Constructing Spatial Concepts from Universal Primitives

Yang Xu* **and Charles Kemp**†
Machine Learning Department*
School of Computer Science*
Department of Psychology†
Carnegie Mellon University
{yx1@cs.cmu.edu, ckemp@cmu.edu}

Abstract

Spatial terms such as *on* and *in* are found in every language, and psychologists have suggested that the meanings of these terms may be constructed from a universal set of spatial primitives. We develop a computational version of this idea and explore whether the primitives typically proposed are sufficient to account for the meanings of spatial terms across languages. We compare a model where spatial terms correspond directly to primitives with models that represent spatial terms as discrete or weighted combinations of primitives. Our results suggest that combinations play an critical role, and we find limited evidence for weighted combinations.

Keywords: spatial cognition; cross-cultural; semantics; computational model.

Every documented language includes some machinery for describing spatial relationships. For example, an English speaker might say that the cup in Figure 1b is *on* the table and that the spoon is *under* the cloth. Spatial terms like these are acquired relatively early by children (Antell & Caron, 1985) and are used so frequently that they may come to seem unremarkable. Researchers have found, however, that it is surprisingly difficult to specify the meanings of spatial terms (Brown, 1994), and that different cultures make use of very different spatial concepts (Levinson & Meira, 2003; Levinson & David, 2006). This paper presents computational models that explore how spatial concepts might be constructed from more basic components, and that help to establish whether spatial concepts across cultures are constructed from a universal set of spatial primitives.

Many previous researchers have discussed the idea that spatial concepts might be constructed as combinations of primitive notions such as "support", "contact" and "containment". (Piaget & Inhelder, 1956; Jackendoff, 1983; Feist, 2000) For example, Figure 1b suggests that *on* in English may be roughly defined as the conjunction of "support" and "contact". Although this basic proposal is very familiar, there have been few sustained attempts to evaluate how well it can account for cross-linguistic data. Here we focus on primitives gathered from the existing literature and ask whether the distinctions that they capture are sufficient to account for spatial concepts across 25 different languages. Future work in this area can compare different sets of candidate primitives and compare how well they account for the data.

Any attempt to study semantic primitives must include some proposal about how these primitives combine to create spatial concepts. Here we compare proposals that vary along three dimensions. One of these dimensions specifies whether combinations of primitives are or are not allowed. A simple baseline approach assumes that every concept in every language corresponds to one of the semantic primitives, and we compare this approach to alternatives which assume that concepts correspond to combinations of primitives. In Figure 1b, for example, "on" is defined as the conjunction of support and contact. A second dimension specifies whether primitives are differentially weighted. In Figure 1b, all combinations are assumed to be conjunctions, and we compare this approach with an alternative that relies on weighted combinations. The final dimension specifies whether or not negations of primitives are allowed—for example, whether "no contact" is included in addition to "contact." Our three dimensions produce a collection of eight possible models, and we explore the five most interesting cases (Table 1). Comparing the performance of these models suggests that combinations of primitives are important, but we find only limited evidence for weighted combinations. None of the models we consider is rich enough to capture the true complexity of spatial cognition, but these simple models are a useful starting point for the computational approach that we advocate.

Our work is inspired in part by several recent studies of cross-cultural spatial cognition (Feist, 2000; Bowerman & Choi, 2001; Levinson & Meira, 2003; Feist, 2008; Khetarpal, Majid, & Regier, 2009). A consistent theme in the previous literature is that spatial concepts correspond to regions in some kind of similarity space. To mention just two examples, Bowerman and Choi (2001) suggest that scenes described using "on" and "in" by English speakers can be arranged along a similarity gradient, and that different languages carve up this similarity space in different ways. Levinson and Meira (2003) propose that spatial terms correspond to attractors in a similarity space, and use multidimensional scaling to support their proposal. Approaches like these have helped to illuminate the basis of spatial cognition, but they rely on a notion of similarity that is rarely made precise, and are unable to explain exactly how humans recognize similarities between spatial configurations. Our work is compatible with many of the insights that have emerged from these previous approaches, and could be viewed as an attempt to ground the notion of similarity in terms of concrete spatial primitives. We prefer, however, to treat similarity as an epiphenomenon, and expect that similarity will play no explanatory role once the building blocks of spatial concepts are understood.

