ParaBLoCC: Parallel Basic Locative Constructions Corpus

Peter Viechnicki and Anthony Kostacos
Johns Hopkins University
{pviechn1,akostac1}@jh.edu

Abstract

We introduce ParaBLoCC, the Parallel Basic Locative Construction Corpus, the first multilingual compendium of this important grammatico-functional construction, and particularly the first such corpus containing semantically equivalent BLCs in source/target language pairs. The data —taken from bitext corpora in English paired with twenty-six typologically diverse languages —are likely to prove useful for studying questions of cognitive underpinnings and cross-linguistic usage patterns of spatial expressions, as well as for improving multilingual spatial relation extraction and related tasks. The data are being made available at https://github.com/pviechnicki/parablocc.

1 Introduction: Definition and Importance of Basic Locative Constructions

Basic Locative Constructions (BLCs) are a sentence type identified through the specific pairing of syntactic form and usage purpose (Sadock and Zwicky, 1985). BLCs —identified notably by Levinson and Wilkins (2006) —are statements used to answer questions of the form, *Where is the figure object in relation to the ground object?*[1] Cross-linguistically, BLCs are taken to be ubiquitous – no languages have been reported which are unable to answer such a question. Languages do vary in choice of syntactic forms used to express BLCs (Fortis, 2010). In English, canonical syntax for BLCs is `[NP Copula PP]`: 'The figure object is over/on/under/behind the ground object.' In other languages, BLC syntactic form may be very different, for example in KwaKwala, BLCs are expressed through locative suffixes (Rosenblum, 2015).

2 Importance of BLCs for Cognitive Science and Linguistics

BLCs have been considered important tools for several decades by cognitive scientists who have used them to elicit cross-linguistic properties of spatial expressions. The prominence of BLCs in studies of spatial cognition was enabled by Bowerman and Pederson's (1992) Topological Relations Picture Series, a set of 71 spatial scene cartoons, each depicting a spatial relation between a figure and ground. A number of studies have used controlled elicitation with BLCs to shed light on psycholinguistic topics: for example BLCs have been used to explore core versus peripheral spatial references (Landau et al., 2016); evidence from spatial relations for the Sapir-Whorf hypothesis (Tseng et al., 2016); language acquisition patterns in infancy (Lakusta et al., 2021); and supposed 'natural concepts' in the spatial domain (Levinson and Meira, 2003).

Relatively few studies have looked at usage patterns of BLCs in uncontrolled settings; to our knowledge only (Viechnicki et al., 2024) have done so. The ParaBLoCC corpus aims to allow such work to proceed, by making available a large corpus of English BLCs paired with parallel text from a typologically diverse set of twenty-six languages. The data are publicly available at https://github.com/pviechnicki/parablocc.

3 Related Research

BLCs differ from two closely related expression types: geospatial expressions and spatial relation triples, both of which have more extensive corpora available. Geospatial expressions in text, which have been studied in the context of georeferencing techniques, are commonly defined as spatial relations whose ground object is located out of doors and is not mobile, and whose spatial relation is expressed within a geospatial coordinate

[1] We follow (Talmy, 1983) in referring to *figure* and *ground;* other terms for the same concepts are *theme* and *relatum,* or *trajector* and *landmark.*

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 18–23
May 3, 2025 ©2025 Association for Computational Linguistics

reference system (Stock et al., 2021). BLCs, by contrast, may reference ground objects of any size, interior/exterior status, and mobility. Spatial relation triples – often discussed in the context of techniques for extracting such relations from text, e.g. (McNamee et al., 2020); (Hassani and Lee, 2017) – are a superset of BLCs. Spatial relation triples include both locative and path expressions, whereas BLCs are restricted to static constructions. Spatial relations also include a wide variety of syntactic forms in whichever language is being studied, whereas BLCs are typically restricted to a single canonical syntactic form, such as `[NP Copula PP]` in English.

Our work in extracting a parallel corpus of BLCs is similar in spirit to other recent efforts to use web-scale usage data to inform theoretical linguistic or psychological research. For example, Hale and Stanojevic (2024) use data from five languages to investigate syntactic universals; and Beekhuizen et al. (2017) use parallel usage patterns from thirty languages to study cognitive properties of indefinite pronouns. This work is therefore part of the larger trend that has been called the 'quantitative turn' in linguistic research (Kortmann, 2021).

