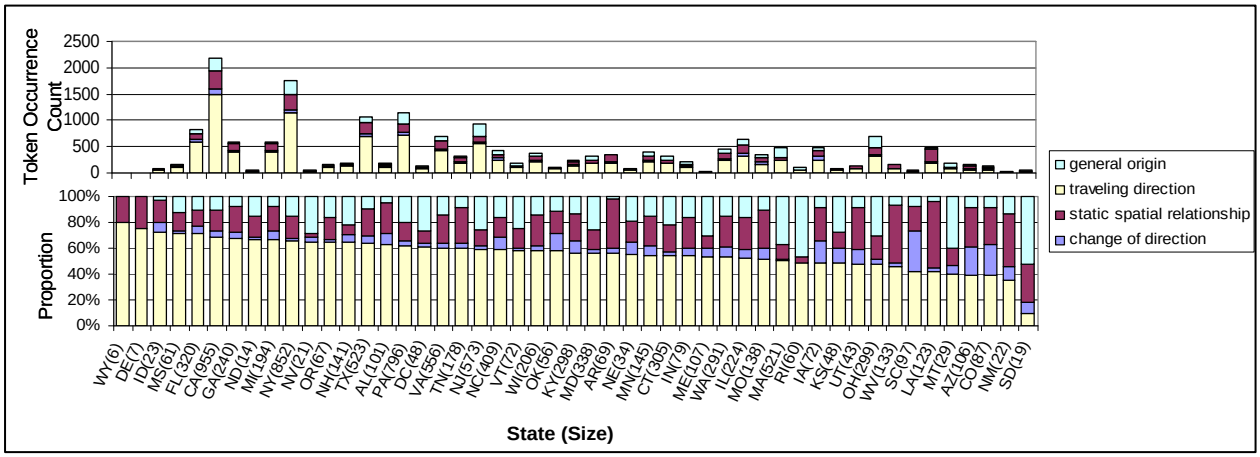

Figure 12. Regional level histogram of CD usage in the U.S. (Top: token occurrence; Bottom: proportion)

From the bottom proportion histogram in Figure 12, the majority in semantic categories is indicating *traveling direction* (takes up more than 40% of cardinal direction usages in 28 out of 49 States). The cardinal directions representing *change of directions* are the least frequently used category. 44 out of 49 states contain less than

10% usage of cardinal directions for *change of directions.* Cardinal directions used for *static spatial relationship* and *general origin* vary from state to state.

For the U.K., relative direction usages are shown in Figure 13. Proportions of relative direction usage are found to be similar to the U.S. The semantic category *change of direction* constitutes 40% to 80% of all relative direction usages, while the proportion is around 60% in 11 out of 13 regions. The proportion of *static spatial relationship* usage is around 20% for 9 out of 13 regions. *Driving aid* is the least commonly seen semantic usage. The high proportion of *driving aid* in Exeter and Northern Ireland are due to the small sample size in these regions.

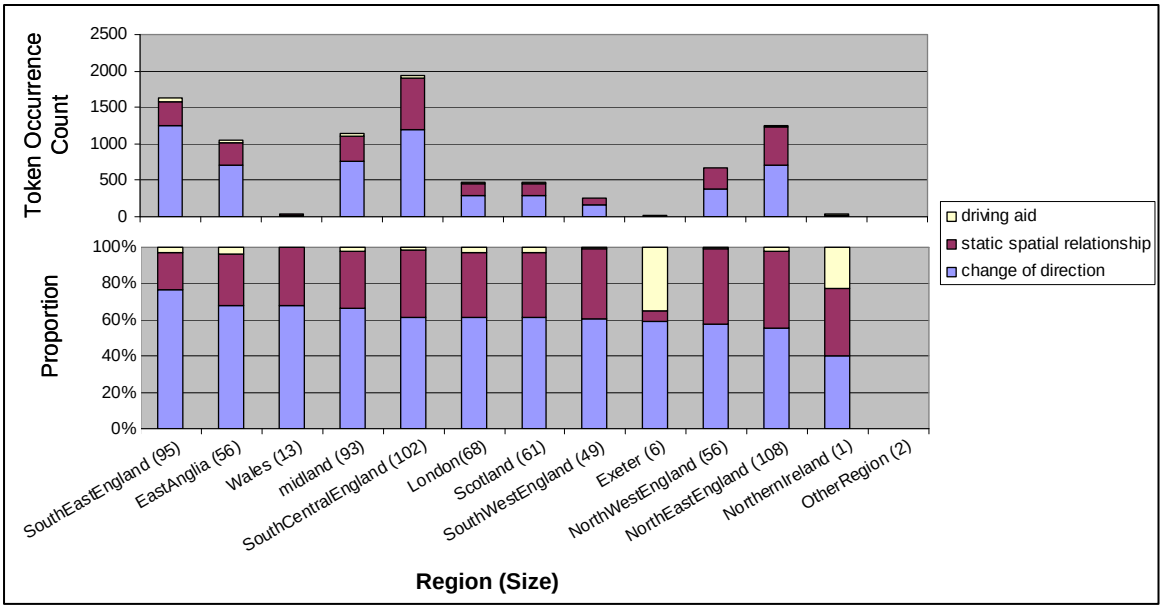

Figure 13. Regional level histogram of RD usage in the U.K. (Top: token occurrence; Bottom: proportion)

Figure 14 shows the cardinal direction usages in the U.K., which contains more variation than relative direction usages. It is noteworthy that the total token occurrence count for cardinal directions and relative direction is different. Midlands contains more

cardinal directions although the number of documents (56) is lower than other regions such as Southeast England, South Central England.

The proportions of cardinal directions used to represent *traveling direction* ranges from 40% to 60% in 8 out of the 13 regions. *General origin* takes up around 25% in 8 out of the 13 regions. *Static spatial relationship* takes up less than 25% in 9 out of 13 regions. Contrary to relative direction usages, the least commonly used semantic category for cardinal directions is "change of directions". All 13 regions contain less than 3% of cardinal directions used in this category. It is noteworthy that the usage of cardinal directions in South West England is quite different from the others, with high proportion for *General Origin*. This phenomenon may be the result of multiple route direction to the same destination in one document (refer to Chapter 5).

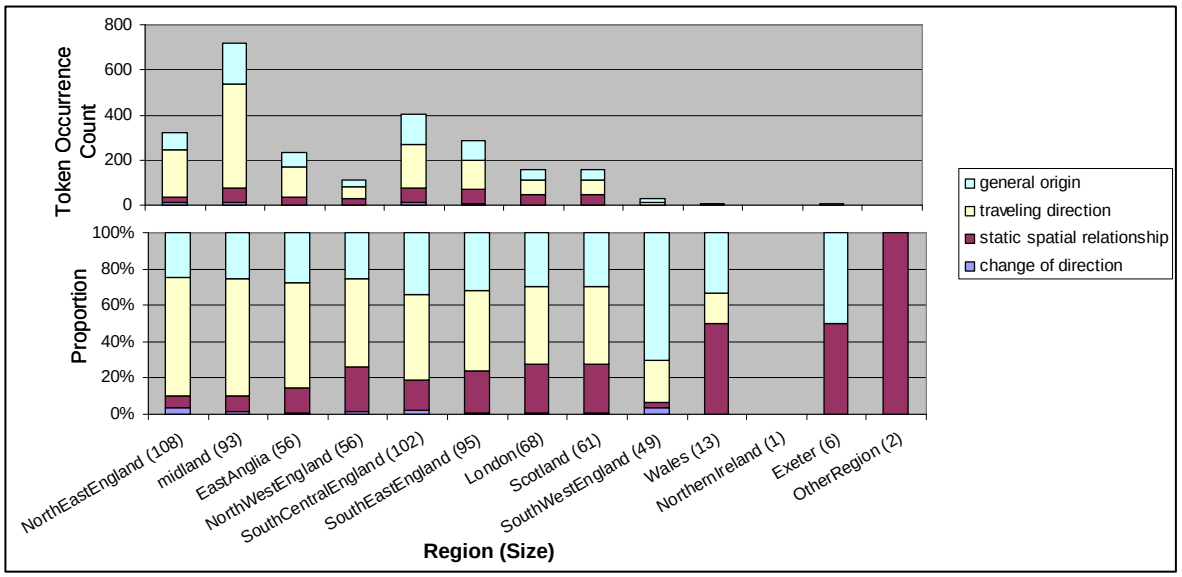

Figure 14. Regional level histogram of CD usage in the U.K. (Top: token occurrence; Bottom: proportion)

For Australia, the number of route direction documents from NSW (213) is much higher than for other regions. This is further evidence that the density of the data source may be related to population (or urbanization) of the region, as NSW contains Sydney, the largest and most populated city in Australia. NSW is also located on the east coast of Australia where the population is higher than farther inland from the West coast. Figure 15 shows a comparison of relative direction usages across all postal regions in Australia. The *Change of direction* category takes up from 70% to 85%; *static spatial relationship* ranges from 15% to 30%; *driving aid* appears least frequent with less than 2% for all regions.

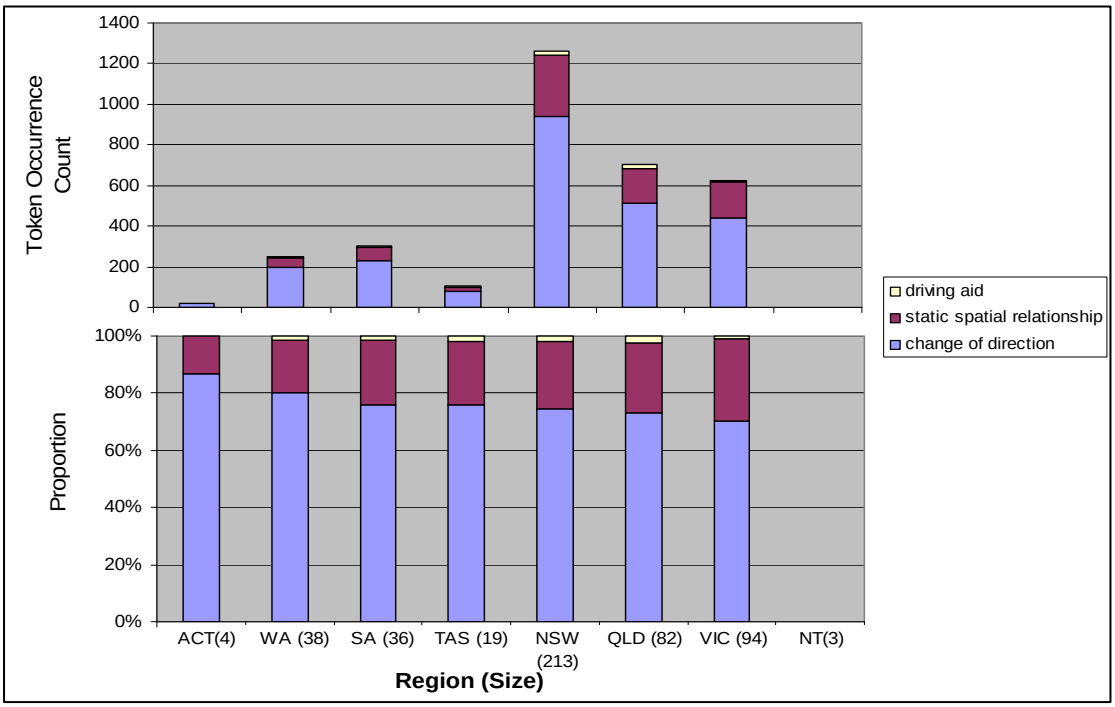

Figure 15. Regional level histogram of RD usage in Australia (Top: token occurrence; Bottom: proportion)

Cardinal direction usages in Australia show more variations than relative direction usages. The top histogram of Figure 16 shows total token occurrence count for cardinal directions do not correlate to the number of route direction documents in one region (SA has much less documents than QLD, VIC and WA but yields more cardinal directions). Cardinal directions in Australia are mostly used for representing *traveling direction* (more than 50% for all regions). *Static spatial relationship* takes up 15% to 40% in 6 out of 8 regions, while the other 2 regions suffer from a low number of documents. The *general origin* category takes up from 10% to 23%. *Change of direction* takes up only less than 7% for all regions.