We begin by introducing the semantic primitives that we will consider and the cross-linguistic data that we will attempt to explain. We then evaluate five simple models which

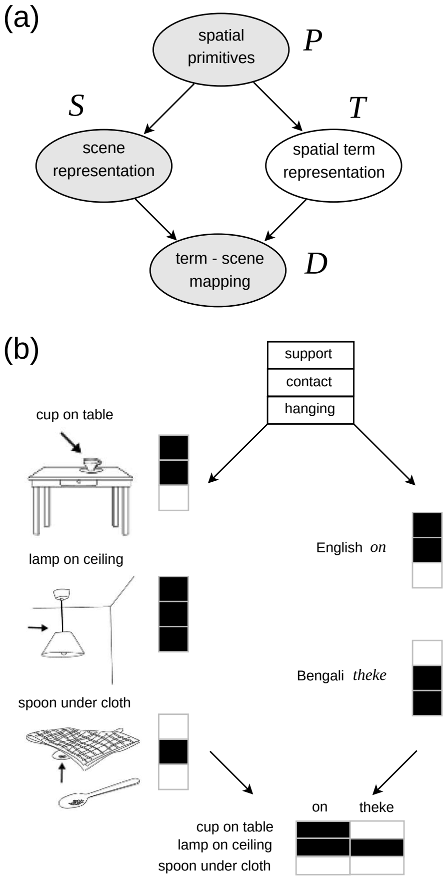

Figure 1: (a) A computational framework for exploring how spatial primitives (P) combine to create the meanings of spatial terms (T). Given information about which primitives characterize a set of scenes (S), the framework predicts which terms apply to which scenes. (b) An illustration of the framework in (a). English "on" is a combination of "support" and "contact," and applies to scenes (like cup on table) where both primitives are present.

make different assumptions about how spatial concepts are constructed from semantic primitives. Each successive model includes one or more previous models as a special case, and we explore whether the additional assumptions made by each model help to account for the cross-linguistic data.

A Computational Approach to Spatial Cognition

Our formal approach is summarized by the graphical model in Figure 1a. Suppose that P represents a set of spatial primitives and that S is a matrix of scene vectors, where column s_i is a binary vector that indicates which primitives apply to scene i. In Figure 1b, for example, the scene vector for "cup on table" indicates that this scene is characterized by "support" and "contact" but not "hanging." Let T be a matrix of term vectors, where vector t_j indicates which primitives contribute to the meaning of term j. In Figure 1b, the term vector for "on" indicates that the meaning of this term is based on the "support" and "contact" primitives. Finally, let D be a binary matrix where entry d_{ij} indicates whether the spatial relationship in scene i can be described by term j.

Table 1: A brief description of the five models and their abbreviations. The two columns on the right compare model scores on the real data to the mean scores on the random sets discussed in Results. D_1 is data from the authors and Levinson and Meira (2003). D_2 is data collected by Feist (2000).

Model	Abbrev.	S(D_1)	S(D_2)
Singleton	BS+	.61 : .39	.61 : .50
Singleton with negations	BS−	.62 : .41	.66 : .53
Conjunction	BC+	.66 : .46	.70 : .58
Conjunction with negations	BC−	.79 : .57	.83 : .68
Weighted combination	WC−	.79 : .54	.80 : .65

The graphical model in Figure 1a can capture at least three kinds of inferences. If asked to decide whether term j applies to scene i, a native speaker can use scene vector s_i and term vector t_j to decide whether $d_{ij} = 1$. When interpreting a description of an unobserved scene i , a native speaker can use term vector t_j along with the information that $d_{ij} = 1$ to predict the scene vector s_i. When learning the meanings of spatial terms, a learner given P, S, and D can infer the term vectors in T. We will address this third problem and the nodes for P, S, and D are shaded in Figure 1a to indicate that these variables are observed for all cases we consider.