4 ParaBLoCC Corpus Characteristics and Data Preparation

4.1 Corpus Characteristics

Data in the ParaBLoCC corpus comprise parallel English and target-language sentence pairs ('bitext') from twenty-six languages. The twenty-six languages were chosen to maximize genetic and areal diversity as well as availability of bitext material. Bitext sentences are harvested from the Opus Machine Translation Portal (Tiedemann et al., 2023), and similar sources, from a wide variety of domains. Numbers of parallel sentence pairs for each language plus domains are shown in Table 1. ParaBLoCC thus contains paired BLCs in English and one of the target languages, for example:

> EN: 'He is still in Serbia.'::HU: 'Szerbiában maradt.'

Table 1: Languages, domains, bitext pairs, and BLCs occurring (plus rate per 1000 bitext sentence pairs) in ParaBLoCC corpus. Domains: a: Bible-UEDIN; d: QED; e: TEDTalks2020; f: Bible-Literal; g: GlobalVoices; h: OpenSubtitles; j: Europarl; k: UN V1 16; m: IWSLT2016; n: Flores200; o: NLLBv1; p: GoURMET; q: CCaligned; s: SETTIME2; t: Tico19; u: Tanzil; v: ntrex128.

Language	Domains	Bitext Pairs (m)	BLCs (k) (*per 1000*)
Bantu			
Swahili	d,e,g,n,o,q,t,v	21.2	135 *(6.4)*
Finno-Ugric			
Finnish	a,d,e,h,j	25.9	132 *(5.1)*
Hungarian	a,d,e,g,h,j	38.4	169 *(4.4)*
Indo-European			
Catalan	d,e,f,g,h,i	7.6	81 *(10.7)*
Czech	a,d,e,h,j	35.5	178 *(5.0)*
Dutch	a,d,e,g,h, j,n,o,q,v	157.5	1924 *(12.2)*
French	a,d,e,g,h,m	36.6	197 *(5.4)*
German	a,d,e,g,h,j	20.0	101 *(5.1)*
Greek	a,d,e,g,h, j,n,o,q,s,v	99.8	740 *(7.4)*
Italian	a,d,e,g,h,j	32.0	159 *(5.0)*
Polish	a,d,e,g,h, j,n,o,v	129.3	1211 *(9.4)*
Russian	a,d,e,g,h,k	35.3	165 *(4.7)*
Spanish	a,d,e,g,h,j, k	55.5	266 *(4.8)*
Swedish	a,d,g,h,j	15.6	80 *(5.2)*
Niger-Congo			
Igbo	d,e,f,n,o,q,v	5.6	32 *(5.7)*
Other, Isolate			
Japanese	d,e,h,n,o,q,v	66.5	511 *(7.7)*
Korean	a,d,e,g,h,n,o,q	28.9	204 *(7.1)*
Quechumaran			
Aymara	d,g,o,q	1.0	9 *(8.9)*
Quechua	d,o	2.2	23 *(10.7)*
Semitic			
Amharic	a,d,g,n,o,p,t	15.7	147 *(9.4)*
Arabic	a,d,e,g,h,k	39.7	149 *(3.8)*
Hebrew	a,d,e,g,h,n,q	34.6	150 *(4.3)*
Tigrinya	d,n,o,q,t	1.1	6 *(5.9)*
Sino-Tibetan			
Chinese	a,d,e,g,h, k,n,o,q,t,v	64.6	926 *(14.3)*
Turkic			
Turkish	a,d,e,g,h, o,p,q,s	99.0	825 *(8.4)*
Uzbek	d,e,n,o,u,v	28.9	245 *(8.5)*

4.2 Parallel BLC Data Preparation

BLCs are selected from the available bitext for each language using a three-stage filtration procedure: 1. lexical filtering, followed by 2. syntactic filtering, followed by 3. spatial sense filtering.

The lexical filter selects sentence pairs whose English sentence contains a locative spatial expression drawn from a reference list of fifty expressions: twenty-nine common English locative spatial prepositions ('above', 'between', 'on', etc) and twenty-one spatial nominals ('in back of', e.g.). Our reference list contains all non-archaic forms from The Preposition Project (Litkowski and Hargraves, 2007), plus spatial nominals. See Appendix A for the complete list. Data files in the ParaBLoCC archive record which lexical item matched each English sentence.

The syntactic filter selects parallel sentences whose English dependency parse structurally matches one of the syntactic parse templates found in English Basic Locative Constructions. We parse the bitext and the BLC templates with the Stanford Core NLP parser (Manning et al., 2014), then perform subgraph matching through depth-first search. In practice, all syntactic patterns for BLCs with the fifty spatial expressions can be expressed using eight unique dependency parse subgraphs. ParaBLoCC data files annotate each bitext sentence with the number of the matching spatial expression subgraph template.