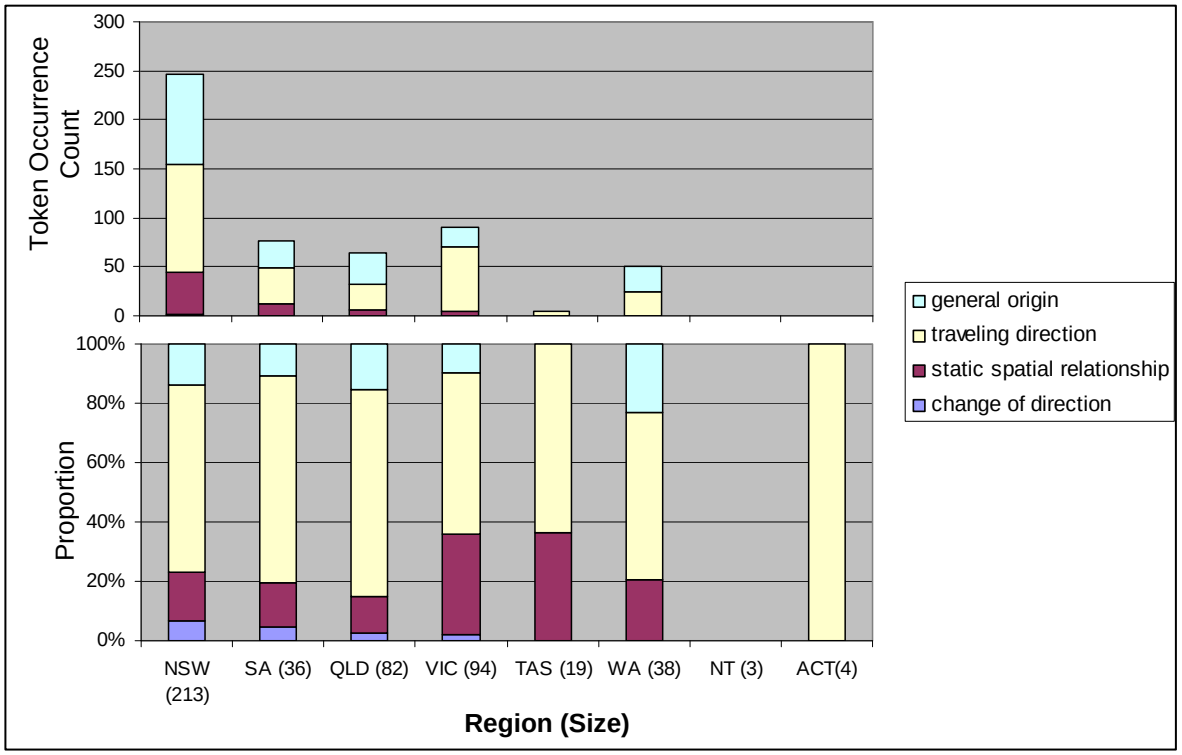

Figure 16. Regional level histogram of CD usage in Australia (Top: token occurrence; Bottom: proportion)

In summary, by comparing cardinal and relative direction usages at the regional level, the following observations are noteworthy. Token occurrence counts for relative directions usually relates to the number of documents in the region, while for cardinal direction it is not always the case (certain regions with less route direction documents contains more cardinal directions). This phenomenon might be affected by different street grids in different regions [Lawton, 2001]. Semantic analysis on the relative direction usages shows similar results for the U.S., the U.K. and Australia: the majority of relative directions are used to indicate *change of directions*, the rest are used for representing *static spatial relationship*; only a very small proportion (mostly less than 3%) are used for representing *driving aid*. For cardinal directions, the majority of regions use more than 50% of cardinal direction to indicate *traveling directions*. The least frequently used category for cardinal directions is *change of directions*, which is just the opposite of relative direction usages. There are more regional variations in cardinal direction usage, which is not easy to interpret from the histogram visualization. Therefore, map visualization of the linguistic characteristics of each region for each direction type is applied in Section 4.2.3, offering a clearer interpretation of regional variations. In Section 4.2.2.2, the token occurrence of relative and cardinal direction usages within the same semantic category are compared. This comparison offers insights on the differences when people use relative directions or cardinal directions to express the same semantic meaning.

4.2.2.2 Comparison between Shared Semantic Categories

In the predefined semantic categories, two of the semantic categories that were defined in Section 3.2 are shared by both cardinal and relative direction (*change of direction* and *static spatial relationship*, underlined in Table 3). In this section, these shared semantic categories are compared across the three countries. This comparison offers insight into cardinal versus relative direction usage from a different angle than Section 4.2.2.1. The previous section answered that if people are using relative direction or cardinal direction, which semantic categories are each direction term used for? This section seeks to provide insight into the situation in which people want to express certain spatial concept (motion or spatial relationship)—whether they choose relative directions or cardinal direction words. The overall goal of the analysis is to investigate the regional variations in the preferences to the two questions above.

For the U.S., Figure 17 shows the comparison within the semantic category *change of directions*. Relative directions take up a significant amount of direction terms (in 46 out of 49 states relative direction is used more than 90% when referring to *change of direction*). Two states (i.e., AZ and MT) use cardinal directions more than others with around 20%. Figure 17 shows a general preference of relative directions across the U.S. when expressing *change of direction.*

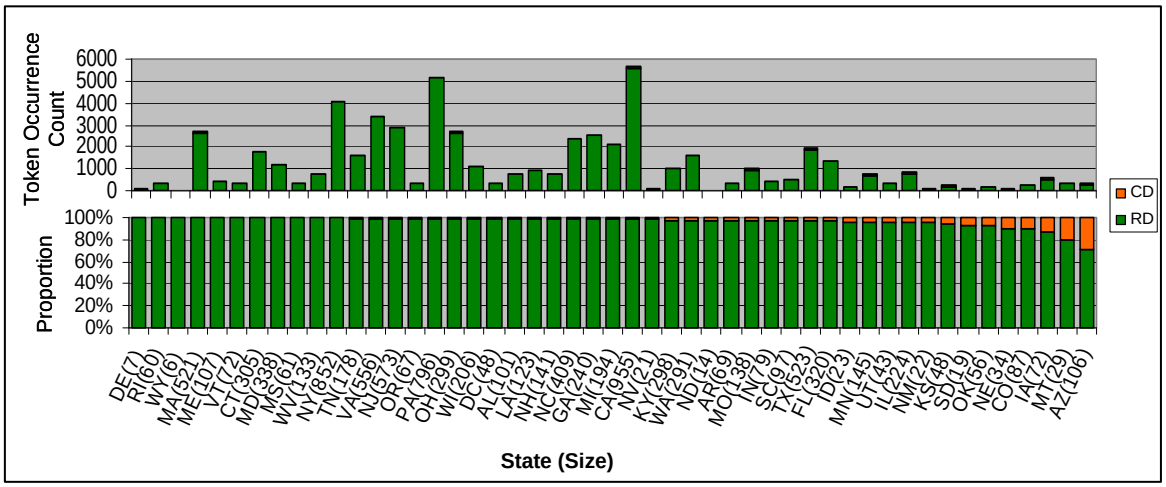

Figure 17. Regional level histogram of CD and RD usage in "change of direction" in the U.S. (Top: token occurrence; Bottom: proportion)

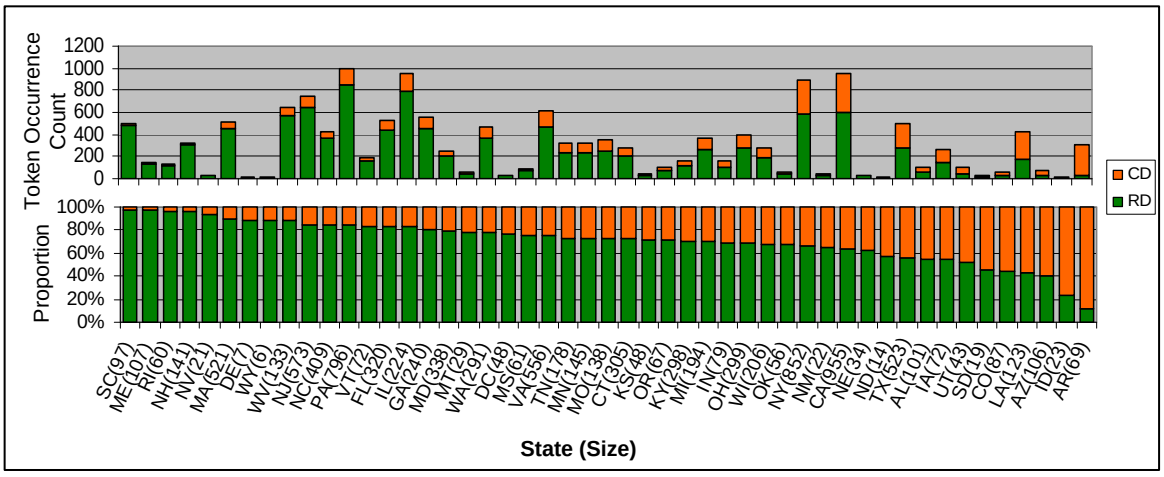

Figure 18. Regional level histogram of CD and RD usage in "static spatial relationship" in the U.S. (Top: token occurrence; Bottom: proportion)

Figure 18 shows the comparison with the semantic category *static spatial relationship* in the U.S., which has more variations than *change of direction* (Figure 17). The proportion histogram shows that although relative direction is preferred in more than half of the states (they take up 70% or more in 28 out of 49 states), several states have a contrasting pattern. 19 States use relative direction for *static spatial relationship*

between 30% and 70%. Moreover, there are extreme contradictions (in the state of AR and ID, cardinal direction takes up more than 70%) as well. This indicates that there is potentially a regional spatial linguistic preference on the choice between cardinal directions versus relative directions in *static spatial relationship.*

For the U.K. (Figure 19 and Figure 20) and Australia (Figure 21 and Figure 22), similar patterns can be observed. When people express *change of directions*, the preference for relative directions over cardinal directions is very obvious. Particularly in the U.K. (Figure 19), cardinal directions take up less than 1% in all regions. In the semantic category of *static spatial relationship*, in the U.K, the majority of the regions still prefer to use relative directions although this preference is not as dominant as for *change of directions* (Figure 20, 10 out of 13 regions represent *static spatial relationship* with relative directions for over 80% of the times). Similar preference can be found in Australia, too (Figure 22).

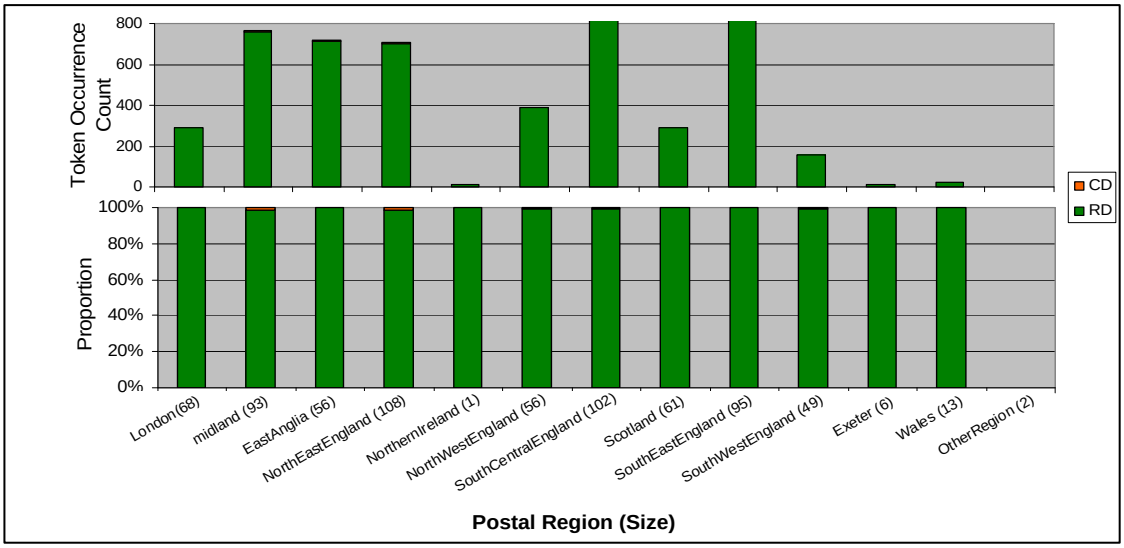

Figure 19. Regional level histogram of CD and RD usage in "change of direction" in the U.K. (Top: token occurrence; Bottom: proportion)

