The Occurrence and Distribution of Spatial Reference Relative to Discourse Relations

Blake Stephen Howald[1]

Georgetown University, Department of Linguistics, ICC 479, 37th and O Streets, NW
Washington, DC 20057-1051
{bsh25}@georgetown.edu

Abstract. I present a descriptive analysis of reference to physical space in the Penn Discourse TreeBank. In particular, I analyze the occurrence of spatial prepositional phrases relative to the discourse relations and semantic senses that hold between two adjacent clauses. The purpose of this investigation is twofold: (1) to better understand how often spatial reference occurs in discourse and (2) to investigate possible relationships between spatial reference and discourse semantics. Overall, the distribution of spatial prepositional phrases and relation-sense pairs are similar. However, statistical evidence suggests that the inclusion of spatial reference in a given clause is independent of the relation-sense of that clause and adjacent clauses. While these results, as applied to the PDTB, indicate the absence of a default pattern of occurrence and discourse semantic function of spatial information, they can nonetheless be extrapolated to provide crucial insights for fully understanding models of spatial representation and interpretation in discourse generally.

Keywords: Spatial Reference, Discourse Relations.

1 Introduction

The semantic and pragmatic functions of discourse relations, which hold between two clauses, contribute to a text's coherence [1]. For example, in the two-line discourse (a) *Lucy is not hungry* (b) *Cati fed her*, (b) is an EXPLANATION for (a) [2]. The inclusion of spatial reference, while accounted for in definitions of discourse relations (e.g., BACKGROUND), is not strictly necessary. However, recent research, grounded in spatial cognitive psychology (e.g., cognitive maps), has suggested that space plays a larger role in discourse structure; in particular, spatial reference organizes narrative discourse into spatially defined groups of events that are temporally linked [3-4]. While this research presents a new analytical perspective, before it can be fully exploited, it is first necessary to better understand what relationships may exist between spatial reference and discourse relations generally.

[1] I would like to thank two anonymous reviewers, my dissertation advisor E. Graham Katz, James Pustejovsky and David Herman for beneficial insights and discussion.

This paper presents the results of a descriptive analysis that evaluates the interface of spatial information and discourse. The particular research question addressed is: Does the occurrence of spatial reference in discourse pattern relative to discourse relations? A negative answer, which is suggested by existing definitions of discourse relations, indicates that spatial reference is independent of discourse relations. An affirmative answer indicates that spatial reference is dependent on (certain) discourse relations. This paper is arranged as follows: Section 2 discusses spatial information (as defined by The Preposition Project [5]), discourse relations (as defined by Penn Discourse TreeBank (PDTB) [6]) and the methodology employed. Section 3 presents the distribution of spatial prepositional phrases relative to discourse relations. Section 4 concludes.

2 Background, Data and Methodology

In this paper, "spatial reference" refers to physical relationships arranged in *figure* and *ground* relationships. For example, *the cup is on the table* locates the figure *the cup* relative to the ground *the table*.[2] A search algorithm was developed to automatically extract 334 different prepositions defined in the Preposition Project [5] (based on a hierarchical network of dictionary entries). 107 of the 334 prepositions have a distinct "spatial" sense. Because prepositions are highly ambiguous (e.g., numerous non-spatial senses), the prepositions extracted from the PDTB were disambiguated by hand.

The PDTB includes annotations of discourse relations in the Penn Treebank II version of the Wall Street Journal (WSJ) corpus [8]. Discourse relations in the PDTB (which hold between pairs of syntactically classified arguments from Penn TreeBank II) ("ArgPairs") are a confluence of connective words, content of the ArgPairs and semantic senses. ArgPairs are either: *Explicit* – a syntactically classified connective word exists in the text (*but*, *and*); *Implicit* – a connective word does not exist in the text but can be inferred; *EntRel* – no relation holds, but the second clause in the ArgPair includes more information about the first clause; *AltLex* – there is no connective word, but a non-connective expression can capture an inferred relation; and *NoRel* – no relation holds. Explicit, Implicit and AltLex ArgPairs co-occur with one of four senses: *Temporal*, *Contingency*, *Comparison* and *Expansion*. The PDTB includes 2159 annotated documents, 40,600 relations and 34,877 senses in total. The overall distribution of the relations and senses in the PDTB provide a baseline of relation-senses. The occurrence or non-occurrence of spatial reference overall, and relative to particular relation-senses and pairs of relation-senses, can then be compared to this baseline to determine relevant (statistically significant) differences and potential patterns.

[2] For sake of brevity, I am restricting the discussion to figure and ground relationships indexed by spatial prepositions [7]. Other sources include motion verbs (*run*, *follow*), deictic verbs (*come*, *go*) and deictic adverbs (*here*, *there*).

3 Results – Distributions and Dependency

200 documents (approximately 10% of the total PDTB), consisting of 5000 relations and 4388 senses, were selected for analysis. If one or both of the arguments in an ArgPair contained one or more spatial prepositions, then these are referred to as Spatial ArgPairs.[3] The occurrence of Spatial ArgPairs is roughly equally distributed between each argument (Arg1 – 54.15%; Arg2 – 45.84%). The average percentage of Spatial ArgPairs per document is 28.90%. The sample selected for analysis conforms to the general relation and sense distributions in the PDTB (Table 1).

Table 1. Distribution of relations and senses.