We report results for two cross-linguistic data sets. The first is based on a triple (P_1, S_1, D_1) that combines data reported by Levinson and Meira (2003) with new data that we have collected. Our second data set is based on a triple (P_2, S_2, D_2) that is taken from the work of Feist (2000). The next sections describe these triples, and we then describe how we used these triples to explore the meanings of spatial terms.

Spatial primitives. The first set of primitives (P_1) is shown in Table 2, and includes 19 primitives that capture position along the vertical axis, position with respect to the observer, and various notions related to contact and inclusion. These primitives were collected from several previous authors, and the set is intended to capture most of the concepts that have previously been proposed as candidate primitives. The second set of primitives (P_2) is based on a set proposed by Feist (2000), and includes primitives like "above," "contact," and "support." The complete set of primitives is shown at the top left of Figure 2b.

Scenes and scene vectors. The scenes we consider are taken from the *Topological Relations Picture Series* designed by Melissa Bowerman. This picture set is composed of 71 different line drawings of a wide range of spatial scenes. Each scene in the picture set represents a spatial relationship between a designated *figure* (indicated by an arrow in the drawing) and a *ground* object. Figure 1 shows a few examples of these drawings. Scene matrix S_1 includes all 71 pictures. We asked three English speakers to code these pictures using the

19 primitives in Table 2. Each primitive was described using a short phrase, and summaries of these descriptions are shown in Table 2. Matrix S_1 was created by merging the three sets of responses using a majority vote, and a subset of this matrix appears in Figure 2a. Scene matrix S_2 includes information for 27 scenes from the picture series. Feist coded each scene in terms of the primitives in her set, and matrix S_2 is based on her codes. A subset of S_2 is shown in Figure 2b.

Scene-term mappings. Matrix D_1 includes results for all 71 scenes. Levinson and Meira (2003) reported data for 4 languages, and we built on this data set by asking one speaker for each of 21 additional languages to label the set of 71 scenes. The languages included are listed in Table 2. Participants were asked to provide a single spatial term for each picture and were allowed to use as many different terms as they liked across the set of 71 scenes. In cases where they were not sure, we asked them to choose the term that seemed best to them. Feist (2000) asked speakers of 16 languages to label the scenes represented in S_2, and the results are collected in data matrix D_2.

Modeling the meaning of spatial terms

The information in a triple (P,S,D) can be used to explore the semantics of spatial terms. We consider a family of five models that make different assumptions about the spatial term representations T and the way in which scene representations (S) and term representations (T) combine to generate the term-scene mappings (D). All of the models assume that spatial term j is represented as a term vector t_j, but the models vary along three dimensions which determine the nature of the entries in each vector.

One of these dimensions—binary (B) or weighted (W)—indicates whether primitives can be differentially weighted. Binary models use term vectors t_j where 0 indicates that a primitive makes no contribution to the meaning of t_j, and 1 indicates that a primitive must be present in order for term j to apply. Weighted models use vectors where each entry is a real number between -1 and 1 inclusive. Weights near 1 indicate that a primitive should be present in order for a term to apply, and weights near -1 indicate that a primitive should be absent. A second dimension—singleton (S) or combination (C)—indicates whether terms correspond to single primitives or combinations of primitives. Singleton models assume that each term vector has exactly one non-zero entry, but combination models allow term vectors to have multiple non-zero entries. The final dimension—positive (+) or negative (-)—indicates whether spatial terms can be defined using negations of primitives. For binary models with negation, we expand the set of primitives so that it includes negated versions of each primitive in Table 2. For weighted models with negation, we keep the original set of primitives and capture negation by allowing term vectors to include negative weights. The three dimensions just introduced generate 8 models in total, and we will focus on the five models in Table 1.

