Large-Scale Bitext Corpora Provide New Evidence for Cognitive Representations of Spatial Terms

Peter Viechnicki[1], **Kevin Duh**[1], **Anthony Kostacos**[2], **Barbara Landau**[2]
[1]Human Language Technology Center of Excellence [2]Department of Cognitive Science
Johns Hopkins University, Baltimore, MD, USA
pviechn1@jhu.edu, kevinduh@cs.jhu.edu, akostac1@jhu.edu,landau@jhu.edu

Abstract

Recent evidence from cognitive science suggests that there exist two classes of cognitive representations within the spatial terms of a language, one represented geometrically (e.g., *above*, *below*) and the other functionally (e.g., *on*, *in*). It has been hypothesized that geometric terms are more constrained and are mastered relatively early in language learning, whereas functional terms are less constrained and are mastered over longer time periods (Landau, 2016). One consequence of this hypothesis is that these two classes should exhibit different cross-linguistic variability, which is supported by human elicitation studies.

In this work we present to our knowledge the first *corpus-based* empirical test of this hypothesis. We develop a pipeline for extracting, isolating, and aligning spatial terms in basic locative constructions from parallel text. Using Shannon entropy to measure the variability of spatial term use across eight languages, we find supporting evidence that variability in functional terms differs significantly from that of geometric terms. We also perform latent variable modeling and find support for the division of spatial terms into geometric and functional classes.

1 Motivation

Understanding the cognitive structures underpinning spatial terms has been an object of inquiry within the broad tradition of the cognitive sciences, e.g. Jackendoff (1983); Talmy (1983); Miller and Johnson-Laird (1976); Landau and Jackendoff (1993); Bloom et al. (1996); Levinson and Wilkins (2006). One key issue concerns the range of spatial relationships that are in fact encoded in the class of spatial terms across languages. There are two ways of framing this question. Scientists who emphasize the universal aspects of spatial language have focused on the idea that non-linguistic spatial representations (which are presumably universal, e.g. containment, support, direction) must provide universal constraints on the spatial properties that are encoded across languages (Landau and Jackendoff, 1993). By contrast, scientists who emphasize cross-linguistic variation across spatial terms focus on the fact that there is substantial variation across languages even in apparently simple domains such as containment, support or direction (see, e.g. Levinson and Wilkins (2006); Bowerman (1996)).

	Sample Spatial Terms	
lang	geometric	functional
EN	*above, below, right...*	*in, on, over...*
FR	*à gauche, à droite...*	*sur, sous, dans...*
FA	... پُشتِ، زِرِ	... دَر، بَر

Table 1: Research Question - Do functional terms have more cross-linguistic variability than geometric terms in corpora, supporting results from cognitive science?

In general, theories and evidence on the issue of universals vs. variation in spatial language have spanned quite different sets of spatial terms and their cross-linguistic equivalents, making broad generalizations across different sets of spatial terms difficult. But some of this debate may be resolved by considering that the answer might be somewhat different in different sub-domains of spatial terms. In this paper, we test a hypothesis that could begin to differentiate between such different subdomains, asking whether there are different patterns of variability across 'geometric' vs. 'functional' spatial terms.

Specifically, some theorists have posited that all spatial terms should be in principle represented as 'geometric', that is, in terms of vectors and their direction (O'Keefe and Burgess, 1996). However, many linguists have argued that the true underlying representation of terms in the domain of containment/support must involve force-dynamic relationships between a target and reference object (Vandeloise, 1991; Coventry and Mather, 2002; Carlson and van der Zee, 2005). That is, for something to be

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 1089–1099

'contained' within an object depends on so-called 'functional' properties, and not simply geometry. Examples abound: flowers 'in' a vase can protrude with most of the flower outside of the vase; a fly 'on' a wall is supported not by simple position or even gravitation, but by force-dynamics between the wall and the fly's foot adhesive pads.

Landau (2016) has built on a broad range of evidence to propose that the spatial terms widely used to examine universal vs. language-specific contributions – both 'geometric' and 'functional'–may have quite different profiles for acquisition, cultural conditioning, cross-linguistic variability, and even neural representation. The differing profiles imply that there should be greater variability in the uses of functional terms across languages than of the geometric terms.

Landau (2016) further argues that the geometric terms will naturally vary only on the choice of reference system relevant for a given term (e.g. for 'above/below', a reference system in which 'above' is represented as lying along the vertical axis centered on a reference object in the upward direction). The choices are relatively few: the reference system could be centered on an object, person, scene for terms 'above/below' but must be centered on the earth for 'north/south/east/west'. By contrast, the dimensions that are relevant for functional terms will be much more numerous and culturally-conditioned. The reference object appropriate for use of 'in' may be concrete or abstract but might also vary by culture/ language. Although 'bird in a tree' is natural to native English speakers, it is not natural to speakers of other languages, for whom trees cannot naturally be conceived of as 'containers' (Munnich and Landau, 2010). Thus, it is predicted that there should be greater variability in the uses of functional terms like 'in' across languages than of the geometric terms like 'above'.

