Spatial Event Language across Geographic Domains

Alexander Klippel, Sen Xu, Jinlong Yang, Rui Li

Department of Geography, GeoVISTA Center,
The Pennsylvania State University, PA, USA
{klippel,sen.xu,jinlong,rui.li}@psu.edu

Abstract. We present first results of an analysis of a corpus of linguistic descriptions that were collected in controlled experiments. This corpus and its analysis add to the body of knowledge on formal models for spatial language, language interpretation and generation. The experiments are grounded in qualitative formalisms (RCC and Intersection Models, IM) that have a long standing tradition as means to bridge formal and linguistic descriptions of space and spatial relations. Our experiments address dynamically changing spatial relations (movement patterns/geographic events). By keeping the formal spatial characterizations identical across experiments but changing the semantics (that is, we used movement patterns across seven different geographic domains such as a hurricane in relation to a peninsula, plus two geometric figure domains) we contribute to disentangling spatial and domain specific aspects of spatial (event) language. We briefly discuss here two aspects: First, we hand examine the corpus by selecting participants that show the same conceptual behavior as identified through RCC/IM; second, we analyze the domain specific sub-corpora to address similarities and dissimilarities between individual domains.

Keywords: Event language, topology, corpus analysis.

1 Introduction

Formal models of spatial language play an import role in several disciplines addressing questions of (natural) language processing, natural language generation, the automatic description of spatial scenes, or the design of unifying frameworks for multimodal information systems and processing [1–4]. While we are in the age of spatio-temporal representation and reasoning, the four-dimensional treatment of spatial language (and information in general) is still a hotly debated topic. With respect to language, research shows that naming of events is more challenging than naming of object [5] and it is therefore not surprising that the insights gained from describing static spatial relations linguistically need to be carefully evaluated and extended to the dynamic domain. This contribution is addressing this issue by combining approaches to model events employing qualitative spatial formalisms with linguistic analysis.

2 Approach

We have developed an experimental paradigm that allows us to evaluate the influences of domain semantics on the conceptualization of movement patterns as well as how movement patterns are linguistically described. Here we focus on the linguistic descriptions. Our framework is based on a topologically defined conceptual neighborhood graph [6–8]. Figure 1 provides an overview of the different semantic domains that we have subjected to behavioral validation. In a nutshell: We distinguish movement patterns on the basis of formal path characteristics as identified by the conceptual neighborhood graph. The shortest path (in each scenario) is a single topological relation, DC (disconnected), the longest path (in each scenario) is defined as follows: DC-EC-PO-TPP-NTPP-TPP-PO-EC-DC. To give an example, a boat that never touches or crosses an area of shallow water will always be disconnected (DC) from it. In contrast, a boat that makes it completely across an area of shallow water will exhibit the long path characteristics with the start and end relation being identical (DC). Our participants have to perform a grouping task as a way to elicit conceptual knowledge. After performing this task, participants are presented with the groups that they created again and are asked to provide linguistic descriptions: a short label and a longer description detailing the grouping rational.