Although some of our models allow term vectors t_j to contain real-valued entries, scene vectors s_i are always represented as binary vectors which specify which primitives apply (1) or do not apply (0) to each scene. Given a scene vector s_i and a term vector t_j, all of our binary models determine whether spatial term j applies to scene i as follows:

$$d_{ij} = \begin{cases} 1, & \text{if } {s_i}^T t_j = |t_j| \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where $|t_j|$ is the number of non-zero entries in term vector t_j. Equation 1 states that term i applies to scene j (i.e. $d_{ij} = 1$) only if all of the constraints specified by term vector t_j are consistent with the scene. Weighted models use a soft version of Equation 1:

$$d_{ij} = \begin{cases} 1, & \text{if } \sigma({s_i}^T t_j) > p \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where $\sigma(\cdot)$ is a sigmoid function (e.g. $\sigma(x) = \frac{1}{1+exp(-x)}$) which maps its argument into a probability (i.e. a number between 0 and 1). The parameter p is a threshold that will be learned from the data sets that we consider.

The models in Table 1 make contact with previous ideas from several fields. The singleton model is based on an idea proposed by Piaget and Inhelder (1956) who claims that there exists a common topology in which spatial languages build on concepts such as proximity and contiguity. Jackendoff (1983) further suggests that spatial semantics are composed of simple primitives such as “on” and “in”, which are directly encoded in languages. We expect, however, that the singleton model is unlikely to prove adequate. Levinson and others (Levinson & Meira, 2003; Levinson & David, 2006) have argued that there is great variation in spatial concepts across cultures, and the singleton model cannot account for this variation without an explosion in the number of primitives.

The combination models are also related to previous work. The discrete combination model captures the familiar proposal that meanings can be represented as conjunctions of primitive concepts, and psychologists have also proposed that spatial terms are represented as sets of weighted attributes (Feist, 2000). The weighted model in Equation 2 is known to statisticians as a logistic regression model, and is equivalent to a single-layer neural network, where the input (s_i) is mapped to the output (d_{ij}) via a layer of weights (t_j) and the sigmoid function.

Inferring term vectors

Our goals can now be precisely formulated. Given a triple (P,S,D) and one of the five models in Table 1, we wish to infer a term matrix T and decide how well S and T account for the data D. For both the singleton and conjunction models, we use a greedy algorithm to infer the term matrix T. For each spatial term we begin with a term vector t_j that includes only zeros, then greedily flip elements to improve a standard precision-recall F-score

$$F = \frac{2 \times \sum_i I(\hat{d_{ij}} = d_{ij} = 1)}{\sum_i I(\hat{d_{ij}} = 1) + \sum_i I(d_{ij} = 1)} \tag{3}$$

Table 2: Lists of author-collected languages (alphabetical), spatial primitives and their descriptions. "*" indicates negatable primitives. "F" and "G" stand for figure and ground.

Language	Primitive	Description
Arabic	above	F higher than G
Bengali	below	F lower than G
Cantonese	vertical equality*	F and G of equal height
Croatian	support*	F supported by G
English	horizontal support*	F supported horizontally by G
Finnish	front	F closer to viewer than G
French	back	G closer to viewer than F
German	viewpoint equality*	F and G equidistant from viewer
Hindi	contact*	F in touch with G
Indonesian	surface contact*	F in surface contact with G
Italian	attachment*	F attached to G
Japanese	adhesion*	F stuck to G
Mandarin	hanging*	F hung from G
Portuguese	piercing*	F pierces through G
Romanian	impaled*	F impaled by G
Russian	proximity*	F in close proximity to G
Slovakian	containment*	F contained by G
Slovene	encircled*	G circles F
Spanish	circlement*	F circles G
Thai		
Vietnamese		

where $\hat{d_{ij}}$ is a prediction based on the term vector t_j and d_{ij} indicates whether term j actually applies to scene i. The F-score will be high if most of the $\hat{d_{ij}} = 1$ entries predicted by t_j are correct (high precision), and if these predicted 1-entries include most of the actual 1-entries for term j (high recall).