Here we pose this question in a wholly new context, in which we are able to examine variability of these two sets of terms across languages using large-scale corpora. The availability of large-scale corpora of translation pairs of sentences offers the possibility of verifying this claim empirically. We now have parallel text corpora wherein we see the linguistic expression of the same semantic structure in multiple different language pairs, allowing us to observe variability in the expression of spatial terms. Our research question, as illustrated in Table 1, is this: Do functional spatial terms exhibit more variability than geometric spatial terms in cross-language corpora for languages such as French (Vandeloise, 1991) and Farsi (Moltaji, 2016)? In other words, do corpus statistics support previous cognitive science studies?

In the remainder of the paper we first review related work investigating cognitive representations of spatial terms. Next we present our method for isolating and analyzing the cross-linguistic equivalents of those terms. Then we present results of our experiments which provide support for the two hypothesized classes and significant differences in variability for functional vs. geometric terms. Finally, we review some of the limitations of our work and how they might be overcome in future studies.

2 Relation to Other Work

Since this work uses computational linguistic techniques in order to provide evidence for a question of cognitive science, it necessarily falls at the intersection of several related sub-disciplines and lines of inquiry. Many cognitive scientists have used experimental techniques in which native speakers of various languages are asked to describe pictures portraying different kinds of spatial relationships. The goal of such studies is to elicit a canonical production of a spatial expression in a constrained setting, to allow cross-linguistic comparison (Levinson and Wilkins, 2006; Bowerman, 1996). This method differs from our current work, in which we deliberately attempt to capture variation between and within speakers of a language by observing multiple target-language usage patterns, all parallel to a particular spatial term in the source language.

A second body of work investigates the structural properties of systems of spatial terms across many languages, developing models of partitions of semantic types (Levinson and Meira, 2003; Khetarpal et al., 2013). By contrast, our work investigates the cross-linguistic correspondences of the tokens of those types within a large-scale parallel text corpus, but without any reference to external representations of spatial arrays.

Building on the observation that spatial terms typically express a core sense which refers to relations between objects in the physical world, but also secondary meanings referring to temporal and other more abstract relations, a third body of research has attempted to build word-sense disambiguation tools to distinguish between spatial and

non-spatial uses of said terms (Hassani and Lee, 2017). Such work has required annotating corpora of text for location phrases, necessitating lists of spatial terms and detailed annotation guidelines (Litkowski and Hargraves, 2007; McNamee et al., 2020). This body of work is similar to ours in that its models learn from usage patterns of spatial terms within a particular language.

Most closely related to our work is a series of studies by Beekhuizen, Stevenson, and colleagues (see e.g. Beekhuizen and Stevenson (2015)), which exploits crowdsourced data and parallel text such as the Bible to understand the cognitive properties of spatial concepts. In contrast to our work, they focus on the interaction between static/dynamic and support/containment spatial markings.

Generally, our work falls within the broader tradition of using multilingual text resources to investigate cognitive science questions: besides spatial terms, examples include the study of color terms (McCarthy et al., 2019), kinship terms (Khalilia et al., 2023), pain predicates (Reznikova et al., 2012), indefinite pronouns (Beekhuizen et al., 2017), and motion verbs (Wälchli and Cysouw, 2012).

3 Methods: Spatial Term Equivalence

Our goal is to extract Basic Locative Constructions (BLCs, i.e. the answer to the question *Where is the object?*) from large-scale bilingual corpora in order to measure the variability in usage. For example, how many different terms are used in French for the concept equivalent to spatial term *above* in English? This requires two things: first, we need to extract BLCs containing spatial terms of interest (e.g. *The urn is above your fireplace*). Second, we need to align French terms corresponding to the identified English terms in pairs of BLCs that are translations of each other. The first requirement is non-trivial to do automatically because non-spatial and metaphorical usages are prevalent in standard usage: in the case of *'above'*, e.g. 'above average profits', 'order from above', 'above the rule of law.'

We propose a pipeline approach to carry out such measurements, as shown in Figure 1. The goal of the pipeline is, from parallel corpora, to filter BLCs – argued to be the clearest contexts for revealing cognitive differences among spatial terms (Levinson and Wilkins, 1999). The pipeline is optimized to run on a research compute cluster using parallelized CPU operations over large-scale parallel text corpora. The pipeline is applicable to any bitext corpus for any pair of source and target languages. Below we describe the components of the pipeline and their interactions.

Preliminaries: Spatial terms – sometimes called Topological Relation Markers (TRMs) (Levinson and Meira, 2003)– consist of one or more morphemes, lexical items, or combinations of these expressing a spatial relationship between objects. A distinction can be drawn between simple spatial terms (often closed-class adpositions or morphemes) and compound, or phrasal spatial terms, including spatial nominals.[1] Given a semantic spatial relation S, we denote each of the k possible types of the expression of S in language $L1$ as $S_k^{L1}, k \in 1, \ldots, K$. Given a pair of parallel sentences containing S_k^{L1} in L1, we can observe the equivalent j realizations in $L2$ as S_{jk}^{L2}.

By observing a number of such realizations we can count the cooccurrence frequency of S_k^{L1} with individual types of S_{jk}^{L2}, and thereby measure the cross-linguistic variability. The cross-linguistic variability we seek to measure involves synchronic usage patterns in specific languages, and does not directly address language change.