Relations	PDTB (%)	Sample (%)	Spatial (%)	Senses	PDTB (%)	Sample (%)	Spatial (%)
Explicit	18459 (45.46)	2311 (46.22)	605 (41.86)	**Expansion**	15432 (44.24)	1832 (41.75)	524 (43.73)
Implicit	16053 (39.54)	2002 (40.04)	596 (41.24)	**Contingency**	8016 (22.98)	1005 (22.90)	255 (21.28)
EntRel	5210 (12.83)	578 (11.56)	209 (14.46)	**Comparison**	7634 (21.88)	940 (21.42)	272 (22.70)
AltLex	624 (1.54)	75 (1.50)	22 (1.52)	**Temporal**	3795 (10.88)	611 (13.92)	147 (12.27)
NoRel	254 (.63)	34 (.68)	13 (.89)				
Total	**40600**	**5000**	**1445**	**Total**	**34877**	**4388**	**1198**

There does not seem to be any independent pattern demonstrated by the Spatial, as compared to Non-Spatial, ArgPairs. This is supported by X^2. H_0 is that the occurrence or non-occurrence of spatial reference is independent of a given relation-sense. For the top six relation-senses occurring in the sample (Explicit-Expansion (EE), Explicit-Comparison (EP), ENT, and Implicit-Contingency (IC)), H_0 can be accepted as the *p*-value is greater than .05 and rejected for the Implicit-Expansion (IE) and Explicit-Temporal (ET) relation-senses as the *p*-value is less than .05 (Table 2).

Table 2. X^2 for spatial and non-spatial relation-senses and pairs.

Relation-Sense	Non-Spatial	Spatial	*p*-value	Relation-Sense Paris	Non-Spatial	Spatial	*p*-value
IE	1073	384	.0002	**IE - IE**	223	102	.6499
EE	796	197	.0546	**EE - IE**	163	70	.9507
EP	630	175	.6575	**IE - EE**	163	72	.8568
ENT	595	180	.5580	**IE - EP**	128	52	.6953
IC	513	157	.5291	**EP - IE**	118	56	.5738
ET	451	99	.0131	**EE - EE**	110	40	.3421

[3] 40 of the 107 Preposition Project prepositions are represented in the analyzed sample (N = 2214) with common prepositions making up the majority (82.92%): *in* – 880 (39.74%); *at* – 335 (15.13%); *to* – 250 (11.29%); *on* – 142 (6.41%); *from* – 130 (5.07%); *of* – 117 (5.28%). The remaining 36 prepositions account for the 17.08% complement.

However, the effect that is being exhibited by the IE and ET relation-senses arguably has more to do with the occurrence of Non-Spatial ArgPairs because of the comparative number (1073 Spatial vs. 384 Non-Spatial for IE and 796 vs. 197 for EE). For pairs of relation-sense s, H_0 can be accepted in all cases (the top six pairs of relation-senses in Table 2) as the p-value is greater than .05. This indicates that, even in greater local context, the occurrence or non-occurrence of spatial information is independent of a given pair of relation-senses.

4 Conclusions and Limitations

In sum, as applied to the PDTB, for the studied sample, there is statistical evidence to support a negative answer to the posed research question: whether or not a figure and ground relationship occurs, indexed by a spatial preposition, is independent of the type of discourse relation. This insight may prove useful in interpreting the results of computational tasks that interpret, represent and analyze spatial information in discourse. The main limitations in this study are the amount of data and scope. Future research will focus on more linguistic spatial phenomenon and larger corpora with varied genres (the WSJ corpus consists of Essays, Summaries, Letters and News; the latter of which accounts for roughly 90% of all text in the corpus [9]). Nonetheless, the present results facilitate a more complete understanding of spatial reference in discourse structure. The occurrence of spatial reference does not appear to be biased by inherent discourse patterning.

References

1. Hobbs, J.: On the Coherence and Structure of Discourse. Technical Report CSLI-85-37, Center for the Study of Language and Information, Stanford University (1985)
2. Asher, N., Lascarides, A.: Logics of Conversation. Cambridge University Press, Cambridge, UK (2003)
3. Herman, D.: Spatial Reference in Narrative Domains. Text 21(4), 515--541 (2001)
4. Howald, B.: Granularity Contours and Event Domain Classifications in Spatially Rich Narratives of Crime. COSIT 2009 Workshop on Presenting Spatial Information: Granularity, Relevance, and Integration, Aber Wrac'h, France, http://repository.unimelb.edu.au/10187/5516 (2009)
5. Litkowski, K.:Digraph analysis of dictionary preposition definitions. In: Proceedings of the SIGLEX/ SENSEVAL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions, pp. 9--16. Association for Computational Linguistics (2002)
6. Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A., Joshi, A., Robaldo, L., Webber, B.: The Penn Discourse Treebank 2.0 Annotation Manual. The PDTB Research Group (2007)
7. Asbury, A., Gehrke, B., van Riemsdijk, H., Zwarts, J.: Introduction: Syntax and Semantics of Spatial P. In: Asbury, A., Dotlacˇil, J., Gehrke, B., Nouwen, R. (eds.) Syntax and Semantics of Spatial P, pp. 1--32. John Benjamins, Amsterdam & Philadelphia (2008)
8. Marcus, M., Santorini, B., Marcinkiewicz, M.: Building a Large Annotated Corpus of English: The Penn Treebank. Computational Linguistics 19(2), 313--330 (1993)
9. Pitler, E., Raghupathy, M., Mehta, H., Nenkova, A., Lee, A., Joshi, A.: Easily Identifiable Discourse Relations. COLING 2008 (2008)