For the weighted combination model, instead of inferring binary vectors we must learn a vector of weights for each term. Choosing the weights to maximize the F-Score is possible in principle (Jansche, 2005), but instead we fit a standard L1 regression model which is equivalent to a Bayesian logistic regression (Genkin, Lewis, & Madigan, 2004) with a Laplacian prior on the weights. For each spatial term, this approach searches for a weight vector t_j such that Equation 2 accurately predicts which scenes can be described by term j. The Laplacian prior captures the idea that term vectors t_j should be as simple as possible, and encourages small entries in t_j to end up as zero weights. In addition to this prior, we use the number of non-zero entries inferred by the conjunction model as an upper bound on the number of non-zero weights for the weighted model. Allowing many of the entries to be non-zero gives the weighted model more flexibility, but enforcing a sparsity constraint enables a direct comparison between the conjunction and weighted combination models. After learning the weights in all of term vectors t_j, we finish by choosing threshold p in Equation 2 to maximize the F-score (Equation 3).

Results

We applied the five models just described to the two triples (P,S,D) mentioned previously. In each case we computed the term matrix T that best accounts for the data. Term vectors for some languages are shown in Figure 2, and are discussed towards the end of this section.

The extent to which each model captures each data set can be captured using the F-score in Equation 3. Scores for the five models are shown in Table 1. To assess whether these scores are better than chance-level performance, we compared them with baseline scores achieved on random data sets. We used three randomization strategies. A *randomized D* set is created by randomizing all entries in D so that the sparsity is preserved (i.e. the number of "1" entries remains the same but all other structure is lost. A *shuffled D* set is created by randomly reordering the rows in D and leaving the scene vectors in S fixed. Finally, a *shuffled S* set is created by permuting the rows in S and leaving D fixed. Note that both shuffled sets leave the columns in D and S unchanged and therefore preserve many characteristics of these matrices, including the extent to which scenes (i.e. columns) tend to fall into clusters. For each triple, we created 20 random sets for each randomization strategy and computed the model scores. We then used t-tests to evaluate the hypothesis that performance on the real sets was significantly higher than performance on the random sets. In all cases we obtained highly significant results with truncated $p < 0.001$ after correction for multiple tests (first five rows of Table 2). These results suggest that all of our models were able to capture the structure in the observed data better than chance.

Although all models appear to capture some structure in the data, it is natural to ask which model performs best. The scores for the individual models do not address this question directly—for example, since the singleton model is a special case of the conjunction model, the conjunction model will always achieve a higher score regardless of whether it is actually the better approach. We therefore compared pairs of models by exploring whether whether the difference between their scores was significantly above chance level. For each pair, we compared the difference in prediction scores on the real data set against the differences achieved on the three random sets. The results appear in the final five rows of Table 2. Rows 6 and 7 suggest that the conjunction models perform better overall than the singleton models. Rows 8 and 9 suggest that allowing negated primitives leads to a significant improvement in performance. Finally, row 10 suggests that the weighted combination model does not perform better than the conjunction model with negations. Note, however, that we also evaluated an alternative weighted model where the sparsity of the weight vectors was not constrained by the conjunctive solution, and where all of the entries in each vector were allowed to be nonzero. This model performed significantly better than the conjunction with negation model on three of the six randomized tests across the two data sets, suggesting that weighted combinations may capture some aspects

Table 3: Significance of model performances and pairwise comparisons from t-tests. The model scores on the real data sets are compared to those on the random sets (1 – *randomized D*, 2 – *shuffled D*, 3 – *shuffled S*). D_1 is data from the authors and Levinson and Meira (2003). D_2 is collected by Feist (2000). For each pairwise comparison, the model on the left scores higher than the model on the right (e.g. BC+ outperforms BS+). ‘*’ indicates statistical significance at $p < 0.05$.

	D_1			D_2		
	1	2	3	1	2	3
BS+	*	*	*	*	*	*
BS−	*	*	*	*	*	*
BC+	*	*	*	*	*	*
BC−	*	*	*	*	*	*
WC−	*	*	*	*	*	*
BC+ *vs* BS+	*	–	–	*	*	*
BC− *vs* BS−	*	*	–	*	*	*
BS− *vs* BS+	*	–	–	*	*	*
BC− *vs* BC+	*	*	*	*	*	*
BC− *vs* WC−	–	–	–	–	–	–

of spatial semantics. Future work can explore this issue in more detail and determine which sparsity assumptions allow weighted models to provide the best account of spatial terms.