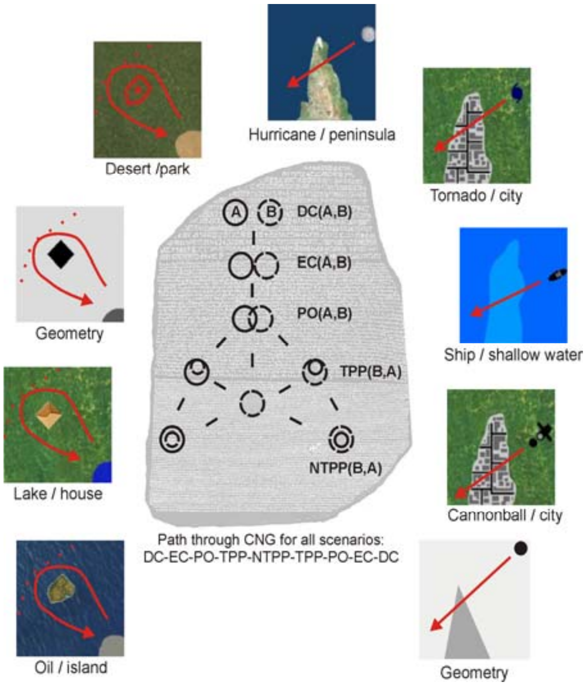

Fig. 1. Nine scenarios from our experiments. Left: four scaling movement patterns: An extending desert in relation to a recreational park, two geometric figures showing a static diamond and an extending/shrinking circle, a lake extending in relation to a house, and an oil slick extending in relation to an island. Right: five translation scenarios: A hurricane in relation to a peninsula, a tornado in relation to a city, a ship in relation to a body of shallow water, a cannonball in relation to a city, and two geometric figures. They are arranged around the Rosetta Stone because all movement patterns in all our experiments are characterized by *topologically equivalent paths* through the conceptual neighborhood graph (which is overlaid on top of the Rosetta Stone).

3 Some results

As rich as our data set is, the flexibility of natural language has made it a challenging task to analyze it. We are presenting two approaches. First we had a look into linguistic descriptions for specific paths. Here we show results from four domains, two from our translation movement patterns (geometry and hurricane) and two from scaling movement patterns (geometry and lake). This path (DC) could be described as *a hurricane not making landfall* or *a lake not flooding the house*. Our goal was to analyze the variety of linguistic descriptions that participants use to this relatively simple scenario. Table 1 provides some representative examples. The important distinction that we made for both scenarios is whether the spatial information (about the movement patterns) in these two scenarios is linguistically encoded using *spatial language*, or, whether this information is encoded using *domain specific language*. The two corresponding geometry scenarios serve as a reference as they obviously do not easily allow for using domain semantics.

With respect to the spatial language we find very diverse ways of conveying spatial information. We do believe that this diversity is fostered by the fact that our research is addressing geographic events / spatio-temporal information (rather than static spatial relations). Especially in the hurricane example we find the following strategies: relative reference frames focusing on the end relations of the geometric characteristics of figure and ground; qualitative distance-based descriptions; negation of what the path does not do; absolute reference; (experiment) context specific descriptions; explicit topological descriptions; intrinsic reference induced by the movement. Interestingly, the explicit spatial descriptions in the lake scenario seem to be less varied, indicating a potential difference between scaling and translation movement patterns that are indistinguishable from a topological perspective.

In both scenarios we also find descriptions that are encoding the spatial event in terms of domain specific language (to different degrees). While, for example, a statement such as “no hit”, “weak hit”, “no landfall” are still rather explicit, a statement such as “weak hurricane” relies heavily on background knowledge of a scenario and is open to interpretation. In case of the lake, the descriptions are much less varied, again, and in most cases refer to a flood not happening.

Table 1. Linguistic descriptions for the shortest possible path (DC).

Hurricane	**Lake**	**Geometry translation**	**Geometry scaline**
Right side stopping circles	Below house Away from the house	Outside right Outside right	Under the box Up and didn't get too far
Hurricane stops short of land	Not touched Not covered	Any part outside the triangle	Team grow Bellow box
Path doesn't cross Completely off east coast Right side	Water reaches short of house	Balls outside triangles Before Fully outside Off to the right	Circle grows beneath box. Half way to the diamond
Outside right		Right	Straight 1/4
Before land	Not flooded No flood	Stopped on right of triangle	Fall short Expand before
Don't make it	Dry house	To the right	Below
Hurricanes that never made it to shore	Short flood No flooded house	Far outside on the right Ball on triangle	Before diamond stop Expanding short stop.
No hit or weak hit	No house flood	Too short	Expand halfway
Calm right before the storm	No flooding Lower risk	Outside Outside right	No contact Stop short
Weak hurricane	Tiny lakes		Grow stop between
Weak hurricanes			Out not close to square
No landfall			Far away
No landing			
Pre-landfall hurricanes			

These findings led us to explore differences between the sub-corpora (the nine different scenarios). First, some domain-corpus properties can be extracted using AntCont [9]. The token occurrences are visualized using Wordle (http://www.wordle.net/), seeFigure 2.