Filtering Stages: In the first stage we apply consistent tokenization to the sentences in L1 and L2 of the parallel text corpus, and save those token sequences so all remaining stages can access them as needed (see §4 for details.)

After tokenizing, we apply string search using a spatial term reference file over the L1 token sequences to filter sentences containing a spatial terms such as *'above'* or *'on'*.

We then select English sentences whose syntax matches that of the English basic locative constructions. The pipeline performs syntactic filtration by applying dependency parsing to the source language sentence token sequences selected at the previous stage, then searching each dependency parse graph for specific node patterns, to select sentences of the syntactic form in Figure 2.

The next stage of the pipeline filters out sentences whose spatial relation arguments are abstract, because we expect that abstract extensions of spatial terms could introduce noise into our understanding of the cross-linguistic variability of those

[1] Other realizations of TRMs are common in the world's languages, including spatial verbs (Ameka and Levinson, 2007), but are not considered here.

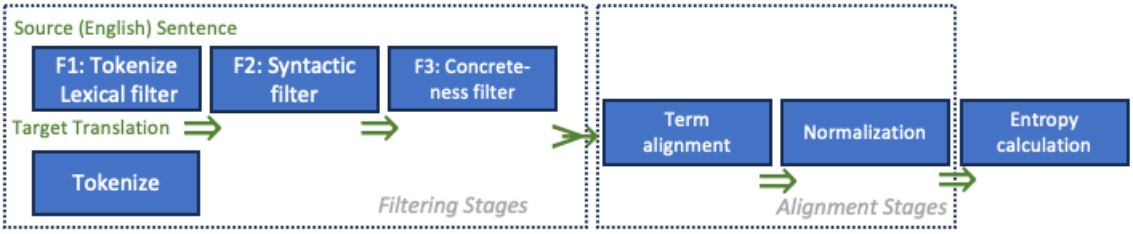

Figure 1: Pipeline for filtering bilingual corpora, aligning spatial terms, and computing variability via entropy.

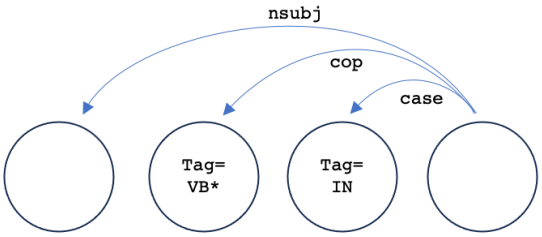

Figure 2: Dependency parsing pattern for filtering basic locative constructions

Example sentence	F1	F2	F3
You're still at work, aren't you?	Y	N	N
Just there's a lot of blood on these sheets.	Y	N	N
Some Kuwaiti monitors and activists were at the protests too...	Y	Y	N
The, uh ... explosive charge was in the receiver itself...	Y	Y	N
He was at a petrol pump and it blew up.	Y	Y	Y

Table 2: Sample sentences passing some or all filtration stages in BLC extraction pipeline; **F1**=lexical filter; **F2**=syntactic filter; **F3**=concreteness filter.

terms. We therefore apply a concreteness classifier for English terms in the form of a multi-layer perceptron trained on human-labeled concreteness judgements (see appendix A for implementation.)

Our use of concreteness as the third filtration criterion in our pipeline (Fig. 1, F3) is a design choice, whose consequences are discussed in §6 and §8 below.

Table 2 shows example English sentences from the bitext corpus which passed or did not pass the various stages of filtration. The syntactic filter is tuned for high precision, and rejects dependency parses which include adjuncts or non-expected structures (as in the first row of Table 2). Results of this design choice are discussed in §6 and §8 below. The concreteness filter rejects sentences unless at least one argument of the spatial relation is categorized as 5 on the five-point concreteness scale described above. For example, the third row of Table 2 shows a sentence that did not pass the concreteness filter because one of the spatial term arguments ('protest') was not categorized as concrete.

Alignment Stages: The next stage of the pipeline aligns the spatial term from L1 to the corresponding token sequence from the target language L2. As discussed in §4 below, we use a standard statistical word alignment package to accomplish this step. The output of this stage is a table of coocurrences of raw target-language spatial types S_{jk}^{L2} with source equivalents $S_k^{L1}, k \in 1, \ldots, K$.

The final stage of the pipeline seeks to minimize noise from orthographic and morphological variation by mapping raw L2 types to canonical forms. For example, the French spatial terms 'au-dessus' and 'au dessus' are in free variation with and without the hyphen (Vandeloise, 1991), but do not convey distinct meanings. We map both to a single canonical form. Similarly, the Greek preposition 'se', corresponding to a range of English prepositions including 'in', 'on', and 'at', appears in contracted form with an inflected following definite article variously as 'sto', 'ston', 'stin', 'sti', etc. We map all such raw types to the single canonical form 'se'. This mapping is currently performed using string substitutions after inspection of the raw spatial term equivalence tables, in consultation with native speaker informants and reference grammars. Consequences and limitations of the normalization are discussed below in §6.