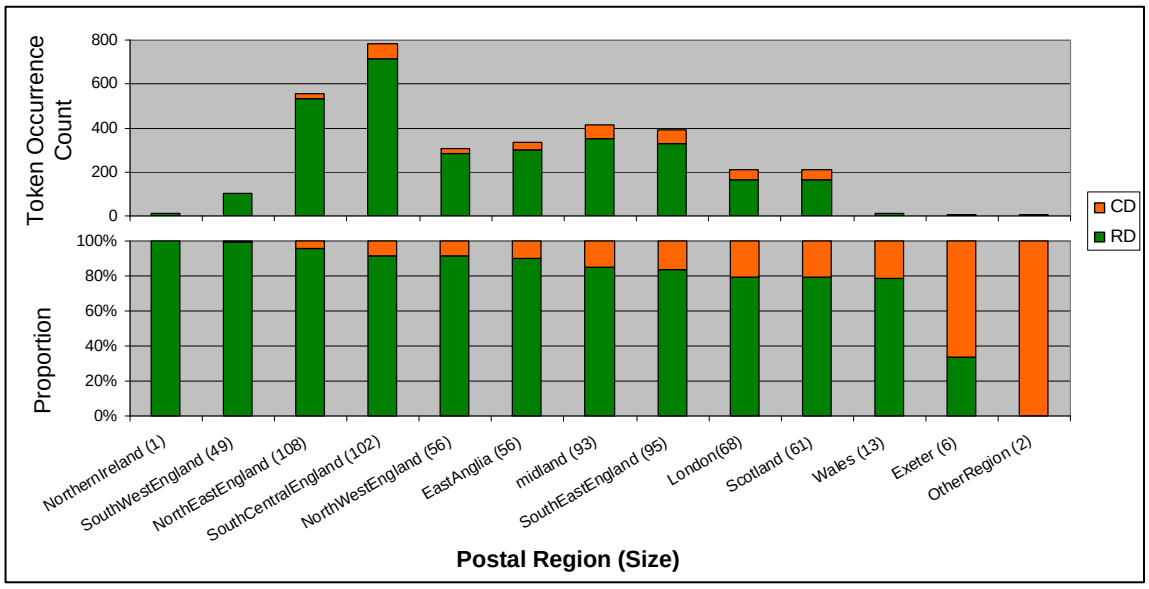

Figure 20. Regional level histogram of CD and RD usage in "static spatial relationship" in the U.K. (Top: token occurrence; Bottom: proportion)

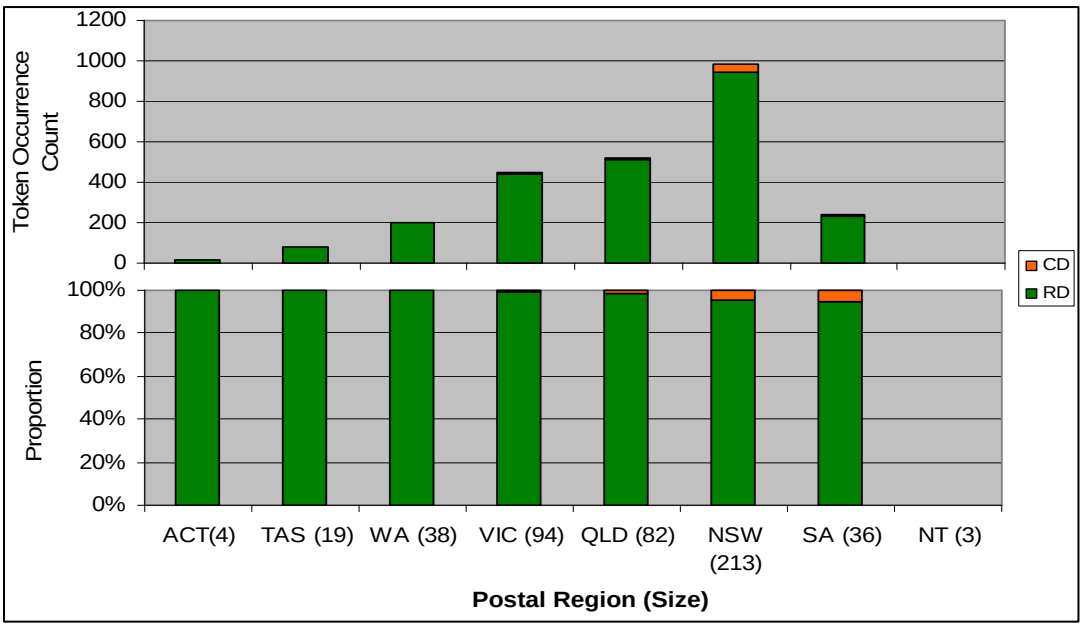

Figure 21. Regional level histogram of CD and RD usage in "change of direction" in Australia (Top: token occurrence; Bottom: proportion)

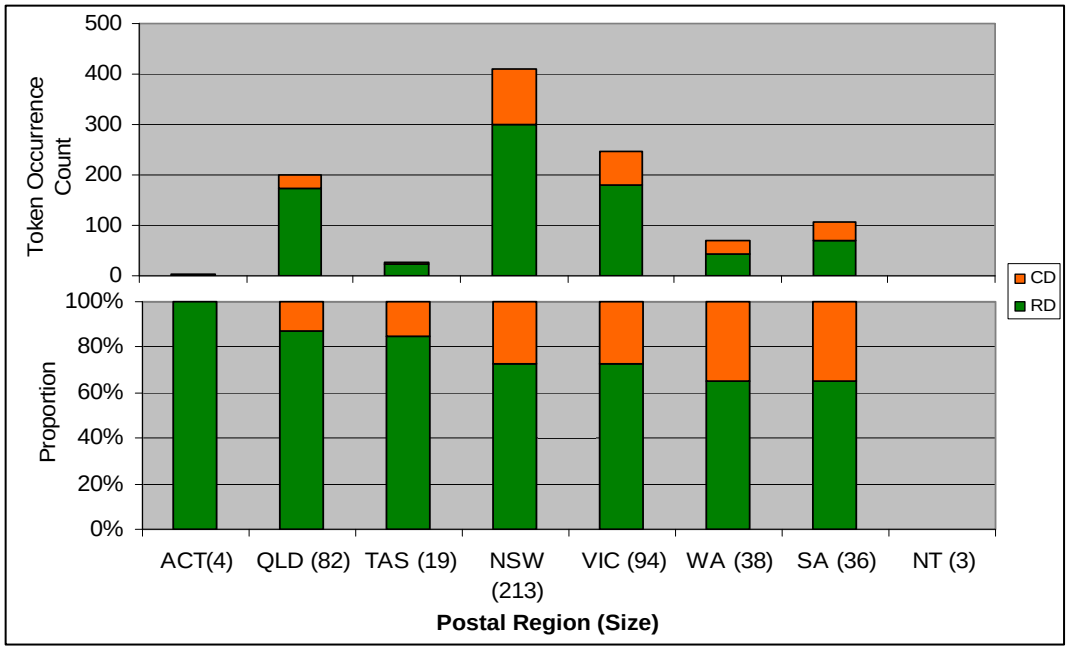

Figure 22. Regional level histogram of CD and RD usage in "static spatial relationship" in Australia (Top: token occurrence; Bottom: proportion)

In summary, when people express *change of direction* and *static spatial relationship* for route directions in the three countries studied, relative direction words are usually preferred. The preference is more obvious in the *change of direction* category. There are some subtle regional differences in the last two sections' analysis, which are not easy to identify in histograms. Maps, K-means analysis and spatial analysis could assist the interpretation, the results of which are presented in Section 4.2.3.

4.2.3 Regional Level Map Comparison

To get a better understanding of the regional variation of relative versus cardinal direction usages, each semantic category's proportion is plotted on a map for comparison. The plotted map can provide geographical knowledge about the regions, such as adjacency of regions, which may assist the analyst to detect regional variations in the linguistic characteristics. Figure 23 shows that the two most dominant usages as noted at the national-level (relative directions used for "change of direction", cardinal directions used as "travelling direction") are used more frequently in most states in the U.S. (Figure 23a, 23c). For cardinal direction usage, there is a geographic pattern (South Dakota to Kansas, Wyoming to Iowa, blue circled area) that differs from its surroundings states in every semantic category. The regional pattern detected is comparable to the Colorado West and Central West region in the map of U.S. dialect [Smith, 2006, p.186]. A possible explanation for this observation may lie in the correlation between the regional linguistic preference and regional geographical features, which is yet to be investigated.

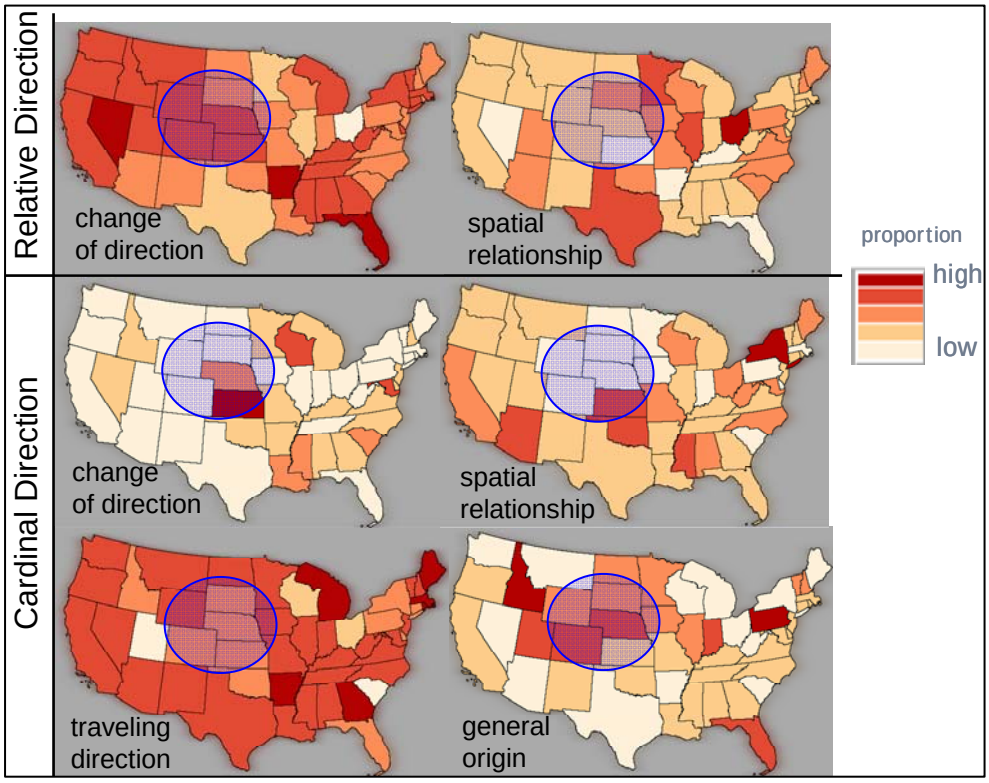

Figure 23. Region-level comparison of RD and CD usages in the U.S.

Unlike the 48 states in the U.S., the shapes of the postal regions in the U.K. are more irregular and most have coastlines. From Figure 24, we can see that the northwest regions of the U.K. differ from the east regions (including Northern Ireland). There is a regional pattern in the Midlands, East Anglia and Northeast England which stands out from their surrounding regions in Figure 24b, 24d, 24e and 24f (blue circle). The regional difference noted might be caused by culture or linguistic (dialectal) differences. However, some abrupt difference in cardinal direction usages might be caused due to regions with a small number of route direction documents. The effect of sample size difference is discussed in Chapter 5.