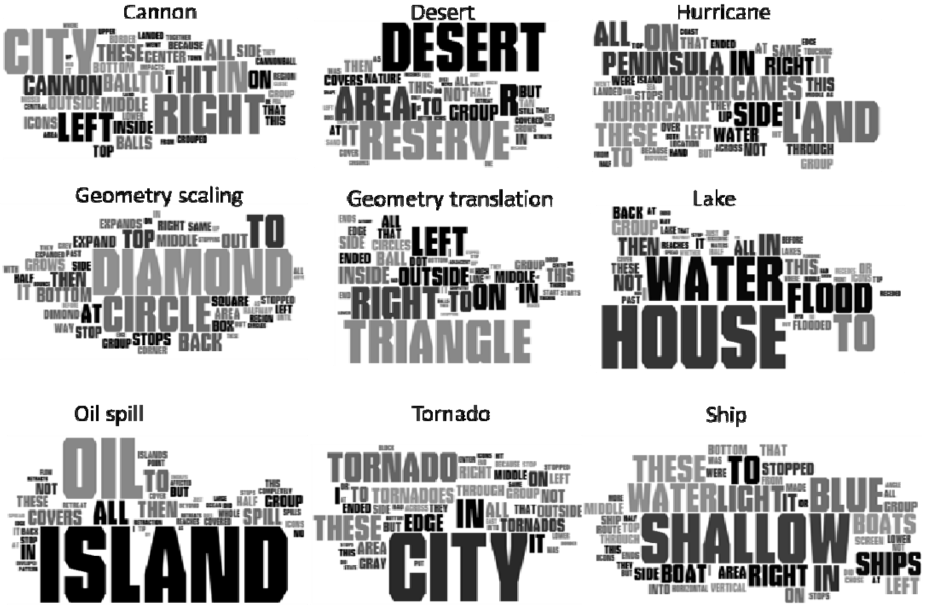

Fig. 2. Frequently appearing words in 9 corpora

From the tag clouds in Figure 2, we can see that top frequent words are mostly related to domain specific semantics. For example, *city*, *desert*, and *island* are referring to objects illustrated in each scenario. It is not surprising that participants make use of the domain semantics for reference to objects in the scenario, as it is a direct and succinct way to describe an object and distinguish it from surroundings. However, for the topological change depicted in different icons, participants would have to use more complicated descriptions such as verb phrases and prepositional phrases. This is the reason that spatial language terms such as *on*, *middle*, *at*, *outside*, *left*, *right*, *through*, and *ended* also appear prominent in the tag cloud. Our analytical question is: given scenarios where only domain semantic is different, how different will the descriptions be?

In the next step, we used the Stanford POS tagger [10]. We investigate the most frequently appearing nouns, verbs, adjectives, and prepositions:

- The most frequently appearing nouns are domain specific ones (see Figure 1). Domain specific nouns with top frequency in one corpus are often never found in other corpora, such as *tornado*, *oil*, and *desert*. Nouns that can be found across domains are common referral terms, such as *side*, *icons*, *middle*, and *bottom*.
- Frequently appearing verbs seem to be not as domain specific as nouns. Common verbs are various forms of *be*, *end*, *touch*, *have*, and *go*. However, there are a few verbs that appear frequently in some corpora but not in others. *Hit* and *miss* frequently appear in the Cannon, Hurricane, and Tornado corpus. *Cover*, *expand* and *grow* frequently appear in the Desert, Oil, and Lake corpus. It is not surprising because *hit* and *miss* can be naturally used for describing “translation” while *cover*, *expand* and *grow* naturally relate to “scaling”, which is the major difference in the above two corpora sets. There are also cases where verbs are specific to a domain. *Landed* used as a verb frequently appears

in Cannon and Hurricane. *Flooded* used as a verb appears exclusively in Lake. *Sailed* exclusively appears in Ship. This shows that domain semantic also influences verb usage, but not as explicit as nouns. More examples are *recede*, *retreat*, *leave*, *surrounded*, *shrink*, and *disappear*.