The third and final filtration stage selects only sentences whose lexical match from the first filtration state has a spatial sense in context, vice a temporal or other sense. Many of the lexical items from our reference list are highly polysemous – in fact at least twenty common English prepositions have six or more spatial and non-spatial senses (Litkowski and Hargraves, 2021). We train a 'glossbert'-style neural word sense disambiguation model (Huang et al., 2019) as a binary classifier and infer spatial/non-spatial sense for each ParaBLoCC English sentence. Model architecture is shown in Figure 1. The spatial sense classifier is trained with 8,111 sentences exemplifying the senses extracted from The Preposition Project dictionaries (Litkowski and Hargraves, 2021). The model uses the ADAM optimizer, batch size of 16, and is trained for 10 epochs.

4.3 Spatial Sense Filter Performance

We assess the performance of the spatial sense filter using held-out validation data from the aforecited Preposition Project and 200 hand-labeled in-domain sentence pairs (Table 2).

While performance of the spatial filter is not as high as state-of-the-art word sense disambiguation (WSD) models tested on less challenging test sets (Yigzaw and Assefa, 2024), we consider spatial sense disambiguation problem to be particularly challenging and the current model to be adequate for the large-scale filtration task at hand, while recognizing the challenges described in §5.

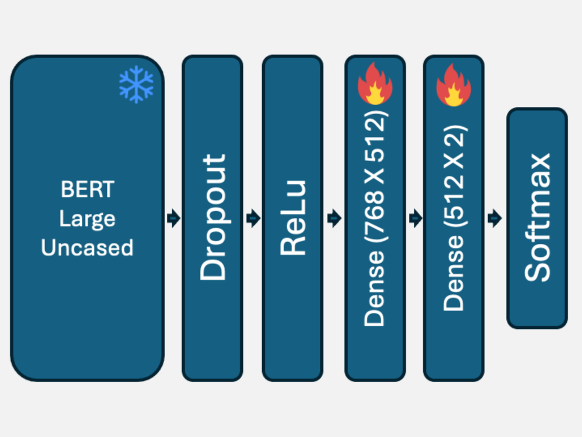

Figure 1: Model architecture for spatial sense disambiguation classifier, based on (Huang et al., 2019).

Table 2: Spatial Sense Classifier Performance: recall, precision, and macro-averaged F1.

	Precision	Recall	F1
Validation Set	.69	.70	.66
In-Domain Set	.56	.80	.66

4.4 BLC Occurrrence Rates

The mean observed rate of BLC occurrence per thousand original sentences, taken across ParaBLoCC languages, is 7.2 (see Table 1). While domain differences may explain some extreme values, we continue to investigate outliers to rule out processing errors resulting in artificially low or high values. Extrapolating from our small set of hand-labeled validation sentences, we believe the ParaBLoCC estimate of 7.2/1000 is three times less than the true BLC occurrence rate, measured at twenty per thousand. See §5 for discussion of the reasons for the underestimate.

5 Sources of Error in ParaBLoCC

The ParaBLoCC data contain noise from two main sources. First due to errors in harvesting target-language sentences from the web, some ParaBLoCC entries will contain target sentences which are not exact semantic equivalents of the source BLC, or which contain other types of bitext alignment errors. In spite of improvements in identifi-

Table 3: BLC Detection Error Analysis and Error Modes

Metric	Value
BLC Detection Rate	.095
False Positive Rate	.016
False Positives	.016
Syntax Errors	-
Spatial Sense Errors	.016
False Negatives	.018
Syntax Errors	.012
Spatial Sense Errors	.006

cation of parallel text for harvest (cf. Paracrawl, (Bañón et al., 2020)), source-target sentence pair mismatches are common in the corpora from which ParaBLoCC draws: recent estimates of error rates in bitext corpora vary from as low as 24% to as high as 76% of sentence pairs (Kreutzer et al., 2022).

A second source of errors in the ParaBLoCC entries comes from BLC detection errors, either Type 1 (false positives) or Type II (false negatives). Those errors in turn can be grouped into errors from the syntactic matching filter and errors from the spatial sense disambiguation filter. (It is assumed that lexical matching errors are negligible, since matching is deterministic.)