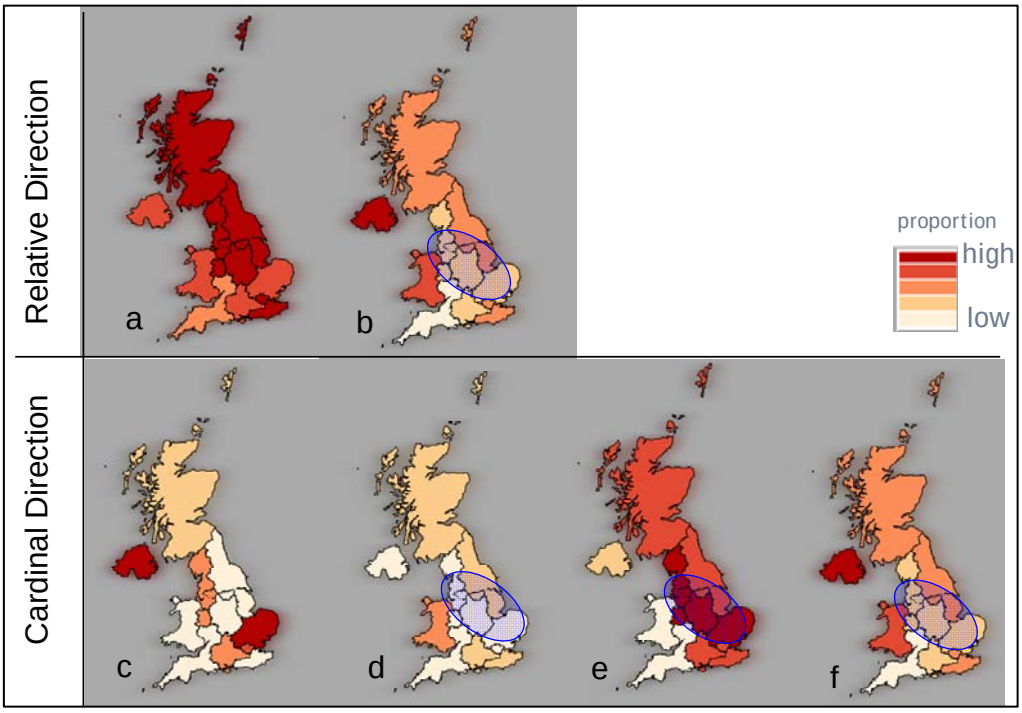

Figure 24. Region-level comparison of RD and CD usages in the U.K. Relative directions used as (a) “change of direction”, (b) “static spatial relationship”; Cardinal direction used as (c) “change of direction”, (d) “static spatial relationship”, (e) “traveling direction”, (f) “general origin”

For Australia, maps of region level comparisons in linguistic characteristic in route directions are shown in Figure 25. Although there is no distinctive region pattern in the Figure, the map shows regional differences. It is noteworthy that the proportions of cardinal direction representing *static spatial relationship* and *general origin* (Figure 25) are higher in Australia than in the U.S. and the U.K.

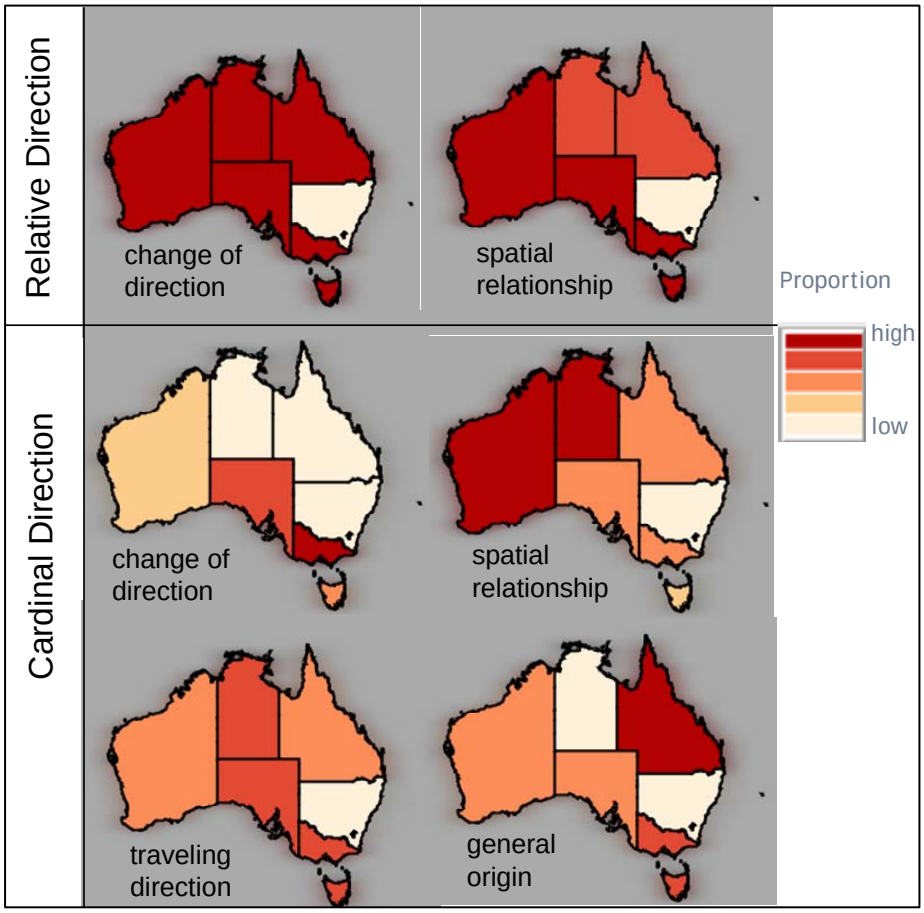

Figure 25. Region-level comparison of RD and CD usages in Australia

4.2.4 Result from K-means analysis

Based on the semantic categorical data used to characterize linguistic characteristics on the postal region level, it is possible to apply unsupervised cluster analysis on the linguistic attributes of each postal region. Cluster analysis calculates the distance of objects in projected attribute space and group objects with short distances into clusters. K-means [MacQueen, 1967] is one of the most popular cluster analysis methods. It can assist the interpretation by assigning regions into groups, rather than having analysts looking at numerical data. To apply K-means analysis, postal regions

are considered instances while the 7 semantic categorical data types (excluding cardinal direction *used in POI names*) are considered attributes to the instances. Cardinal direction *used in POI name* is related to place name conventions, which is a different topic than referencing frame difference in route directions. After calculating the clusters in the 7 dimension attribute space, postal regions with similar linguistic attributes are grouped together as a cluster. Mapping the cluster results offers another perspective to identify linguistic preference across different regions. K-means cluster analysis is applied to the U.S. data only because there are too few postal regions in the U.K. (13 postal regions) and Australia (8 states). For a higher number of K, that is, the number of clusters, the K-means for the U.K. and Australia would yield mostly each cluster with one postal region while each cluster in the U.S. would contain a higher number of postal regions, which is more sensible. After applying K-means algorithm, the results are mapped and shown in the Figure 26.

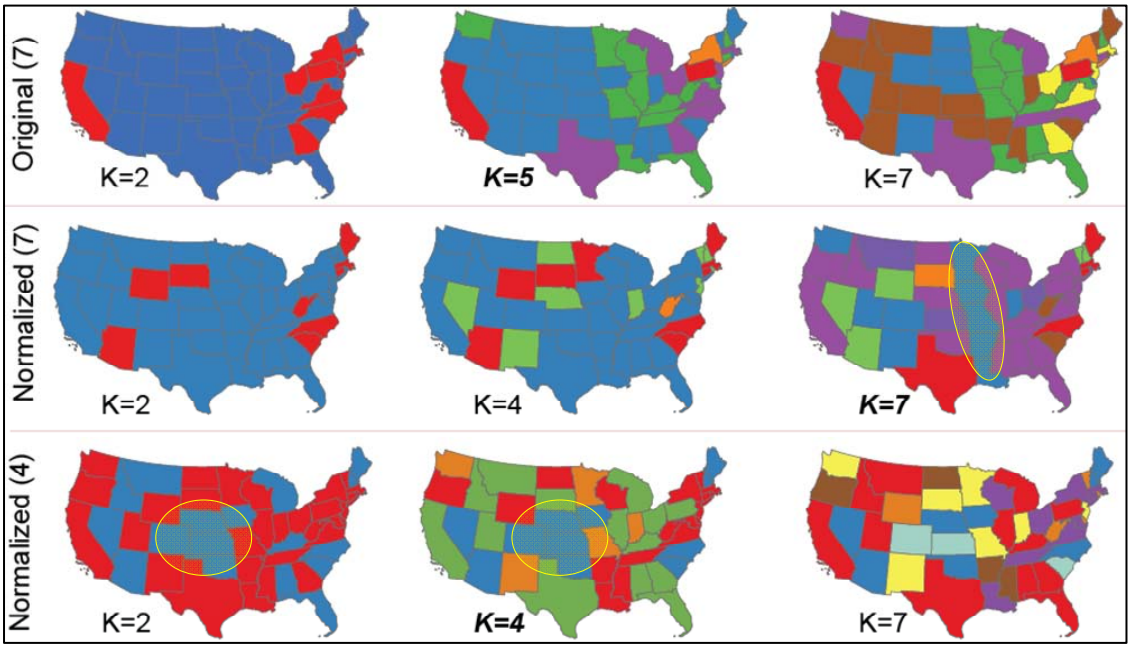

Figure 26. Mapping cluster analysis result using RD and CD usage in the U.S. with different K. Top row: original data; mid row: normalized data; bottom row: normalized data with only the 2 common semantic categories: "change of direction" and "static spatial relationship"

In Figure 26, three different sets of data are selected for carrying out the K-means analysis. The top row provides results from using the raw token frequency count of the seven semantic categories (see Section 4.2.2.1 and Appendix C, Table 11). The mid row shows results from using proportions of the seven semantic categories (see Section 4.2.2.1). The bottom row shows results from using proportions of the two shared semantic categories (*change of direction* and *static spatial relationship* in both relative directions and cardinal directions, that is, 4 regional linguistic characteristics, see Section 4.2.2.2). Also, three different numbers of K (number of cluster) are used to conduct K-means cluster analysis on different levels of granularity. The different choice

of data and K in this cluster analysis aims for providing grouping of regions from different perspectives.

Using the elbow method [Aldenderfer and Blashfield, 1984], the most suitable cluster number K was found to be 7 for the proportional linguistic data, 5 for the token occurrence data, and 4 for the proportional linguistic data using only the shared semantic categories (**bold** in Figure 26). Results of cluster analysis with K from 2 to 10 are mapped to find the clustering that is most suitable not only regarding numeric attributes, but also potential geospatial explanations. From the selected maps shown in Figure 26, it is noteworthy that we find a regional cluster along the west of the Mississippi river (from LA to MN, yellow circled in the mid right map). The region cluster in the Mideast (Figure 26 yellow circled in the bottom row K=2, K=4 maps) is also noted in Section 4.2.3 Figure 23. The explanation of the linguistic phenomenon may lie in the correlation between regional linguistic preferences and geographical features, which are discussed in Chapter 5. As a result of the analysis, regional linguistic patterns are shown to exist in cardinal/relative direction usages.

4.2.5 Result from Spatial Autocorrelation – Moran's I

The first law of geography states: "everything is related to everything else, but near things are more related than distant things" [Tobler, 1970, p. 236]. The regional patterns observed in the analysis support Tobler's law: using regional linguistic characteristics as attributes of a postal region, certain regions with similar attributes are adjacent to each other. To evaluate the significance of the observed pattern, spatial

autocorrelation is applied. Spatial autocorrelation is often used to evaluate if attributes at proximal locations are correlated, either positively or negatively. Spatial autocorrelation is a systematic pattern of a variable observed in its spatial distribution: positive autocorrelation corresponds to clusters; negative autocorrelation corresponds to a dispersed distribution. Random patterns exhibit no spatial autocorrelation. In other words, spatial autocorrelation can be used to test the assumption if the regional linguistic attributes are the results of an IRP/CSR (Independent Random Process/Complete Spatial Randomness). In this section, the spatial pattern using the regional linguistic characteristics in the U.S. is evaluated with a spatial autocorrelation measure—Moran's I. The U.S. is selected instead of the U.K and Australia because the number of postal regions in the latter two nations is too small for spatial autocorrelation analysis.

Moran's I [Moran, 1950] is a common spatial autocorrelation measure. A Moran's I value near +1.0 indicates positive spatial autocorrelation while a value near –1.0 indicates negative spatial autocorrelation, z-scores are used to indicate statistical significance. Proportion data are used to avoid the bias introduced by differences in the number of documents in a postal region. The software GeoDA [Anselin et al., 2010] is used in this section to conduct Moran's I spatial autocorrelation analysis. A weight matrix is created using Rook contiguity. First the semantic usages within each direction type are investigated. Moran's I values of the seven semantic categories in the U.S. are listed in Table 9.