Entropy Calculation: The result after this final processing stage is a cooccurrence matrix of correspondences between $S_{English}$ and S_{L2} spatial terms, for all language pairs in the corpus. We conceive of each column of this matrix as the outcome of a process whereby a speaker of the target language L2 is asked to translate an English sentence,

and selects a fitting spatial term equivalent in the target language. Over a number of trials, then the correspondence between S_k^{L1} and $\{S_{jk}^{L2}\}$ equivalents can be modeled as a discrete random variable with unknown distribution, i.e. we compute probability $p(S_{jk}^{L2})$ as the number of cooccurences between S_{jk}^{L2} and S_k^{L1}, divided by the total count of S_k^{L1}. Then we calculate the Shannon entropy (Shannon, 1948) of this correspondence:

$$H_{S_k^{L1}} = -\sum_j p(S_{jk}^{L2}) \times log(p(S_{jk}^{L2})) \quad (1)$$

Finally we compare $H_{S_k^{L1}}$ of functional terms with that of geometric terms, testing if there is higher variability in one class. In practice, computing Equation 1 directly with plug-in estimators (using the maximum likelihood estimates of probability from raw counts) may lead to negative bias, underestimating the true entropy. So we use the Miller-Madow estimator which adds to Equation 1 a correction term that grows with the number of classes and decreases with the number of samples (Arora et al., 2022). The results from both estimators differ in magnitude but not in overall pattern.

4 Experiment Setup

The proposed method of measuring cross-linguistic variability has been applied to large parallel text corpus of pairs of sentences from English and seven Indo-European plus one Finno-Ugric language: Spanish (ES), Greek (EL), German (DE), French (FR), Dutch (NL), Italian (IT), Farsi (FA), and Hungarian (HU). We chose these languages based on two criteria: (1) the availability of large amounts of data in multiple domains and (2) the availability of language informants to perform the manual normalization step in our pipeline.

For tokenization, we use the Stanford CoreNLP tokenizer when available (English, Spanish, German, and French) and the Moses tokenizer otherwise.[2] For the syntax match component of the pipeline, we use Stanford CoreNLP 4.5.1 (Manning et al., 2014). Specifically, we use the neural network transition-based dependency parser (Chen and Manning, 2014) trained on English University Dependencies.[3] For word alignment, we use `giza++` (Och and Ney, 2003) from the Moses package run up to IBM Model 4.[4]

For the experiments reported here, the list of English spatial terms from the SEMEVAL project (Litkowski and Hargraves, 2007) was used as the starting point, supplemented with common spatial nominals such as 'in front of', and minus any kinetic (path oriented) terms (Levinson and Wilkins, 2006)). Six terms which did not occur frequently enough in the corpus to calculate entropy scores were also dropped, yielding a final reference list of twenty-two English static locative spatial terms (Table 4).

Table 3 summarizes the statistics of our dataset. We begin with millions of sentences pairs from a collection of parallel text corpora obtained via the OPUS portal (Tiedemann, 2012) and obtain approximately tens of thousands of BLCs for each language pair. These BLCs form the basis of our entropy study.

Lang	#Sent	#BLC	Sources
DE	25.3M	30,851	b, e, os, gv, q, t
EL	42.7M	58,581	b, e, os, gv, q, t
ES	65.4M	70,693	b, e, os, gv, q, t, u
FA	7.5M	8,882	tz, os, gv, q, t
FR	62.6M	35,229	b, p, i, os, u
HU	43.9M	59,579	b, e, os, gv, q, t
IT	38.2M	50,839	b, e, os, gv, q, t
NL	40.0M	55,853	b, e, os, gv, q, t

Table 3: Count of sentence pairs in millions in the original corpus (#sent) and count of Basic Locative Constructions (#BLC) after filtering for each language. Keys for sources/domains for bitext corpora: **b**: Bible, **e**: Europarl v7 or v10, **os**: Open Subtitles 2018, **gv**: Global Voices, **q**: QED corpus, **t**: TedTalks 2020, **tz**: Tanzil, **u**: United Nations, **i**: IWSLT 2022.

5 Results

5.1 Do functional terms exhibit more cross-lingual variability than geometric?

To investigate this question, each preposition on the reference list (Table 4) was labeled as either 'functional' or 'geometric' using *a priori* knowledge of linguistic-semantic literature. Entropies per spatial term class were computed next, and are shown in Figure 3.

Mean Miller-Madow entropy for the geometric spatial terms across eight languages was $H = .46$, while for the functional terms mean entropy cross

[2] https://github.com/moses-smt/mosesdecoder/

[3] https://nlp.stanford.edu/software/nndep.html

[4] Though deep neural aligners have become available in recent years (e.g. Dou and Neubig (2021)), our experience is that `giza++` still achieves comparable results on variable-sized corpora from different languages.

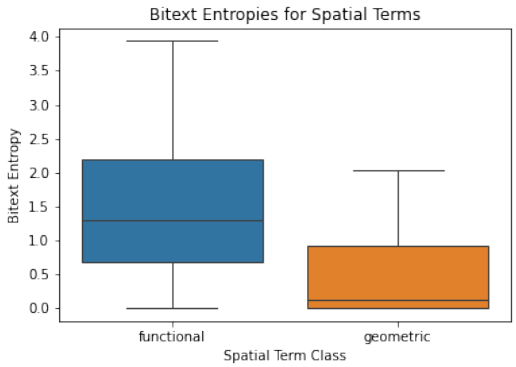

Figure 3: Distribution of entropies with the Miller-Madow estimator for functional and geometric spatial terms defined by *a priori* class labels. Box plot shows median and quartiles.

eight languages was $H = 1.47$. To assess significance we performed a *t*-test for independent distributions, assuming unequal variance: $t = 7.764$, $p < 1.81 \times 10^{-12}$ This result indicates that the means of the geometric and functional classes are significantly different, which we interpret as offering initial support for the hypothesis that functional spatial terms show greater cross-linguistic variability than do geometric ones.