- Adjectives seem to even less domain specific. Common adjectives across all corpora are *middle*, *same*, *right*, and *lower*. The few cases where adjective are domain specific are the use of colors. *Blue*, *grey*, and *red* appears as to provide additional referral information respectively in Ship, Geometry, and Desert corpus. Exclusively in the Ship corpus, *shallow*, *light* and *dark* are frequently used to refer to the boundaries or the center of the water body. Adjectives about size were also used. *Large* appear more often in Oil.
- Prepositions are the least domain specific lexical category. Few prepositions are domain specific. *Across* frequently appears in translation scenarios but not in scaling ones.

In sum, POS-tagging offers possibilities to examine linguistic usages by lexical categories. Examining the nine corpora, frequently appearing nouns are highly domain specific; a few verbs and adjectives are domain specific and a general difference in translation vs. scaling can be found; prepositions are least domain specific, only the word "across" is found to be differentiable between translation scenarios vs. scaling scenarios.

The last analysis step here involves topic modeling [11,12]. It is a method for discovering "topics" shared among documents within a corpus. It can be viewed as cluster analysis for documents. Applying topic modeling to all documents (one for each participant, 20 documents per scenario) in the nine corpora (180 documents in total), we can evaluate whether documents might be clustered based on their domain. Mallet (Machine Learning for LanguagE Toolkit) [13] is used to realize topic modeling. Setting the "number of topics" to be nine, we can see if the nine topic models correlate with the nine domains (scenarios). Because topic models are data-driven and don't imply any predefined knowledge, we want to compare the topic modeling result with domain semantics and see if they are comparable to each other. Each topic is defined by the keywords appearing most frequently and most distinctively.

Table 1. Keywords for the nine topic models (TopicID).

Topic ID	Keywords
0	bottom diamond circle stops top stop back grows expand box expands halfway corner middle diamond grey expanded moves touches
1	middle ended lower blue light ships upper boats side boat left top corner chose screen section hand cross horizontal
2	area half touch past stopping retreat point square part retreats short tan position contact full expanding small space pass
3	left side triangle inside ball end center circles middle ends line landed start high dot touching fell location images
4	water house flood back entire recedes reaches flooded touching halfway spread lake lakes front receded show past starting receeding
5	group desert reserve stopped fully put based touched nature partially chose groups animations criteria reached red shape sand choose
6	island oil covers completely covered cover spill stop reach ocean tip covering islands barely large pattern reaches spills animation
7	city edge icons tornado cannon balls region tornadoes border gray grouped tornados east boundary enter missed southwest town block
8	shallow land hit peninsula hurricanes hurricane moving made mid close west coast ship part central move hits low

Table 1 shows the keywords that identify each topic model. Unsurprisingly, domain specific nouns are distributed across topic models. These topic models can be used to evaluate the probability of one document (descriptions created by one participant) being associated with a specific topic model (ideally catching the domain). Assigning the most probable topic ID to a document allows for using topic models for document classification. To evaluate the correlation between topics and domain semantics further, we use the already built topic models to classify each document. The results and evaluations are shown in Table 2.

Table 2. Matching nine topic models to the nine domains in a confusion matrix.

Topic ID **Domain**	0	1	2	3	4	5	6	7	8
Geometry_translation	**16**	0	0	1	1	0	1	1	0
Desert	1	**14**	0	0	1	0	2	0	2
Lake	0	0	**19**	0	0	0	1	0	0
Cannon	3	0	0	3	3	1	0	**10**	0
Hurricane	0	0	0	1	**16**	0	1	1	1
Geometry_scaling	0	4	0	1	0	**14**	1	0	0
Oil	5	0	1	3	0	0	**11**	0	0
Tornado	0	0	0	0	3	0	2	**15**	0
Ship	0	0	0	0	0	0	1	0	**19**