Table 9. Spatial autocorrelation (Moran's I) result in the U.S. using 7 regional linguistic characteristics (proportion): For relative directions, RD_1: "representing change of directions", RD_2: "representing static spatial relationship", RD_3: "representing driving aid". For cardinal directions, CD_1: "representing traveling directions", CD_2: "representing change of directions", CD_3: "representing static spatial relationship", CD_4: "representing general origin", CD_5: used in POI names

Regional Linguistic Characteristics	RD_1	RD_2	RD_3	CD_1	CD_2	CD_3	CD_4
Moran's I Value	-0.23	-0.19	0.15	-0.08	0.16	0.24	0.11
Z-score	-1.8	-1.42	1.46	-0.48	1.57	2.27	1.16

From Table 9 we can see that Moran's I values for RD_1, RD_2 and CD_1 are close to 0, indicating the pattern is close to IRP/CSR for these regional linguistic characteristics. Moran's I values for RD_3, CD_2, CD_3 and CD_4 indicate positive spatial autocorrelation. However, only CD_3 has a z-score indicating a statistically significant pattern, while the z-scores for the other 6 regional linguistic characteristics are all too low to be considered statistically significant.

To evaluate the regional pattern from another perspective, Moran's I values using the semantic usages of the two shared semantic categories are presented in Table 10. The proportion is calculated in the same way as in the analysis in Section 4.2.1.2 and Section 4.2.2.2. For example, the proportion of relative direction representing *change of direction* equals the token occurrence count of relative direction used in this category divides the sum of the token occurrence count of both relative and cardinal direction, representing *change of directions*. It is noteworthy that the Moran's I value in the semantic category *change of directions* is 0.36 with a high z-score of 3.26.

The positive spatial autocorrelation is statistically significant, which means that when people are expressing change of directions, the regional distinction of preference in using relative directions or cardinal directions is significant. For the *static spatial relationship* categories, the z-score 1.86 (between 1.645 and 1.96) indicates that there is a between 5% and 10% probability that the pattern observed could be the result of a IRP/CSR process. The statistical significance level of Moran's I spatial autocorrelation analysis is much higher when compared between the shared semantic categories, indicating more confident regional variations.

Table 10. Spatial autocorrelation (Moran's I) result in the U.S. using 4 regional linguistic characteristics (proportion)

Regional Linguistic Characteristics	Change of direction		Static spatial relationship	
	RD	CD	RD	CD
Moran's I Value	0.36	0.36	0.2	0.2
Z-score	3.26	3.26	1.86	1.86

4.2.6 Summary of the Analysis Result

The case study of analyzing relative direction and cardinal direction usages for detecting regional variations in route direction language use shows the value of building the SARD Corpus. The identified semantic categories of relative directions and cardinal directions allow for analyzing regional differences within each direction type and between shared semantic categories. From histograms and Map visualization of the regional linguistic characteristics, interesting patterns are observed. Preference of relative directions in *change of directions* is common in the U.S., the U.K. and Australia

while variations exist on the regional level. K-means analysis and Moran's I spatial autocorrelation analysis allowed for a better interpretation of the regional variations. As a result of the many analysis methods applied, it is found that regional variations do exists in route direction language usage, and certain variations (for example, relative directions for representing *change of directions*) are statistically significant.

5 Discussion

Because of the exploratory nature of this study, there are interesting findings as well as space for improvement that are worth discussing. Deeper investigations by manually examining route direction documents from particular regions are presented in Section 5.1. The pros and cons of data collection and data analysis schemes are discussed in Section 5.2 and Section 5.3 respectively. Although the analysis in this thesis focuses on relative versus cardinal directions, several other potential analysis opportunities are noted during the process of conducting linguistic analysis of route directions. These potential analysis directions are discussed as future work in Section 5.4. Related work to this thesis and an overview of the GeoCAM project (which this thesis is part of) is discussed in Section 5.5.

5.1 Interpretation of the Analysis Results on Cardinal vs. Relative Direction Usages

Although regional variations are detected as a result of many analysis methods applied to regional linguistic characteristics in this thesis, providing reasons for the variation requires looking into the route direction documents in regions that have interesting patterns. For example, the abrupt difference in *static spatial relationship* usages noted in Figure 18 (comparing AR with 90% cardinal direction proportion and ME with 98% relative direction proportion). To find out the underlying reason of the

spiky shift, examination of documents within each postal region individually has been carried out. Because the occurrence was not normalized per document or per route description, it is possible that one frequently occurring usage in one document may skew the analysis. Take the region AR for example, the high occurrence of cardinal directions representing static spatial relationship is partially due to the fact that there is one long document that contains a lot of cardinal directions used for representing static spatial relationship. There are 20 occurrences in one document, in forms of "*Eight miles south of Fayetteville*", "*approximately 17 miles south of Paris*[4] *or from Danville*". Although this is a bias that should be avoided in future analysis, taking the skewed result out still will not affect the result of the analysis (total occurrence become 251 instead of 271, the proportion of cardinal direction in static spatial relationship usage changes from 89% to 88%). Hence the proposed result from the case study is still valid for identifying regional linguistic variations. However, from this example, token occurrence per route direction appears to be a more sensible choice for linguistic characteristics. Because the route direction documents in the SARD Corpus are collected from the WWW, even though they contain route directions, other types of text still exists in them (for example, addresses or advertisements). The challenges in analyzing token occurrence per route direction would require a sentence-level classifier which could separate the unrelated text from route directions, then divide multiple route

[4] A small town in Arkansas which shares the same name as the Capital of France.

directions into single ones. This challenge can be overcome by text classification, which is discussed further in Section 5.4 future work.

From investigating original route direction documents in the SARD Corpus, the following usage is worth noticing:

Turn *right* (*south*) at S 1st E and drive ….

Cardinal direction and relative direction can be used together in provide additional information in route directions. Providing additional information is a typical feature in human generated route directions. Take the use of landmarks as an example: humans usually include landmark information in route directions, which provides a way for the traveler to make sure he or she is making turns at the correct decision points or is traveling on the correct route. The use of both relative and cardinal directions has a similar function. It indicates that the route describer is aware of the difference in using these two types of directional terms and using them both can complement each other in providing the reader route direction information.

5.1 Pros and Cons of the Data Collection Scheme

The 3 steps data collection scheme (crawl, classify, geo-reference) using postal code databases has advantages of collecting spatially distributed data with low spatial ambiguity; However, there are several pros and cons that are worth mentioning when comparing this data collection scheme and other candidates. The following three aspects

of the data collection scheme will be discussed: regional density bias and its reasons, ambiguous route direction document handling, and the sacrifice of certain types of useful route direction documents.

First, after building the SARD Corpus with the proposed data collection scheme, there is a clear regional density bias in the postal code location where each route direction document comes from (see Section 4.1). The density bias comes from the data collection scheme: for each postal code queried, the first 20 results are retrieved and then classified. This results in a varying number of route direction documents in each postal code. A possible modification to partly prevent such regional density bias is to set a fixed number of route direction documents for each postal code to be crawled—there are still postal codes that do not return any route direction documents, so the density bias still exists in these regions. However, there are several reasons the former scheme was chosen rather than the latter: first, from my experience in processing route direction documents, route directions are more likely to be generated in highly populated areas. The high density along the eastern and western coast line represents this phenomenon well. The question becomes: what should be favored, representativeness of regions or representativeness of usages? If forcefully retrieving the same number of route direction documents from each postal code, it would be unfair to the higher populated regions as route directions are used more frequently in these regions. For example, in an area with low population, 50 hits were classified resulting in five route direction documents. On the contrary, in a highly populated area, the first ten hits sometimes already provide five route direction documents. The forced equation is unfair to the latter area since it may

need a much larger number of route direction documents to obtain representative linguistic features. Comparatively, the data collection scheme carried out in this thesis provides a relatively balanced way to obtain representativeness of overall route direction usages. In turn, size of the document set in regions with low populations (examples are Wales, Exeter, Northern Ireland and other regions in the U.K., refer to Figure 13) are sacrificed. A remedy for preventing low numbers of documents is a two way approach: setting of threshold number of documents per postal code, if a postal code returns more than the threshold number of documents after the first 20 hits, stop; otherwise, continue crawl until the threshold number of documents is reached or all returned documents are classified. This modification will be applied for future research.

Second, machine-generated route directions by Natural Language Generation (NLG) systems and human-generated route directions can both be found on the WWW, of which we are only interested in the latter. The machine-generated route directions have a rigid style, which sometimes is in contrast to human-generated ones. For example, a lot of usage in precise travel distance (e.g., *drive for 10.2 miles*) can be found in machine-generated route directions while distance measurements are not frequently found in human-generated route directions. In human-generated route directions, vague distances (drive approximately 2 miles) are used instead of overly exact distance measurements. Most map service providers, such as MapQuest, Google Maps, Yahoo! Maps, do not apply landmarks in their machine-generated route directions. On the contrary, landmarks are used frequently in human-generated route directions (for example, “after the traffic light you will see a church on your right”).

Without ensuring the corpus is built with only human-generated text, any analysis built on that has little credibility. To prevent the machine-generated language from causing a major effect on the analysis, all documents used in the training set are selected human-generated route directions. Examination of the final SARD Corpus shows that only a small amount 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 result. From another perspective, it is hard to say that the human-generated route directions used online nowadays are not affected by the machine-generated style at all. Some route directions, although generated by a human (with landmarks), clearly borrow some machine-generated distance measures and combine them together with describer's own text. On the other hand, the on-line map service providers are improving their technology in generating route directions in a more human-like way as well. Bing Maps, for example, has started to provide landmark information in the machine generated route directions, in form of "Turn left onto Colonnade Blvd -- 0.3 miles. OUTBACK STEAKHOUSE on the corner". There is some other additional information (such as "The last intersection is Waddle Rd. If you reach Theatre Dr, you've gone too far") that is learned from the style in human generated route directions. It is evident that the spatial database of on-line map service providers are evolving together with the Natural Language Generation systems for route directions.

Third, the data collection scheme sacrifices certain types of route direction documents to ensure the purity of the resulting corpus. Admittedly, not all route direction documents contain postal codes or addresses. Without a mature automated

geo-referencing scheme with high precision, route direction documents without addresses or postal codes are very hard to plot with the correct location. The proposed data collection scheme ignores such documents in order to avoid the geographic name entity recognition problem and deliver correct location information of the documents. Additionally, through the data classification and location validation process, a lot of route direction documents are thrown away (e.g., 1403 route direction documents from the U.S. are excluded from the SARD Corpus) because they contain multiple postal codes. Although it may seem to be overkill, for example, a route direction document with a detailed address of an unrelated company in another state would be thrown out by this process—it is a simply step to ensure that the documents being analyzed are unambiguous with regard to regions. A more sensible, yet computationally complicated way to address this issue is to develop a sentence-level text classifier and route-associater to sort out the unwanted text in the route direction documents and organize the route directions in sets. The sentence-level text classification and geo-referencing is a necessity for regional linguistic analysis on a finer scale, which is being developed and tested (refer to Section 5.4).

Similar data collection ideas on different topics in the following studies are worth mentioning. Work by [Jones et al., 2008] shares some characteristics with the data collection scheme introduced in this thesis. They used queries such as "hotels in <placename>" to harvest geo-referenceable information from documents on the WWW. [Tomai and Kavouras, 2004] utilizes text documents from the WWW to reveal geospatial knowledge by analyzing the semantics of spatial languages. This trend of

utilizing various kinds of volunteered or pseudo-volunteered spatial data from the WWW shows the power of crowd sourcing, which may subvert (in a good way) the current data collection method of many research fields.

5.2 Pros and Cons of the Data Analysis Scheme

The case study of investigating regional variation in relative direction and cardinal direction usages offers a preliminary take on designing an analysis scheme on a regional level. The semantic categories are designed to capture the linguistic characteristics of the two direction types, while the text processing tool (TermTree Tool) and the visual analytic tool (VIT) are used to assist the interpretation. There are several important factors to consider in designing the analysis scheme which are discussed below.

For analyzing the SARD Corpus, token occurrence rather than token frequency (for example, occurrence per million words) is chosen as a linguistic measure because of the nature of documents from the WWW. All documents in the corpus are original web pages where the route directions are found. As a common webpage design, they may include headers, advertisements and other components that are not related to route directions. These unrelated texts 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 influences from the unrelated text. Given a sentence level route direction classifier, token frequency would become another

potential linguistic measure. Other potential measure could also be chosen for the regional linguistic characteristics, such as token occurrence per document and token occurrence per route direction. Token occurrence in all documents from a region is chosen for its reliability and to shorten the time for examination of semantic categorical usages.

