Exploring regional variation in spatial language
-A case study on spatial orientation with spatially-stratified web-sampled documents

Sen Xu*, Alexander Klippel*, Alan MacEachren*, Prasenjit Mitra†, Ian Turton*, Xiao Zhang†, Anuj Jaiswal‡

*GeoVISTA Center, Department of Geography, Pennsylvania State University, Pennsylvania, U.S.A.
‡Information Science and Technology, Pennsylvania State University, Pennsylvania, U.S.A.

Introduction

How does spatial language differ regionally? Spatial language contains expressions that characterize static and dynamically changing spatial relations. It has long been noted, that a) different languages refer to the same spatial information differently and that b) these differences can occur even within the same language but in different regions. Collecting data has primarily been accomplished by individuals via time consuming studies. With the advent of methods and tools developed in computational linguistics and in geographic information retrieval (e.g., geo-referencing), new opportunities arise and allow for efficiently collecting and analyzing large amounts of data.

As a case study using a corpus that we created, tools, and analysis scheme, we address the often raised question of the usage of cardinal/relative directions in route directions both to indicate actions as well as static spatial relations (detailed semantic categories defined in Table 1). The Spatially-strAtified Route Direction Corpus (SARD Corpus) consisting of more than 10,000 web documents from three countries (U.S., U.K., and Australia) is used in this analysis. Linguistic variations on both national level and regional level are observed. Similarities and differences of cardinal/relative direction usage are shown at both national and regional level: geographical patterns are identified at the regional level. The explanation of the observation may lie in the correlation between the regional linguistic preference and geographical feature.

Methods

The WWW is an ideal data source to sample data with extensive geographical coverage. By sourcing from the World Wide Web with a spatially-stratified data collection schema (see Figure 1), the SARD Corpus has been built to investigate regional variations in route directions. We used a machine learning-based document classification approach [1] to crawl and identify web documents that contain route directions and associated them with postal codes. Text processing tools (Term Tree Tools [2]) assisted in the analysis by providing a spatial language usage taxonomy from region to region for the large quantity of text data. Computational analysis results are complemented by visual analytics tools to allow for a more precise analysis of linguistic usage patterns.

Figure 1. Data flow diagram for building motion-descriptive language corpus

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. This visualization provides geographical information to assist analysts to detect regional variations in the linguistic characteristics. For example, in cardinal direction usage in Figure 3, there is a geographic pattern (South Dakota to Kansas, Wyoming to Iowa, blue circled) that differs from its surrounding states in every semantic category. The regional pattern detected is comparable to the Colorado West and Central West regional pattern of U.S. dialects [3]. This finding shows that the difference in spatial language usage may correspond to dialect differences. From Figure 4, we can see that the northern regions of the U.K. differs from the southern regions. There is a regional pattern in the Midlands, East Anglia and Northeast England which stands out from their surrounding regions (blue circled). The regional difference noted might be caused by culture or linguistic (dialectal) differences or from correlation between the regional linguistic preference and regional geographical features, which is yet to be investigated.

This case study offers a novel perspective on performing spatial linguistic studies. The design and implementation of building a geo-referenced large-scale corpus from Web documents offers a methodological contribution to corpus linguistics, spatial cognition, and GI-Sciences.

Results & Discussion

The regional linguistic characteristics are presented and compared at both the national level and regional level. Figure 2 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. There are more distinctive usage differences at the regional level (refer to regional histogram in handout) which is important for detecting regional variations.

Figure 2. National level histogram of relative direction (RD) (left) and cardinal direction (CD) (right) usages (Top: token occurrence count, Bottom: Proportion)

Table 1. Semantic categories for cardinal direction and relative direction

<table>
<thead>
<tr>
<th>Semantic Categories</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Represent change of direction</td>
<td>like a left, bear right</td>
</tr>
<tr>
<td>2. Represent static spatial relationship</td>
<td>comes to the left of</td>
</tr>
<tr>
<td>3. Represent traveling aid</td>
<td>keep to the left lane, merge right</td>
</tr>
<tr>
<td>4. Indicate general origin</td>
<td>comes from, if coming from south of</td>
</tr>
</tbody>
</table>

*used in POI names

Figure 3. Region-level comparison of RD and CD usages in the U.S.

References