Out of 20 documents from each domain, we evaluated the proportion of documents being classified into the same topic model, which ideally should correspond to the domain (this correspondence worked except for tornado and cannon, where most of both are assigned to Topic ID 7). The **bold** numbers in Table 2 shows the topic model (see also Table 1) that most documents from a domain are assigned to. As shown in Table 2, it is reasonable to match each topic ID to one domain semantic and the matching proportion (sum of diagonal cells divided by total) is 70.56%. Cross-examining the domain semantic with keywords from corresponding topic models (see Table 1) we find that a large proportion of documents are classified correctly.

However, the above matching of topic models and domain semantics may be skewed by the high volume of domain specific nouns. Hence, as a comparison, we removed all the domain specific nouns from all corora and rebuilt the topic models.

Table 3. Keywords for nine topic models (excludes domain specific nouns).

TopicID	**Keywords**
0	area hit grows box half gray touch tan past enter missed grow leaving boxes consumes retreat green direction paths
1	half covers cover covered fully entire recedes tip touch reach ocean covering recede oil starting sand island retraction affected
2	land icons touching chose center grouped landed hand border put close section location barely didn mass shore passed landing
3	left side middle top bottom end corner ends start high starts drop adjacent moved inbetween flush receds hang till
4	shallow stopped blue light moving made screen based cross horizontal vertical route low make angle map body path sailed
5	edge inside ended city line mid dot west east fell ball boundary central criteria south images portion impacts southwest
6	completely point flood stopping square touched icons retreat receded show house groups large pattern space receeding grass part recession
7	back stops stop halfway reaches past short grey retreats expanded front full moves touches position hits expanding reached disappears
8	group lower upper expand region part expands spread animations straight red shape shrink partially grew diamond slightly icon movement

As shown in Table 3, because all domain specific nouns are excluded, the keywords are not as clearly correlated to domain semantics. Nouns that are not domain specific, verbs, adjectives, prepositions and all other words are still kept and were used for building another topic model. In the following we analyze if these words can create document clusters that correlate to domain semantics, too.

Table 4. Matching nine topic models (excluding domain specific nouns) to the nine domains in a confusion matrix.

Topic ID **Domain**	0	1	2	3	4	5	6	7	8
Geometry_scaling	5	5	0	1	1	2	6	0	0
Hurricane	1	1	4	3	4	2	1	4	0
Tornado	1	0	10	1	1	2	4	0	1
Lake	3	1	0	7	0	2	3	0	4
Ship	8	3	0	0	9	0	0	0	0
Desert	0	0	0	1	0	5	3	1	10
Geometry_translation	1	2	1	1	1	4	2	7	1
Cannon	2	2	6	1	1	1	0	6	1
Oil	3	0	0	5	0	3	1	0	8

From Table 4, we can see that document from each domain are being classified as belonging to various topic models. It is a stretch to relate topic IDs from this topic model to the nine domains and the highest possible matching proportion is only 29.44%. This result shows that excluding domain specific nouns lets the correspondance between topic models and domain semantics disappear.

4 Conclusions

Two observations are important: In the first part of this paper we showed an analysis by hand that allows for relating a qualitative formal description of a movement pattern to a linguistic description. The linguistic descriptions are varied and participants used manifold strategies to characterize formally identical movement patterns. However, we seem to be able to clearly reveal domain specific differences, especially if we look into whether or not domain semantics is present. In the second part of this paper we tried to use this insight and compared the documents from each domain (one document with all linguistic descriptions per participant, 20 documents in each domain). We found that figure and ground (moving entity and reference entity) are the dominating linguistic features used and that these nouns allow for classifying documents largely correctly. However, once we remove these obvious, domain specific features, classification and identification of documents becomes very inaccurate despite the differences we found in the first part.