Considering the entropies for functional versus geometric spatial terms derived from *a priori* labels, noteworthy are not just the differences in the means of the two distributions, but also the higher variance among the functional terms. This observed distribution parameter fits the hypothesized properties of functional (force-dynamic) spatial term cognitive representations. Because this class of terms shows a relatively lengthy developmental profile (Landau et al., 2017) and their usage patterns are more culturally conditioned, greater variance in this class makes sense.

5.2 Do entropies for specific spatial terms match expectations?

The class-level box plot covers some complexity in the behavior of individual terms. Table 4 presents entropies for individual spatial terms. The table is sorted by term entropy, low to high. In all cases Miller-Madow estimates are slightly higher in magnitude than plug-in estimates, but do not change their relative rankings. As expected, the putative geometric terms cluster at the top, while the putative functional terms cluster at the bottom.

While 'in' and 'on' have most often been discussed in the context of functional or force-dynamic cognitive representations – and in fact they do display high cross-linguistic variability – the highest variability is from the term 'over.' 'Over' has also been argued to be functionally defined (Coventry and Mather, 2002), and is unusual in having a high degree of polysemy; 'over' conveys three distinct spatial senses including covering, aboveness, and above-acrossness (Brugman and Lakoff, 1988). The degree of polysemy no doubt contributes to the cross-language variability. Word senses disambiguation and role labeling of spatial terms (Kordjamshidi et al., 2010) are potentially useful in obtaining more fine-grained analyses; we leave this as future work.

A few departures from initial expectations in Table 4 are noteworthy. 'Against' is labeled *a priori* as 'functional' because of its requirement for a very specific kind of support from one object relative to the other; note that Levinson and Meira (2003) consider 'against' to be an interstitial and hence unusual English blend somewhere in topological space between the more focal 'on' and 'near.' The term 'down' has a surprisingly high H value, implying non-spatial usages may have muddied the analysis of this particular term. Partial review of the BLCs containing 'down' confirms that metaphorical uses such as 'down the tubes' are included, as

Spatial Term	Term Class	$\bar{H}$	MM
to the left of	G	0	0
to the right of	G	0	0
in back of	G	0	0
in the front of	G	.24	.25
behind	G	.25	.26
between	G	.29	.29
below	G	.37	.39
against	F	.41	.42
on the bottom of	F	.44	.47
under	G	.51	.52
above	G	.53	.55
on the top of	F	.62	.66
in front of	F	.77	.78
inside	F	1.06	.1.08
on top of	F	1.26	1.30
in	F	1.43	1.43
down	G	1.47	1.54
off	F	1.58	1.65
at	F	1.75	1.76
on	F	2.12	2.13
by	F	2.34	2.39
over	F	2.45	2.49

Table 4: Individual Spatial Term Entropies: $\bar{H}$ is the mean of Equation 1 of a spatial term S_k^{L1} over 8 languages using the plug-in estimator; MM is the mean of the Miller-Madow estimate of entropy; Term class = functional (F) or geometric (G) based on evidence from linguistic-semantic literature.

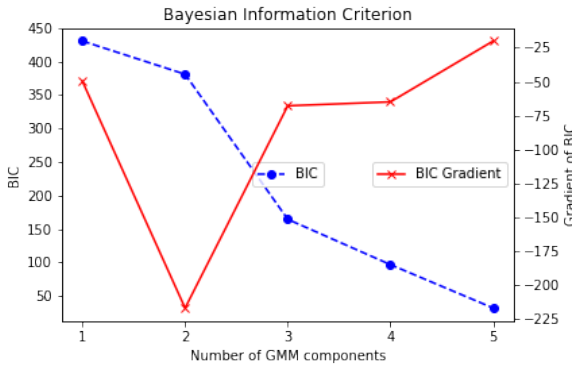

Figure 4: Number of Components in GMM analysis of Spatial Term Entropies

	GMM Labels	
	Low	High
G	above, behind, below, between, in back of, in front of, in the front of, to the left of, to the right of, under	down
F	against, on the bottom of, on the top of	at, by, in, inside, off, on, on top of, over

Table 5: Agreement matrix between a priori spatial term class labels (G=geometric, F=functional) and labels derived from GMM ({High, Low}).

ES	EL	DE	NL	FR	IT	FA	HU
.84	.87	.94	.94	.95	1.00	1.32	1.43

Table 6: Mean Entropy, Miller-Madow estimator, of all Terms by Language

are kinetic usages such as 'I was halfway down the stairs when...'

5.3 Does clustering reveal the same kind of classes?

To check whether similar results could be obtained without assumed spatial term class labels, we use mixture modeling to identify the number and composition of latent components in the entropy data.