The scale of the regional linguistic analysis is set to postal regions and has its pros and cons as well. First, most route directions, even though they have a destination that can be refined to a postal code-level, describe routes with greater coverage than postal code-level. It is safer and more reasonable to carry out the analysis on postal region-level because the scales of most route directions are smaller than a postal region (e.g., state). Second, the number of postal regions is relatively smaller than the number of regions of a finer granularity (e.g., postal code area or county). As the analysis scheme requires analysts to calculate the token frequency of each semantic category in every region, a lower number of regions shortens the analysis process—considering the size of the SARD Corpus, conducting linguistic analysis over 10,000 documents on a postal code scale (41,119, 8860 and 3312 postal codes from the U.S., the U.K, and Australia respectively) would result in a much longer analysis period. However, postal regions (e.g., states in the U.S.) are divided by political or cultural boundaries, which may not offer the best aggregation scheme for analyzing linguistic characteristics. Attempts to define linguistic regions can be found in existing researches, for example, [Zelinsky, 1955] uses population of towns and “frontiers of settlement” [Zelinsky, 1955, p.347] for investigating place-term distributions. As a preliminary case study, postal

regions are used for the regional analysis in this thesis. However, various regional aggregations could be applied and offer insight into regional variations from different perspectives.

5.3 Potential Analysis Opportunities and Future Work

Besides regional variations in referencing frame preference as analyzed in this thesis, there are several other potential analysis opportunities. Some of these potential analysis opportunities are discussed below.

As previously stated, landmarks are a set of geo-referenceable features in route direction documents and appear frequently in human-generated route directions. Considering different route environments, the type of landmarks used may vary. For example, buildings may be the most effective type of landmarks in local-scale (e.g., in metropolitan environments such as New York City) environments but they may be less useful when describing region-scale routes (e.g., traveling on interstate highways). Landmark classification has been studied thoroughly in [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 (for example, traffic lights, buildings), line-like (for example, roads, streets), area-like (for example, parks, campuses) and others. The occurrence frequency of landmark categories in local-scale versus region-scale route directions document sets can be

compared to provide an understanding of how humans use landmarks across various environments.

Travel distance and time are important elements for representing how far and how long a traveller should expect to travel during a specific travel period, which may be written in different ways based on mode-of-transportation. For example, the use of travel distance (for example, “drive for approximately 6 miles on US 322”) is appropriate for automobile travelling but it is less preferred than travel time when referring to Mass Transit (for example, “The travel time is 20-30 minutes, longer during peak hours”). Moreover, the importance and effectiveness of these elements may vary as well (for example, while a digital map provides route descriptions using travel distance for automobile travel, a traveller may prefer to use landmarks or other indicators, for example, “take a left after the hospital”, to describe routes). Analysis of the frequency of travel distance and time in each mode-of-transportation on large spatial language corpora will provide a more concrete idea on how humans use them.

The above two analysis directions can be illustrated in a data collection and data analysis process together with regional variations of relative direction and cardinal direction usage. The whole process is illustrated in Figure 27.

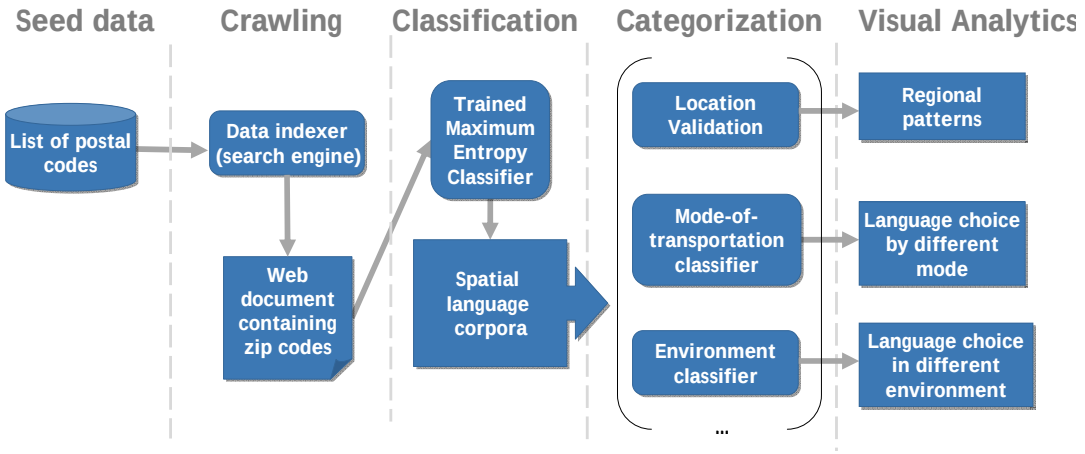

Figure 27. Data collection and analysis schemes for analysis regional, environmental, and transportational differences in route directions

5.4 The Big Picture – GeoCAM Project

This study is part of the GeoCAM (Geographic Contextualization for Accounts of Movements) project. Given a huge amount of text data, a system that is capable of extracting, visualizing, and analyzing the linguistically encoded movements is the goal of the GeoCAM project. Computer linguistics methods are integrated with geographic databases to build functional components for the GeoCAM system, including Route Document Classifier [Zhang et al., 2009], Sentence Classifier [Zhang et al., 2009], text analyzing tool [TermTree Tool: Turton & MacEachren, 2008], and Route Sketcher. The whole system is being developed to enable spatial analysis of linguistically encoded movement patterns at various scales.

The study in this thesis, from data collection to data analysis, utilizes the tools developed in GeoCAM to assess regional variation in cardinal and relative direction

usage within route directions. I participated in the process of developing the route direction classifier on both a document and a sentence level (only document-level route direction classifier is used in this thesis) with regard to building the training set, evaluating precision of different text classification models and applying the route direction document classifier recursively to build the SARD Corpus. The postal code based crawling scheme for building the SARD Corpus introduced in this thesis is developed by me. The SARD Corpus as a route direction document repository has been used in test cases for testing RouteSketcher, which is an on-line geovisualization tool for route directions in text. The study in this thesis interacts with other components in the GeoCAM project and becomes inseparable. The RouteSketcher, which offers geovisualization of geo-referenced text route directions on the sentence-level at a finer spatial granularity, will be used for future analysis of regional linguistic characteristics.

5.5 Summary

The pros and cons of the data collection and analysis scheme as well as potential future works are discussed in this chapter., Although there are drawbacks regarding the data collection schema, such as excluding route directions without postal codes, using postal codes to locate route direction document and organizing route direction documents in postal regions is shown effective for regional analysis on linguistic characteristics. In the analysis results, some noted abrupt differences are investigated by looking at the original route documents, which poses new challenges for analyzing the regional linguistic characteristics. Besides token occurrence, other potential measure for

regional linguistic characteristics is discussed and challenges are put forward. The data collection and data analysis schemes presented are discussed to be effective for detecting regional variations in relative and cardinal direction usages.

Further potential analysis opportunities on spatial language usage in route directions, such environmental and transportation related differences, are introduced and analysis schemes have been designed. These potential research opportunities offers multiple perspectives for understanding route direction usages. The GeoCAM project, which this thesis is part of, aims at a much more challenging goal: visualizing linguistically characterized routes automatically in a visual analytics workbench and processing text information of route directions on various spatial scales. Geographic name entity recognition schemes and sentence-level route direction classifier have been designed, offering additional potentials for analysis.

6 Conclusions

The research question of regional variation in route direction usage within English is addressed in this thesis. Important insights into route direction usages across the U.S., the U.K. and Australia have been provided. On a national level, relative directions are preferred to represent *change of directions*, while cardinal directions are preferred to represent *traveling directions*. Regional patterns using linguistic characteristics have been observed. Spatial autocorrelation analysis using global Moran's I demonstrates that some of the regional patterns observed (such as difference in cardinal vs. relative directions in representing *change of direction*) are statistically significant. The primary findings of this study—that linguistic preferences across regions exist in spatial language—not only add to the growing literature on spatial language, but also have practical implications. In other words, exploring regional spatial linguistic preference may offer insight to improve wayfinding performance by creating route directions with fitting regional language usages.

This study also offers a methodological contribution: a paradigm to build a geo-referenced corpus from documents on the WWW. The proposed workflow provides an inexpensive, fast, and reliable way to acquire a spatially distributed corpus with innovative schemes and computational tools to conduct text analysis. This methodology in data collection and data analysis can be extended for text analysis research on other topics as well. Potential analysis opportunities on route directions with regard to

landmark usages and travel distance and time usages are designed for future works. The contributions in both data collection and data analysis schemes introduced in this thesis advance evaluation possibilities for linguistic patterns existing within the same language describing movement patterns across different regions.

Reference

[Aldenderfer and Blashfield, 1984] Aldenderfer, M. S. and Blashfield, R. K. (1984). *Quantitative Applications in the Social Sciences 44: Cluster Analysis*, volume 31. Sage Publications (CA), Beverly Hills.

[Allen, 1997] Allen, G. (1997). From knowledge to words to wayfinding: Issues in the production and comprehension of route directions. In Hirtle, S. C. and Frank, A. U., editors, *Spatial Information Theory A Theoretical Basis for GIS*, volume LNCS 1329, pages 363–372, Laurel Highlands, Pennsylvania, USA. Springer.

[Ambati et al., 2010] Ambati, V., Vogel, S., and Carbonell, J. (2010). Active learning and crowd-sourcing for machine translation. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

[Anselin et al., 2010] Anselin, L., Syabri, I., and Kho, Y. (2010). GeoDa: An introduction to spatial data analysis. In *Handbook of Applied Spatial Analysis*, pages 73–89. Springer Berlin Heidelberg.

[Anthony, 2006] Anthony, L. (2006). Concordancing with antconc: An introduction to tools and techniques in corpus linguistics. In *Proceedings of the JACET 45th Annual Convention*, volume 155, pages 218–219.

[AustraliaPost, 2009] AustraliaPost (2009). postal code for Australia. World Wide Web electronic publication. Australia Post.

[Banko and Brill, 2001] Banko, M. and Brill, E. (2001). Scaling to very very large corpora for natural language disambiguation. In *ACL '01: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 26–33, Morristown, NJ, USA. Association for Computational Linguistics.

[Bateman et al., 2007] Bateman, J., Tenbrink, T., and Farrar, S. (2007). The role of conceptual and linguistic ontologies in interpreting spatial discourse. *Discourse Processes*, 44(3):175 – 212.

[Beesley, 1988] Beesley, K. (1988). Language identifier: A computer program for automatic natural-language identification of on-line text. In *Language at Crossroads: Proceedings of the 29th Annual Conference of the American Translators Association*, pages 12–16.

[Biber, 2006] Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. John Benjamins Pub Co.

[Burenhult and Levinson, 2008] Burenhult, N. and Levinson, S. C. (2008). Language and landscape: a cross-linguistic perspective. *Language Sciences*, 30(2-3):135 – 150. Language and landscape: geographical ontology in cross-linguistic perspective.

[Chakrabarti et al., 1999] Chakrabarti, S., Berg, M., and Dom, B. (1999). Focused crawling: A new approach to topic- specific web resource discovery. In *Computer Networks*, volume 31, pages 1623–1640.

[Chen et al., 2007] Chen, J., MacEachren, A. M., and Guo, D. (2007). Visual inquiry toolkit - an integrated approach for exploring and interpreting space-time, multivariate patterns. Technical report, GeoVista Center and Department of Geography Pennsylvania State University, Department of Geography University of South Carolina.

[Chrisment et al., 2004] Chrisment, C., Dousset, B., Karouach, S., and Mothe, J. (2004). Information mining: extracting, exploring and visualising geo-referenced information. In *Workshop on Geographic Information Retrieval, SIGIR*.

[Dabbs et al., 1998] Dabbs, J. M., Chang, E.-L., Strong, R. A., and Milun, R. (1998). Spatial ability, navigation strategy, and geographic knowledge among men and women. *Evolution and Human Behavior*, 19(2):89–98.

[Daniel and Denis, 1998] Daniel, M.-P. and Denis, M. (1998). Spatial descriptions as navigational aids: a cognitive analysis of route directions. *Kognitionswissenschaft*, 7:45–52.