To assess the accuracy and sensitivity of the BLC labels in the ParaBLoCC corpus, we used a hand-labeled reference set of 1,000 ParaBLoCC sentences which passed the lexical filter, from sixteen of the included languages.[2] The authors independently coded the sentences and discarded any where we did not agree. Inter-annotator agreement was $\kappa = .55$ —in the 'moderate' range. BLC detection error rate and Type I and II errors are reported in Table 3. The observed BLC detection rate in ParaBLoCC is estimated at .095, quite low with a balance of false positives and false negatives. The false positive rate is .016. Post-hoc analysis of error modes shows that the spatial sense filter did not perform well on longer sentences, particularly those with multiple clauses, which are common in the ParaBLoCC corpus. We leave improvements to the spatial sense filter for future work.

6 Likely uses of ParaBLoCC

We created the ParaBLoCC data to appeal to a wide variety of scholars interested in spatial language, and by making them available we hope to encourage additional study in this area. The primary utility of the data are to allow study of usage patterns for parallel spatial expressions in twenty-six genetically and typologically diverse languages. Through automated alignment and span detection, silver labels for BLCs in the target languages can be extracted and studied themselves or used for downstream tasks.

Likely secondary uses for the ParaBLoCC data will be to enable work on multilingual aspects of spatial relation extraction (Rawsthorne et al., 2023). Until very recently, text corpora annotated for spatial relation triples were limited to the most high-resource numbers of languages, though this situation is starting to improve (Wang et al., 2023) so the multilinguality of ParaBLoCC should be welcome. The data can be used to improve current models of geospatial expression resolution (Wang et al., 2024). Finally we expect multilingual image caption models (Ramos et al., 2023) will benefit from the parallel data collected by ParaBLoCC.

7 Limitations

We acknowledge several limitations of the ParaBLoCC corpus. The selection of languages is limited to those with adequate bitext availability. In practice, this limits us from collecting BLCs in languages whose spatial expression systems are most formally distinct from English and European languages. For example, languages with only absolute reference frames, lacking intrinsic or relative frames (Fortis, 2010), are conspicuously absent from ParaBLoCC.

Granularity of annotation is another limitation of ParaBLoCC. Because of the method of collecting and labeling the sentence pairs in the corpus, text spans representing BLCs are not overtly annotated in either source language (English) or target language. Explicit span annotations for BLCs would provide additional training and test data veracity. Furthermore, ParaBLoCC would ideally provide semantic role annotations for sub-spans of source and target-language BLCs as `<figure>`, `<ground>`, and `<spatial relation>`. While stochastic methods of labeling subspans of BLCs have been demonstrated (Viechnicki et al., 2024), they are noisy. Explicit annotation of this nature would allow more in-depth analysis of the kinds of syntactic variation found in BLCs 'in the wild.' We leave annotation improvements to future community efforts.

[2] am, ar, ay, cs, de, es, fi, hu, ig, ko, nl, qu, sw, ti, uz, zh

8 References

References

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

B. Beekhuizen, J. Watson, and S. Stevenson. 2017. Semantic typology and parallel corpora: Something about indefinite pronouns. *CogSci*.

M. Bowerman and T. Pederson. 1992. *Topological relations picture series*, chapter 1.2. Max Planck Institute for Psycholinguistics, Nijmegen.

Jean-Michel Fortis. 2010. Space in language. *Leipzig Summer School 2010 Part 1*.

John T. Hale and Miloš Stanojević. 2024. Do LLMs learn a true syntactic universal? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17106–17119, Miami, Florida, USA. Association for Computational Linguistics.

K. Hassani and W. Lee. 2017. Disambiguating spatial prepositions using deep convolutional networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. GlossBERT: BERT for word sense disambiguation with gloss knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3509–3514, Hong Kong, China. Association for Computational Linguistics.

B Kortmann. 2021. Reflecting on the quantitative turn in linguistics. *Linguistics*, 59:1207–1226.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

L. Lakusta, Y. Hussein, A. Wodzinski, and B. Landau. 2021. The privileging of 'support-from-below' in early spatial language acquisition. *Infant Behav Dev.*

B. Landau, K. Johannes, D. Skordos, and A. Papafragou. 2016. Containment and support: Core and complexity in spatial language learning. *Cognitive Science: A Multidisciplinary Journal.*

S. Levinson and S. Meira. 2003. 'natural concepts' in the spatial topologial domain–adpositional meanings in crosslinguistic perspective: An exercise in semantic typology. *Language*, 79:485–516.