First, we investigated how many latent classes could be identified in the 8-dimensional spatial term × language entropy score matrix. Over the course of twenty trials, we estimated Gaussian Mixture Models (GMMs) with number of components varying between one and six. Figure 4 plots the mean Bayesian information criterion (BIC) for each number of components on the left-hand y-axis, and the corresponding gradient of the BIC on the right-hand vertical axis. Lower BIC means a more informative mixture model, and locations of steep gradient BIC are good cut points for number of components (Neath and Cavanaugh, 2012).

Figure 4 shows a clear drop in BIC between one and two mixture components, and a corresponding steep BIC gradient. We take these results to mean that the data are best described as composed of mixtures of two underlying distributions.

We next estimate a two-component GMM, labeling the cluster with higher mean vector as "high" and the other as "low." We repeat this for 20 trials and report the most frequent label for each term. Table 5 shows results for twenty-two spatial terms, and their corresponding *a priori* class labels.

Of the twenty-two spatial terms on the reference list, eighteen (82%) show agreement between the *a priori* class labels and the labels derived organically from mixture models. (These are the terms in the top-left and lower right cells of Table 5.) We interpret this result as suggesting a significant but not perfect overlap between the sets of terms belonging to each class, and the corresponding categories of {Functional, Geometric} as defined by the cognitive science community.[5]

6 Discussion and Analysis

Generally, our corpus-based results provide support for the two classes of spatial language shown in the cognitive science literature. We now turn our discussion to the more fine-grained nuances regarding the findings.

Linking hypothesis to results: Our general hypothesis is that there are differences in cross-lingual variability between different classes of spatial terms. The results in §5.1 show functional terms have significantly different mean entropy compared to geometric terms. The direction of the difference is consistent with the hypothesis that functional terms should exhibit more cross-lingual variability than geometric ones. The larger variance of the functional class also matches the initial prediction.

All results are anchored with the same set of English terms as S_k^{L1} in Equation 1. So when we say the English term 'on' has higher cross-lingual variability than 'below', we are only comparing between terms in the same language (English). Our results say nothing about the inherent variability of words in, e.g., Hungarian vs. German or Hungarian

[5]With regard to mismatches between *a priori* and GMM labels, two of six ('against', and 'on top of') are boundary cases, which likely would appear in the expected *a priori* classes given additional data. See §5.2 for discussion of 'down' and its unexpectedly high variability.

vs. English.[6] One assumption is that the choice of languages in L2 should not impact our comparison of L1 term entropies as long as as the set of L2 languages are held constant in our analysis. This is reasonable, but in future work we would like to confirm with a broader set of L2 languages that are either related and unrelated to L1.

BLC data quantity and quality: The current experiments demonstrate that basic locative constructions are comparatively rare syntactically and semantically. Given that BLCs are the clearest indications of core meaning of static spatial terms in a language, and given that a reliable estimate of cross-linguistic variability requires a certain number of observations of each spatial term in a BLC, it is only because of the size of the available bitext corpora that the current analysis has become possible.

We attempt to characterize our BLC data quality by performing a manual *post-hoc* annotation of random samples of 1,000 sentences before and after filtration. Samples were coded by one of the authors as either containing or not containing a BLC. Among the 1,000 sampled sentences determined to be BLC by our pipeline, 79% are coded as containing a BLC by the human annotator. This gives a precision of 0.79, which we believe is sufficiently high for the entropy studies. Note that it is not easy to estimate recall in this setup; we suspect it is not high, due to the strictness of our syntactic filter.

Two of the design choices in our pipeline (Fig 1) – the precision-tuned syntactic filter and the use of concreteness as a proxy for spatial sense – lead to results which have high precision, but whose recall is low and whose sample size is also low. The small sample sizes made it difficult to estimate H for more rare terms, particularly for some of the geometric terms. We compensated for this issue by using corpora large and diverse enough to provide sufficient estimate of terms in both classes. In future replications of this research, we hope to increase the recall of the pipeline without diminishing precision.

Entropy distribution by language: The distribution of mean H for each language (Table 6) is noteworthy. Miller-Madow estimates of mean entropy by language cluster around $H = .9$. Farsi and Hungarian are clear outliers with mean entropies of $H = 1.32$ and $H = 1.43$ respectively. For Hungarian, high overall entropy would appear to correlate with complexity of inflectional morphology. Hungarian which has the highest mean entropy score of 1.43, also has the most complex morphological system of any of the eight languages (Keresztes, 1995). Because the analytic pipeline is not optimized for automatic morphological parsing, some of the morphemes marking spatial relations in Hungarian likely have not been normalized to canonical forms. Hence there are more L2 spatial term types in the current Hungarian sample, and by implication higher entropy scores.

For Farsi we suspect diglossia as a factor contributing to its higher entropy score of $H = 1.32$. The Farsi data in our sample consist of both Iranian and Afghan Persian, which use different lexical, morphological, and orthographic conventions (Windfuhr, 2009). A lower proportion of written vs. spoken texts in the Farsi sample (see Table 3) may also have contributed to the higher observed entropies.