There could be a number of reasons for this. Instead of comparing all documents of a particular domain, which contains linguistic descriptions of several, topologically distinguishable paths, we may need a finer granularity for the analysis. For example, we could extract all DC descriptions from all domains and focus only on these. Likewise, we could extract all descriptions for movement patterns that could be labeled *across* in the translation scenarios and *expand-and-retreat* in the scaling scenarios. We could perform this analysis for all topologically equivalent movement patterns that we used to design our experiments.

It also could be that the topic modeling approach we used needs refinement. Topic models make use of terms and co-occurrences with documents to discover topics. It is an effective method for knowledge discovery from large corpora without predefined knowledge. However, we are specifically looking for spatial language usage in this study. In order to reduce the influence of domain specific nouns, we use a crude method which is removing the domain specific nouns. Integrating predefined knowledge (in our case, specific target language and contexts) into topic models would allow an analysis to focus on certain term usages, which would enhance the capability of topic modeling.

To sum up, we presented a first exploratory analysis of a corpus that is the result of the conceptualization of movement patterns in different semantic domains. The unique aspect of our experiments is that grounding the design in qualitative spatial representation and reasoning frameworks allows for keeping the spatial information identical across domains only changing the semantic (domain specific) context. We are hopeful that this corpus can contribute to a better understanding of the relation between formal/computation models and spatial language across different domains.

References

1. Bateman, J.A.: Language and space. a two-level semantic approach based on principles of ontological engineering. International Journal of Speech Technology **13**, 29–48 (2010).
2. Galton, A.: Spatial and temporal knowledge representation. Earth Science Informatics **2**, 169–187 (2009).
3. Kordjamshidi, P., Otterlo, M. von, Moens, M.-F.: From language towards formal spatial calculi. In: Ross, R.J., Hois, J., Kelleher, J. (eds.) Computational Models of Spatial Language Interpretation (CoSLI) Workshop at Spatial Cognition 2010, Mt. Hood, Oregon, pp. 17–24. CEUR Workshop Proceedings (2010).
4. Ross, R.J., Hois, J., Kelleher, J. (eds.): Computational Models of Spatial Language Interpretation (CoSLI) Workshop at Spatial Cognition 2010, Mt. Hood, Oregon. CEUR Workshop Proceedings (2010).
5. Gentner, D., Boroditsky, L.: Individuation, relativity, and early word learning. In: Bowerman, M., Levinson, S.C. (eds.) Language acquisition and conceptual development, pp. 215–256. Cambridge Univ. Press, Cambridge (2001).
6. Egenhofer, M.J., Al-Taha, K.K.: Reasoning about gradual changes of topological relationships. In: Frank, A.U., Campari, I., Formentini, U. (eds.) Theories and methods of spatio-temporal reasoning in geographic space, pp. 196–219. Springer, Berlin (1992).
7. Freksa, C.: Temporal reasoning based on semi-intervals. Artificial Intelligence **54**, 199–227 (1992).
8. Randell, D.A., Cui, Z., Cohn, A.G.: A spatial logic based on regions and connections. In: Proceedings 3rd International Conference on Knowledge Representation and Reasoning, pp. 165–176. Morgan Kaufmann, San Francisco (1992).
9. Anthony, L.: AntConc (version 3.2.2). Waseda University, Tokyo, Japan (2011). available from http://www.antlab.sci.waseda.ac.jp/.
10. Toutanova, K., Klein, D., Manning, C., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of HLT-NAACL 2003, pp. 252–259 (2003).
11. Blei, D.M., Lafferty, J.D.: Topic models. In: Srivastava, A.N., Sahami, M. (eds.) Text mining. Classification, clustering, and applications, pp. 71–93. CRC Press/Taylor & Francis, Boca Raton, Fla (2009).
12. Steyvers, M., Griffiths, T.L.: Probabilistic topic models. In: Landauer, T.K., McNamara, D.S., Dennis, S., Kintsch, W. (eds.) Handbook of Latent Semantic Analysis. Lawrence Erlbaum, Mahwah, NJ (2007).
13. McCallum, A.K.: MALLET. A machine learning for language toolkit (2002). http://mallet.cs.umass.edu/.