S Levinson and D. Wilkins. 2006. *Grammars of Space: Explorations in Cognitive Diversity*. Language Culture and Cognition. Cambridge University Press.

Ken Litkowski and Orin Hargraves. 2021. The preposition project. *Preprint*, arXiv:2104.08922.

Kenneth C. Litkowski and Orin Hargraves. 2007. SemEval-2007 task 06: Word-sense disambiguation of prepositions. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 24–29, Prague, Czech Republic. Association for Computational Linguistics.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Paul McNamee, James Mayfield, Cash Costello, Caitlyn Bishop, and Shelby Anderson. 2020. Tagging location phrases in text. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4521–4528, Marseille, France. European Language Resources Association.

Rita Ramos, Bruno Martins, and Desmond Elliott. 2023. Lmcap: Few-shot multilingual image captioning by retrieval augmented language model prompting. *Preprint*, arXiv:2305.19821.

Helen Mair Rawsthorne, Nathalie Abadie, Eric Kergosien, Cécile Duchêne, and Éric Saux. 2023. Automatic nested spatial entity and spatial relation extraction from text for knowledge graph creation: A baseline approach and a benchmark dataset. In *Proceedings of the 7th ACM SIGSPATIAL International Workshop on Geospatial Humanities*, GeoHumanities '23, page 21–30, New York, NY, USA. Association for Computing Machinery.

Daisy Rosenblum. 2015. *A grammar of space in Kwakwala*. Ph.D. thesis, University of California Santa Barbara.

Jerrold Sadock and Arnold Zwicky. 1985. Speech act distinctions in syntax. In Tim Shopen, editor, *Language, Typology, and Syntactic Description*, volume 1, pages 155–196. Cambridge University Press, Cambridge.

K. Stock, C. Jones, S. Russell, M. Radke, P. Das, and N. Aflaki. 2021. Detecting geospatial location descriptions in natural language text. *International Journal of Geographical Information Science*, 36(3).

Leonard Talmy. 1983. How language structures space. In Herbert L. Pick and Linda P. Acredolo, editors, *Spatial Orientation: Theory, Research, and Application*, pages 225–282. Springer US, Boston, MA.

Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato, Yves Scherrer, Raul Vazquez, and Sami Virpioja. 2023. Democratizing neural machine translation with opus-mt. *Preprint*, arXiv:2212.01936.

Christine Tseng, Alexandra Carstensen, Terry Regier, and Yang Xu. 2016. A computational investigation of the sapir-whorf hypothesis: The case of spatial relations. In *Proceedings of the 38th Annual Meeting of the Cognitive Science Society, CogSci 2016*, Proceedings of the 38th Annual Meeting of the Cognitive Science Society, CogSci 2016, pages 2231–2236. The Cognitive Science Society. Publisher Copyright: © 2016 Proceedings of the 38th Annual Meeting of the Cognitive Science Society, CogSci 2016. All rights reserved.; 38th Annual Meeting of the Cognitive Science Society: Recognizing and Representing Events, CogSci 2016 ; Conference date: 10-08-2016 Through 13-08-2016.

Peter Viechnicki, Kevin Duh, Anthony Kostacos, and Barbara Landau. 2024. Large-scale bitext corpora provide new evidence for cognitive representations of spatial terms. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1089–1099, St. Julian's, Malta. Association for Computational Linguistics.

W. Wang, J. Li, W. Ku, and H. Wang. 2024. Multilingual spatial domain natural language interface to databases. *GeoInformatica*.

Zixiang Wang, Jian Yang, Tongliang Li, Jiaheng Liu, Ying Mo, Jiaqi Bai, Longtao He, and Zhoujun Li. 2023. Multilingual entity and relation extraction from unified to language-specific training. *Preprint*, arXiv:2301.04434.

R. Yigzaw and B. Assefa. 2024. State-of-the-art approaches to word sense disambiguation: A multilingual investigation. *Communications in Computer and Information Science*.

A Appendix: Prepositions and Spatial Nominals used in Lexical Filtration Step

English Spatial Expressions used in lexical filtration stage.

A.1 Prepositions

`above, across, against, along, alongside, amid, amidst, among, around, at, atop, before, behind, below, beneath, between, by, down, in, inside, near, off, on, over, toward, towards, under, underneath, with`

A.2 Spatial Nominals and Spatial Phrases

`out of, in back of, in the front of, on the top of, on top of, in front of, to the right of, to the left of, right of, left of, north of, south of, east of, west of, in the middle of, on the bottom of, next to, outside of, in the back of, on the left of, on the right of`