[Davies and Pederson, 2001] Davies, C. and Pederson, E. (2001). Grid patterns and cultural expectations in urban wayfinding. In Montello, D., editor, *Spatial Information Theory*, volume LNCS 2205, pages 400–414, Morro Bay, CA, USA. Springer.
[Dumais et al., 2002] Dumais, S., Banko, M., Brill, E., Lin, J., and Ng, A. (2002). Web question answering: is more always better? In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 291–298, New York, NY, USA. ACM.
[freethepostcode.org, 2009] freethepostcode.org (2009). postal code for UK. World Wide Web electronic publication. freethepostcode.org.
[Gladwin, 1970] Gladwin, T. (1970). *East Is a Big Bird: Navigation and Logic on Puluwat Atoll*. Cambridge, MA: Harvard University Press.
[Goker and Davies, 2009] Goker, A. and Davies, J. (2009). *Information Retrieval: Searching in the 21st Century*. John Wiley and Sons.
[Goodchild, 2007] Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4):211–221.
[Goodchild and Glennon, 2010] Goodchild, M. F. and Glennon, J. A. (2010). Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth*, 1:1753–8947.
[Gumperz and Levinson, 1996] Gumperz, J. J. and Levinson, S. C. (1996). *Rethinking linguistic relativity Studies in the Social and Cultural Foundations of Language*. Cambridge: Cambridge University Press.
[Hansen et al., 2006] Hansen, S., Richter, K.-F., and Klippel, A. (2006). Landmarks in OpenLS - a data structure for cognitive ergonomic route directions. In *Geographic, Information Science*, volume 4197 of *Lecture Notes in Computer Science*, pages 128–144. Springer Berlin / Heidelberg.
[Haugen, 1957] Haugen, E. (1957). The semantics of icelandic orientation. *Cognitive anthropology*, pages 330–42.
[Hearst, 1999] Hearst, M. A. (1999). Untangling text data mining. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 3–10, Morristown, NJ, USA. Association for Computational Linguistics.
[Heer and Bostock, 2010] Heer, J. and Bostock, M. (2010). Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *CHI '10: Proceedings of the 28th international conference on Human factors in computing systems*, pages 203–212, New York, NY, USA. ACM.
[Herskovits, 1986] Herskovits, A. (1986). *Language and Spatial Cognition - An Interdisciplinary Study of the Propositions in English*. Stiudies in Natural Language Processing. Cambridge UK: Cambridge University Press.
[Howe, 2006] Howe, J. (2006). The rise of crowdsourcing. *The Wired Magazine*, 14.06.
[Howe, 2008] Howe, J. (2008). *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*. Crown Business.
[Hudson-Smith et al., 2009a] Hudson-Smith, A., Batty, M., Crooks, A., and Milton, R. (2009a). Mapping for the masses: Accessing web 2.0 through crowdsourcing. *Social Science Computer Review*, 27(4):524–538.
[Hudson-Smith et al., 2009b] Hudson-Smith, A., Milton, R., Crooks, A., and Batty, M. (2009b). Crowd sourced data for the social sciences: Web based services and real-time geographic surveys. In *5th International Conference on e-Social Science*, Cologne, Germany.
[Ishikawa and Kiyomoto, 2008] Ishikawa, T. and Kiyomoto, M. (2008). Turn to the left or to the west: verbal navigational directions in relative and absolute frames of reference. In Cova, T. J., Miller, H. J., Beard, K., Frank, A. U., and Goodchild, M. F., editors, *Geographic Information Science*, volume LNCS 5266, pages 119–132, Park City, UT, USA. Springer.
[Jackendoff, 1983] Jackendoff, R. (1983). *Semantics and cognition*. MIT press.
[Jackendoff and Landau, 1991] Jackendoff, R. and Landau, B. (1991). Spatial language and spatial cognition. In Napoli, D. J. and Kegl, J. A., editors, *Bridges between psychology and linguistics: a swarthmore festschrift for Lila Gleitman*, pages 145–169. Lawrence Earlbaum Associates, Inc., Publishers.

[Jaiswal, 2010] Jaiswal, A. R. (2010). Automatic semantic tagging of transportation mode for digital documents. Technical report, College of Information Science and Technology, Pennsylvania State Unversity.

[Jones et al., 2008] Jones, C., Purves, R., Clough, P., and Joho, H. (2008). Modelling vague places with knowledge from the web. *International Journal of Geographical Information Science*, 22(10):1045–1066.

[Kilgarriff and Grefenstette, 2003] Kilgarriff, A. and Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Computational Linguistics*, 29(3):333–347.

[Klein and Levelt, 1978] Klein, W. and Levelt, W. J. M. (1978). Sprache und kontext. *Naturwissenschaften*, 65(6):328–335.

[Kobayashi and Takeda, 2000] Kobayashi, M. and Takeda, K. (2000). Information retrieval on the web. *ACM Comput. Surv.*, 32(2):144–173.

[Labov et al., 2006] Labov, W., Ash, S., and Boberg, C. (2006). *The atlas of North American English: phonetics, phonology, and sound change: a multimedia reference tool*. Walter De Gruyter Inc.

[Lamarre, 2008] Lamarre, C. (2008). The linguistic categorization of deictic direction in chinese - with reference to japanese. In *Space in Languages of China*, pages 69–97. Springer Netherlands.

[Landau, 1998] Landau, B. (1998). Spatial cognition and spatial language: What do we need to know to talk about space? AAAI Technical Report WS-98-06 AAAI.

[Lawton, 2001] Lawton, C. A. (2001). Gender and regional differences in spatial referents used in direction giving. *Sex Roles*, 44(5-6):321–337.

[Lee and Kretzschmar, 1993] Lee, J. and Kretzschmar, W. A. (1993). Spatial analysis of linguistic data with gis functions. *International Journal of Geographical Information Systems*, 7(6):541–560.

[Lee and Lee, 2007] Lee, S. and Lee, G. G. (2007). Exploring phrasal context and error correction heuristics in bootstrapping for geographic named entity annotation. *Information Systems*, 32(4):575–592.

[Lewis, 1972] Lewis, D. (1972). *We, the navigators*. Hawaii: University Press of Hawaii.

[Lewis, 1976] Lewis, D. (1976). Observations on route finding and spatial orientation among the aboriginal peoples of the western desert region of central australia. *Oceania*, 46:249–282.

[Li et al., 2003] Li, H., Srihari, R., Niu, C., and Li, W. (2003). Infoxtract location normalization: a hybrid approach to geographic references in information extraction. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references-Volume 1*, pages 39–44. Association for Computational Linguistics Morristown, NJ, USA.

[List_of_crowdsourcing_projects,] List_of_crowdsourcing_projects. http://en.wikipedia.org/wiki/list_of_crowdsourcing_projects.

[LIU, 2008] LIU, D. (2008). Syntax of space across chinese dialects: Conspiring and competing principles and factors. In *Space in Languages of China*, pages 39–67. Springer Netherlands.

[Lovelace et al., 1999] Lovelace, K. L., Hegarty, M., and Montello, D. R. (1999). Elements of good route directions in familiar and unfamiliar environments. In Freksa, C. and Mark, D. M., editors, *Spatial Information Theory - Cognitive and Computational Foundations of Geographic Information Science*, volume LNCS 1661, pages 65–82. Springer Berlin.

[Luhn, 1957] Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM J. Res. Dev.*, 1(4):309–317.

[MacMahon et al., 2006] MacMahon, M., Stankiewicz, B., and Kuipers, B. (2006). Walk the talk: Connecting language, knowledge, and action in route instructions. In *Twenty-First National Conference on Artificial Intelligence (AAAI-06)*, pages 1475–1482.

[MacQueen, 1967] MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1:281–297.

[Mark and Gould, 1992] Mark, D. and Gould, M. (1992). Wayfinding as discourse: A comparison of verbal directions in english and spanish. *Multilingua*, 11(3):267–291.

[Mark and Egenhofer, 1995] Mark, D. M. and Egenhofer, M. J. (1995). Topology of prototypical spatial relations between lines and regions in english and spanish. In *Proceedings of the Twelfth International Symposium on Computer-Assisted Cartography*, pages 245–254, Charlotte, North Carolina.

[Mark and Frank, 1989] Mark, D. M. and Frank, A. U. (1989). Concepts of space and spatial language. In *Proceedings of the 9th International Symposium on Computer-Assisted Cartography*, pages 538–556, Baltimore, MD.

[Martins et al., 2006] Martins, B., Silva, M., Freitas, S., and Afonso, A. (2006). Handling locations in search engine queries. In *Workshop on Geographical Information Retrieval, SIGIR'06*, Seattle, Washington. ACM.

[McCallum, 2002] McCallum, A. K. (2002). Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

[McConchie, 2002] McConchie, A. (2002). The pop vs. soda map. http://popvssoda.com:2998/.

[Montello, 1993] Montello, D. R. (1993). Scale and multiple psychologies of space. In Frank, A. U. and Campari, I., editors, *Spatial Information Theory: A Theoretical Basis for GIS*, volume LNCS 716, pages 312–321, Marciana Marina, Elba Island, Italy. Springer.

[Montello, 2001] Montello, D. R. (2001). Spatial cognition. In Smelser, N. J. and Balters, P. B., editors, *International Encyclopedia of the Social & Behavioral Sciences*, pages 14771–14775. Oxford: Pergamon Press.

[Montello et al., 1999] Montello, D. R., Lovelace, K. L., Golledge, R. G., and Self, C. M. (1999). Sex-related differences and similarities in geographic and environmental spatial abilities. *Annals of the Association of American Geographers*, 89(3):515–534.

[Mooney and Roy, 2000] Mooney, R. J. and Roy, L. (2000). Content-based book recommending using learning for text categorization. In *DL '00: Proceedings of the fifth ACM conference on Digital libraries*, pages 195–204, New York, NY, USA. ACM.

[Moran, 1950] Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37:17–33.

[Munnich et al., 2001] Munnich, E., Landau, B., and Dosher, B. A. (2001). Spatial language and spatial representation: a cross-linguistic comparison. *Cognition*, 81(3):171 – 208.

[Nigam et al., 1999] Nigam, K., Lafferty, J., and McCallum, A. (1999). Using maximum entropy for text classification. In *In IJCAI-99 Workshop on Machine Learning for Information Filtering*, pages 61–67.

[Retz-Schmidt, 1987] Retz-Schmidt, G. (1987). Deictic and intrinsic use of spatial prepositions: A multidisciplinary comparison. In Kak, A. C. and shing Chen, S., editors, *Spatial Reasoning and Multi-Sensor Fusion: proceedings of the 1987 workshop*, pages 371–380, Pleasan Run Resort, St. Charles, IL. Morgan Kaufmann Publishers.

[Richter, 2005] Richter, K. (2005). Route direction structure diagrams. In v. Nes, A., editor, *Proceedings of 5th International Space Syntax Symposium*, TU Delft, Netherlands.

[Richter et al., 2008] Richter, K., Tomko, M., and Winter, S. (2008). A dialog-driven process of generating route directions. *Computers, Environment and Urban Systems*, 32(3):233–245.

[Richter, 2008] Richter, K.-F. (2008). Context-specific route directions: Generation of cognitively motivated wayfinding instructions. In Barkowsky, T., Freksa, C., Holscher, C., Krieg-Bruckner, B., and Nebel, B., editors, *the Monograph Series of the Transregional Collaborative Research Center SFB/TR 8*, volume 3.

[Sebastiani, 2002] Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47.

[Smith, 2006] Smith, J. (2006). *Bum bags and fanny packs: a British-American, American-British dictionary*. Carroll & Graf Publishers.

[Tobler, 1970] Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46:234–240.

[Tomai and Kavouras, 2004] Tomai, E. and Kavouras, M. (2004). "where the city sits?" revealing geospatial semantics in text descriptions. In *7th AGILE Conference on Geographic Information Science*, pages 189–194, Heraklion, Greece. Association of Geographic Information Laboratories for Europe.

[Turney, 2002] Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 417–424, Morristown, NJ, USA. Association for Computational Linguistics.

[Turton, 2008] Turton, I. (2008). A system for the automatic comparison of machine and human geocoded documents. In *GIR '08: Proceeding of the 2nd international workshop on Geographic information retrieval*, pages 23–24, New York, NY, USA. ACM.

[Turton and MacEachren, 2008] Turton, I. and MacEachren, A. (2008). Visualizing unstructured text documents using trees and maps. In *GIScience workshop*, Park City, Utah.