DE	EL	ES	FA	FR	IT	HU	NL
.89	.72	.84	.92	.90	.70	.68	.85

Table 7: Spearman's ρ rank order correlation of spatial term entropies with and without orthographic and morphological normalization

Impact of Manual Normalization: One bottleneck of our approach is the manual nature of the orthographic and morphological normalization applied at the end of the pipeline before calculating entropy scores. This step was highly labor intensive, required consultation with native speaker informants in some cases, and limited the current analysis to only eight languages. To identify potential bias we performed an ablation study by re-measuring spatial term entropies without the final orthographic and morphological normalization. Specifically we measured the Spearman rank order correlation coefficient ρ in spatial term entropy scores for the eight languages both with and without the normalization applied. A high degree of correlation between the entropy ranks with and without morphological normalization would indicate that the normalization stage is not introducing bias, and potentially could be skipped in future versions of the pipeline. We report the results in Table 7.

Table 7 shows that seven of eight correlation

[6] We also do not answer any questions about the Shannon entropy of running text in different languages, as done in predictive language modeling (e.g. entropy of probability of word3 given word2 and word1).

coefficients are in the high ($>= .7$) or very high ($>= .9$) ranges, based on a common interpretation criteria (Akoglu, 2018). The eighth, Hungarian, is 2 percentage points below the high range. We infer that the orthographic and morphological normalization process did not introduce significant bias into the overall spatial term entropy scores.

7 Conclusion

We find that cross-linguistic variability in spatial term usage is consistent with the hypothesis (Landau 2016) of two distinct cognitive representations of spatial terms: one *functional* (dependent on the force-dynamic interactions between the figure and ground), and the other *geometric* (defined by the distance and direction of the figure from the ground along primary, secondary, or tertiary axes). The current study adds a new type of evidence for the existence of the two distinct classes to prior studies based on child language development patterns and adult elicitation paradigms.

This initial finding would seem to motivate various future investigations. It is desirable to scale up the current analysis and validate it against a larger and more typologically diverse set of languages, requiring automatic morphological parsing to be added to the filtration pipeline. Languages using combinations of morphemes to mark spatial relations will require extension of the current Shannon entropy measure (equation 1), to allow for joint bigram and trigram probabilities in addition to the unigram probabilities.

Lastly the data presented here suggest a possible connection between the grammatical categories used by languages to express the two classes of spatial terms, and the theory of formal markedness (Jakobson). Some languages like English use a single grammatical category (prepositions) to express both putative classes of spatial relations. However other languages such as Hungarian use both case markings and postpositions to express spatial relations. Our data suggest that Hungarian case markings are more closely associated with functional spatial relations, while postpositions are associated with geometric ones. This finding, if extended to other languages with complex nominal and verbal morphological strategies for marking spatial relations, suggest a new way of understanding the range of formal strategies employed by languages to express spatial concepts.

8 Limitations

The first limitation of this work is the small number of languages used, all of which except Hungarian are Indo-European. It is important to verify the conclusions hold given a more typologically and areally diverse sample.

The second limitation is the spatial term coverage in the selected corpora. Though our corpora are large and chosen from diverse domains, certain spatial terms particularly from the geometric class were not well represented. For example, in the English-French bitext corpus of 62 million sentences from diverse genres, no basic locative sentences occurred containing the spatial term 'east of.' Poor coverage of rarer spatial terms disproportionately affects geometric terms, and could bias our results. Directional terms were particularly rare in our sample, and for future analyses may need to be harvested from specialized genres such as travel guides. We hope to increase the recall of our pipeline to reduce potential bias in estimating entropies of comparatively rare terms.

Rare terms could potentially introduce more bias in the entropy estimates compared to frequent terms. This work attempts to mitigate such bias with the Miller-Madow correction. But it is still important to be careful when comparing entropy estimates between words that drastically different occurrences.

A third limitation of this analysis comes from the manual morphological normalization which was implemented as the last stage of the pipeline before entropy estimation. This stage limited the analysis to languages for which we had access to fluent informants and convenient reference grammars. While we show in §6 that this normalization did not introduce bias into our results, we nevertheless hope future iterations of this work will avoid it through automated morphological segmentation for a larger set of languages.

Acknowledgements

The authors gratefully acknowledge the assistance of Istvan Pajor, Nathan Viechnicki, and Mahsa Yarmohammadi for their assistance with Hungarian, Spanish, and Farsi spatial expressions respectively, and technical assistance from David Hunter. We also are grateful for the thoughtful suggestions of several anonymous reviewers.

References

Haldun Akoglu. 2018. User's guide to correlation coefficients. *Turk J Emerg Med.*, 18(3):91–93.

Felix Ameka and Stephen Levinson. 2007. The typology and semantics of locative predicates: posturals, positionals, and other beasts. *Linguistics*, 45-5/6.

Aryaman Arora, Clara Meister, and Ryan Cotterell. 2022. Estimating the entropy of linguistic distributions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 175–195, Dublin, Ireland. Association for Computational Linguistics.

Barend Beekhuizen and Suzanne Stevenson. 2015. Crowdsourcing elicitation data for semantic typologies. In *CogSci*.

Barend Beekhuizen, Julia Watson, and Suzanne Stevenson. 2017. Semantic typology and parallel corpora: Something about indefinite pronouns. In *39th Annual Conference of the Cognitive Science Society (CogSci)*, page 112–117.

Lois Bloom, Cheryl Margulis, Erin Tinker, and Naomi Fujita. 1996. Early conversations and word learning: Contributions from child and adult. *Child Development*, 67(6):3154–3175.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Melissa Bowerman. 1996. The origins of children's spatial semantic categories: Cognitive versus linguistic determinants. In *Rethinking linguistic relativity*, pages 145—-176. Cambridge University Press.

Claudia Brugman and George Lakoff. 1988. Cognitive topology and lexical networks. In *Lexical ambiguity resolution*, pages 477–508. Elsevier.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, 46.

Laura Anne Carlson and Emile van der Zee. 2005. *Functional features in language and space: Insights from perception, categorization, and development*. Oxford University Press.

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar. Association for Computational Linguistics.

Kenny Coventry and Gayna Mather. 2002. The real story of "over"? In *Spatial Language: Cognitive and Computational Perspectives*. Kluwer Academic Publishers.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Kaveh Hassani and Won-Sook Lee. 2017. Disambiguating spatial prepositions using deep convolutional networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Ray Jackendoff. 1983. *Semantics and cognition*. The MIT Press.

Laszlo Keresztes. 1995. *A practical Hungarian grammar*. Debreceni nyari egyetem.

Hadi Khalilia, Gábor Bella, Abed Alhakim Freihat, Shandy Darma, and Fausto Giunchiglia. 2023. Lexical diversity in kinship across languages and dialects. *Front. Psychol.*, 14(2023).

N. Khetarpal, G. Neveu, A. Majid, L. Michael, and T. Regier. 2013. Spatial terms across languages support near-optimal communication: Evidence from peruvian amazonia, and computational analyses. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 35.

Parisa Kordjamshidi, Martijn Van Otterlo, and Marie-Francine Moens. 2010. Spatial role labeling: Task definition and annotation scheme. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Barbara Landau. 2016. Update on "what" and "where" in spatial language: A new division of labor for spatial terms. *Cognitive Science*, 41(S2):321–350.

Barbara Landau and Ray Jackendoff. 1993. "what" and "where" in spatial language and spatial cognition. *Behavioral and Brain Sciences*, 16(2):217–238.

Barbara Landau, Kristen Johannes, Dimitrios Skordos, and Anna Papafragou. 2017. Containment and support: Core and complexity in spatial language learning. *Cognitive Science*, 41(S4):748–779.

S. C. Levinson and D. P. Wilkins. 2006. Grammars of space. *Grammars of space: Explorations in cognitive diversity.*

S.C. Levinson and D. Wilkins. 1999. Hypotheses concerning basic locative constructions and the verbal elements within them. In *Manual for the 1999 Field Season*, page 55–56. Max Planck Institute for Psycholinguistics.

Stephen Levinson and Sergio Meira. 2003. 'Natural concepts'in the spatial topological domain - adpositional meanings in crosslinguistic perspective: An exercise in semantic typology. *Language*, 79(3):485–516.

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 Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Arya D. McCarthy, Winston Wu, Aaron Mueller, William Watson, and David Yarowsky. 2019. Modeling color terminology across thousands of languages. 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 2241–2250, Hong Kong, China. Association for Computational Linguistics.

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.

George A. Miller and P. N. Johnson-Laird. 1976. *Language and perception*. Cambridge University Press.

Niloofar Moltaji. 2016. *An investigation on the locative use of prepositions with comparison to English*. Ph.D. thesis, Department of Linguistics, Stockholm University.

Edward Munnich and Barbara Landau. 2010. Developmental decline in the acquisition of spatial language. *Language Learning and Development*, 6(1):32–59.

Andrew A. Neath and Joseph E. Cavanaugh. 2012. The bayesian information criterion: Background, derivation, and applications. *WIREs Comput. Stat.*, 4(2):199–203.

Franz Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

John O'Keefe and Neil Burgess. 1996. Geometric determinants of the place fields of hippocampal neurons. *Nature*, 381(6581):425–428.

Tatiana Reznikova, Ekaterina Rakhilina, and Anastasia Bonch-Osmolovskaya. 2012. Towards a typology of pain predicates. *Linguistics*, 50(3):421–465.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423.

Leonard Talmy. 1983. How language structures space. In *Spatial Orientation: Theory, Research, and Application*, pages 225–282, Boston, MA. Springer US.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Claude Vandeloise. 1991. *Spatial Prepositions: A Case Study from French*. University of Chicago Press.

Bernhard Wälchli and Michael Cysouw. 2012. Lexical typology through similarity semantics: Toward a semantic map of motion verbs. *Linguistics*, 50-3:671–710.

Gernot Windfuhr. 2009. *The Iranian Languages*. Routledge.

A Appendix: Concreteness Classification

Concreteness classifier architecture and training: To classify English sentences as BLCs, we use a concreteness classifier for the arguments of the spatial relation in the sentence. Specifically, we first train a regression model whose input is a 300-dimensional subword-based FastText word embedding (Bojanowski et al., 2017), with a hidden layer of 100 dimensions, and whose output layer is the concreteness score. The model is trained via L2 loss on data provided by Brysbaert et al. (2014), which includes 14k English nouns rated for concreteness along a 5-point scale by human judges. This model achieves 0.29 mean squared error loss on 20% held-out data. Finally, we threshold the regression model output such that any word with a concreteness score prediction above 4 (in the range 1-5) is determined to be concrete.