[Vanetti and Allen, 1988] Vanetti, E. J. and Allen, G. L. (1988). Communicating environmental knowledge: The impact of verbal and spatial abilities on the production and comprehension of route directions. *Environment and Behavior*, 20(6):667–682.

[Ward et al., 1986] Ward, S. L., Newcombe, N., and Overton, W. F. (1986). Turn left at the church, or three miles north: A study of directions giving and sex differences. *Environment and Behavior*, 18(2):192–213.

[Wise et al., 1995] Wise, J. A., Thomas, J. J., Pennock, K., Lantrip, D., Pottier, M., Schur, A., and Crow, V. (1995). Visualizing the non-visual: spatial analysis and interaction with information from text documents. In *INFOVIS '95: Proceedings of the 1995 IEEE Symposium on Information Visualization*, page 51, Washington, DC, USA. IEEE Computer Society.

[Yahoo, 2009] Yahoo (2009). Yahoo! http://search.yahoo.com/search?p=a.

[Yao and Yao, 2003] Yao, J. and Yao, Y. (2003). Web-based information retrieval support systems: Building research tools for scientists in the new information age. *Web Intelligence, IEEE / WIC / ACM International Conference on*, 0:570.

[Zelinsky, 1955] Zelinsky, W. (1955). Some problems in the distribution of generic terms in the place-names of the northeastern united states. *Annals of the Association of American Geographers*, 45:319–349.

[Zhang et al., 2009] Zhang, X., Mitra, P., Xu, S., Jaiswal, A. R., Klippel, A., and MacEachren, A. M. (2009). Extracting route directions from web pages. In *Twelfth International Workshop on the Web and Databases (WebDB 2009)*, Providence, Rhode Island, USA.

[zipcodeworld.com, 2009]zipcodeworld.com (2009). postal code for the continental U.S. World Wide Web electronic publication. http://www.zipcodeworld.com/zipcodefree.htm.

Appendix A

Pseudo code for the web crawling/classifying process for building the SARD Corpus (refer to Section 3.1):

```
webCrawl_SARDCorpus {

String[] query;     //result of reading the line containing State abbreviation and ZIPcode in the form of "PA
        16803”

Int num_Crawl;   //number of document to be crawled per postal code

File DownloadDirectory; //for storing the returned documents from the Yahoo!.com

File RouteDocDirectory; //for storing the classified positive route direction documents

File listOfPostalCodes;     //stored list of postal codes, organized with one postal code per line

Begin

        Create DownloadDirectory;

        Create RouteDocDirectory;


//Load Training Features from Disk

        trainList = InstanceList.load(training features);

//Build Maximum Entropy Classifier (Initialize classifier before loops else will cause significant
computation time increase)

maxEntTrainer= new DirectionMaxEntTrainer(trainList);

        maxEntTrainer.train();

        RouteDocClassifier = maxEntTrainer.getClassifier();

//Finished Training Classified



While query = readline from listOfPostalCodes)!=null

        Begin_crawl_query

                create Yahoo! returned result URL according to query.
```

Store Yahoo! return result page;

If Yahoo! returns 999 error, pause (random interval between 10 minutes to 30 minutes to simulate human searcher behavior) and continue;

Extract num_Crawl URL from of result from the Yahoo! result page;

Store file to DownloadDirectory;

//start classification. positive classified route direction documents to RoutDocDirectory, delete negative classified result

Begin_classify_DownloadDirectory;

int i=0;

repeat

targetFile=DownloadDirectory.listfile[i];

if (targetFile.ClassLable==1) //when the target document is classified as route direction document

Move(targetFile, RouteDocDirectory)

i=i+1;

Else

Delete(targetFile);

Until (DownloadDirectory.size=0)

End_classify_DownloadDirectory

End_crawl_query

End

Appendix B

List of postal region in the U.K., derived from

http://www.summergardenoffices.com/index.php?ref=POSTCODE-AREAS.

Postal Region Name	First two letters in postal code and corresponding area
East Anglia	AL = St. Albans, Hertfordshire CB = Cambridge, Cambridgeshire CM = Chelmsford, Essex CO = Colchester, Essex EN = Enfield, Middlesex IG = Ilford and Barking, Essex IP = Ipswich, Suffolk LU = Luton, Bedfordshire MK = Milton Keynes, Buckinghamshire NR = Norwich, Norfolk PE = Peterborough, Cambridgeshire RM = Romford, Essex SG = Stevenage, Hertfordshire SS = Southend-on-Sea, Essex WD = Watford, Hertfordshire
London	E = East London EC = City of London N = North London NW = North West London SE = South East London SW = South West London W = West End, London WC = West Central London
The Midlands	B = Birmingham, West Midlands CV = Coventry, West Midlands DE = Derby, Derbyshire DY = Dudley, West Midlands LE = Leicester, Leicestershire NG = Nottingham, Nottinghamshire NN = Northampton, Northamptonshire ST = Stoke-on-Trent, Staffordshire WS = Walsall, West Midlands WV = Wolverhampton, West Midlands
North East England	BD = Bradford, West Yorkshire DH = Durham, County Durham DL = Darlington, North Yorkshire DN = Doncaster, South Yorkshire

	HD = Huddersfield, West Yorkshire HG = Harrogate, North Yorkshire HU = Hull, North Humberside HX = Halifax, West Yorkshire LN = Lincoln, Lincolnshire LS = Leeds, West Yorkshire NE = Newcastle, Tyne and Wear S = Sheffield, South Yorkshire SR = Sunderland, Tyne and Wear TS = Teesside, Cleveland WF = Wakefield, West Yorkshire YO = York, North Yorkshire
North West England	BB = Blackburn, Lancashire BL = Blackpool, Lancashire CA = Carlisle, Cumbria CW = Crewe, Cheshire FY = Blackpool, Lancashire L = Liverpool, Merseyside LA = Lancaster, Lancashire M = Manchester, Lancashire OL = Oldham, Lancashire PR = Preston, Lancashire SK = Stockport, Cheshire SY = Shrewsbury, Shropshire TF = Telford, Shropshire WA = Warrington, Cheshire WN = Wigan, Lancashire CH = Chester, Cheshire
Northern Ireland	BT = Belfast
Scotland	AB = Aberdeen DD = Dundee, Angus DG = Dumfries and Galloway EH = Edinburgh, Midlothian FK = Falkirk, Stirlingshire G = Glasgow, Lanarkshire HS = Isle Of Lewis, Outer Hebrides IV = Inverness KA = Kilmarnock and Ayr, Ayrshire KW = Kirkwall, Caithness KY = Kirkcaldy, Fife ML = Motherwell, Lanarkshire PA = Paisley, Renfrewshire PH = Perth, Perthshire TD = Galashiels, Selkirkshire ZE = Shetland Islands
South Central England	GU = Guildford, Surrey HA = Harrow, Middlesex HP = Hemel Hempstead, Hertfordshire OX = Oxford, Oxfordshire PO = Portsmouth, Hampshire RG = Reading, Berkshire

	SL = Slough, Buckinghamshire SN = Swindon, Wiltshire SO = Southampton, Hampshire SP = Salisbury, Wiltshire UB = Southall and Uxbridge, Middlesex
South East England	BN = Brighton, East Sussex BR = Bromley, Kent CR = Croydon, Surrey CT = Canterbury, Kent DA = Dartford, Kent KT = Kingston-upon-Thames, Surrey ME = Medway, Kent RH = Redhill, Surrey SM = Sutton, Surrey TN = Tonbridge, Kent TW = Twickenham, Middlesex BA = Bath, Avon
South West England	BH = Bournemouth, Dorset BS = Bristol, Avon DT = Dorchester, Dorset GL = Gloucester, Gloucestershire HR = Hereford, Herefordshire PL = Plymouth, Devon TA = Taunton, Somerset TQ = Torquay, Devon TR = Truro, Cornwall WR = Worcester, Worcestershire
Wales	CF = Cardiff, South Glamorgan LD = Llandrindod Wells, Powys LL = Llandudno, Clwyd NP = Newport, Gwent SA = Swansea, West Glamorgan
Other regions	GY = Guernsey, Channel Islands JE = Jersey, Channel Islands IM = Isle Of Man

Appendix C

Table 11. Token occurrence count at the regional level for the U.S.

	Relative Directions			Cardinal Directions			
State(number of document)	change of direction	static spatial relationship	driving aid	traveling direction	change of direction	static spatial relationship	General origin
AL(101)	754	59	14	112	16	43	8
AR(69)	315	85	1	193	14	131	8
AZ(106)	269	88	3	58	31	46	12
CA(955)	5727	901	126	1477	115	348	228
CO(87)	240	27	5	47	28	35	10
CT(305)	1847	271	32	175	9	68	71
DC(48)	343	73	20	81	3	13	35
DE(7)	61	16	3	6	0	2	0
FL(320)	1189	196	33	593	46	105	86
GA(240)	2522	449	60	401	31	117	45
IA(72)	517	143	0	227	81	122	38
ID(23)	156	4	2	55	6	13	2
IL(224)	795	113	12	332	38	159	100
IN(79)	403	109	23	109	13	48	32
KS(48)	203	27	9	45	11	11	26
KY(298)	994	116	30	137	23	50	32
LA(123)	950	182	22	197	15	241	18
MA(521)	2658	456	107	235	7	53	173
MD(338)	1187	200	20	186	7	51	85
ME(107)	441	138	18	16	2	3	9
MI(194)	2101	258	28	393	41	112	46

MN(145)	708	237	25	212	30	90	58
MO(138)	958	250	16	171	30	96	36
MS(61)	323	72	11	115	2	23	20
MT(29)	297	44	2	79	12	27	79
NC(409)	2363	363	57	248	39	66	67
ND(14)	38	12	1	36	1	9	8
NE(34)	76	22	0	45	8	13	16
NH(141)	729	301	18	121	12	14	41
NJ(573)	2831	637	166	551	26	115	235
NM(22)	80	28	1	13	4	15	5
NV(21)	92	28	6	39	2	2	17
NY(852)	4040	586	136	1155	29	307	263
OH(299)	2656	274	30	323	27	125	204
OK(56)	191	43	8	68	16	21	13
OR(67)	314	71	20	107	3	29	27
PA(796)	5113	843	280	714	51	159	232
RI(60)	363	122	4	49	0	5	47
SC(97)	488	139	10	22	16	10	4
SD(19)	49	11	0	4	4	13	23
TN(178)	1606	236	39	192	12	88	29
TX(523)	1880	280	53	682	62	221	106
UT(43)	337	50	15	70	16	47	13
VA(556)	3374	463	102	421	26	156	100
VT(72)	369	158	5	113	2	31	47
WA(291)	1590	365	55	238	37	106	69
WI(206)	1081	192	17	217	14	92	53
WV(133)	768	572	25	76	5	75	11
WY(6)	39	16	0	8	0	2	0

Table 12. Token occurrence count at the regional level for the U.K.

	Relative Directions			Cardinal Directions			
Postal Region (number of document)	change of direction	static spatial relationship	driving aid	traveling direction	change of direction	static spatial relationship	General origin
London(68)	287	165	15	68	1	43	48
Midlands (93)	757	353	25	462	11	64	184
East Anglia (56)	715	300	38	137	1	33	65
Northeast England (108)	699	534	25	210	11	22	78
Northern Ireland (1)	14	13	8	0	0	0	0
Northwest England (56)	389	281	5	55	2	27	28
South Central England (102)	1193	716	30	189	10	67	137
Scotland (61)	287	165	15	68	1	43	48
Southeast England (95)	1254	328	53	128	3	65	92
Southwest England (49)	158	102	2	7	1	1	21
Exeter (6)	10	1	6	0	0	2	2
Wales (13)	23	11	0	1	0	3	2
Other Region (2)	0	0	0	0	0	8	0

Table 13. Token occurrence count at the regional level for Australia

	Relative Directions			Cardinal Directions			
State (number of document)	change of direction	static spatial relationship	driving aid	traveling direction	change of direction	static spatial relationship	General origin
ACT(4)	20	3	0	2	0	0	0
NT (3)	0	0	0	0	0	0	0
TAS (19)	78	23	2	7	0	4	0
WA (38)	200	45	4	67	0	24	27
SA (36)	229	69	4	177	12	37	28
QLD (82)	512	173	17	150	6	26	33
NSW (213)	942	299	24	414	43	110	93
VIC (94)	442	179	6	109	4	67	19