Our analyses so far suggest that the primitives in P_1 and P_2 are able to account for much of the structure in data sets D_1 and D_2. It is important, however, to consider whether our models combine the primitives in psychologically meaningful ways. Figure 2 shows the definitions learned by our models for three languages, and focuses on a subset of 10 scenes that were used in both data sets. Figure 2a shows term vectors and predictions for our data set. Note that the conjunction model captures important aspects of meaning that the singleton model misses. For example, Figure 2a.ii shows that “contact” is included in the meaning of *on* by the conjunction but not the singleton model. The plots also illustrate how negations allow the conjunction model to improve its predictions. Figure 2a.xi shows that the conjunction model makes several predictions about “qian mian” that do not match the true scene-term mapping in Figure 2a.x. “Qian mian” corresponds roughly to the phrase “in front of,” and including “no contact” in the definition of this term allows the negated conjunction model to successfully predict that it will not apply to scenes like “handle on cupboard” or “stamp on letter.”

For our second analysis the term vectors T_1 (Figures 2b.iii and b.iv) can be compared against a gold standard, which is the set of term vectors manually assigned by Feist (Figure 2b.ii). The vectors learned by our model are similar to those specified by Feist, and the predictions that follow from Feist’s representation (Figure 2b.vi) do not appear more accurate overall than the predictions generated by our automatically learned term vectors (Figure 2b.vii).

Conclusion

We presented computational models that explore whether spatial concepts can be constructed by combining a set of universal primitives. Our results suggest that a large proportion of the information in two cross-linguistic data sets can be captured by models that begin with the primitives typically discussed in the literature and combine them using simple operations such as conjunctions and weighted sums. Our general framework (Figure 1a) can be used to address many questions in spatial cognition and we mention just two directions for future work. First, we fit our models to the cross-linguistic data by learning definitions for each spatial term, and future work can use our approach to explore how humans learn spatial concepts. Second, all our analyses used primitives that were specified *a priori*, but it is conceptually straightforward to develop models that learn the primitives that best account for a given data set. Uncovering the nature of spatial primitives presents many challenges, but computational approaches can help address some of these challenges.

References

Antell, S. E. G., & Caron, A. J. (1985). Neonatal perception of spatial relationships. *Infant Behavior and Development*, *8*, 15-23.

Bowerman, M., & Choi, S. (2001). Shaping meanings for language: Universal and language-specific in the acquisition of spatial semantic categories. In M. Bowerman & S. C. Levinson (Eds.), *Language acquisition and conceptual development.* Cambridge: Cambridge University Press.

Brown, P. (1994). The INs and ONs of Tzeltal locative expressions: the semantics of stative descriptions of location. *Linguistics*, *32*, 743-90.

Feist, M. I. (2000). *On In and On: An investigation into the linguistic encoding of spatial scenes*. Doctoral dissertation, Northwestern University.

Feist, M. I. (2008). Space between languages. *Cognitive Science*.

Genkin, A., Lewis, D., & Madigan, D. (2004). *Large-scale Bayesian logistic regression for text categorization* (Tech. Rep.). Rutgers University.

Jackendoff, R. (1983). *Semantics and cognition*. Cambridge, MA: MIT Press.

Jansche, M. (2005). Maximum expected F-measure training of logistic regression models. In *Proceedings of EMNLP*.

Khetarpal, N., Majid, A., & Regier, T. (2009). Spatial terms reflect near-optimal spatial categories. In *Proceedings of the 31st annal conference of the cognitive science society.*

Levinson, S. C., & David, W. P. (2006). *Grammars of space*. Cambridge University Press.

Levinson, S. C., & Meira, S. (2003). ’Natural concepts’ in the spatial topological domain — adpositional meanings in crosslinguistic perspective: an exercise in semantic typology. *Language*, *79*, 485-516.

Piaget, J., & Inhelder, B. (1956). *The child’s conception of space*. London: Routledge and Kegan Paul.

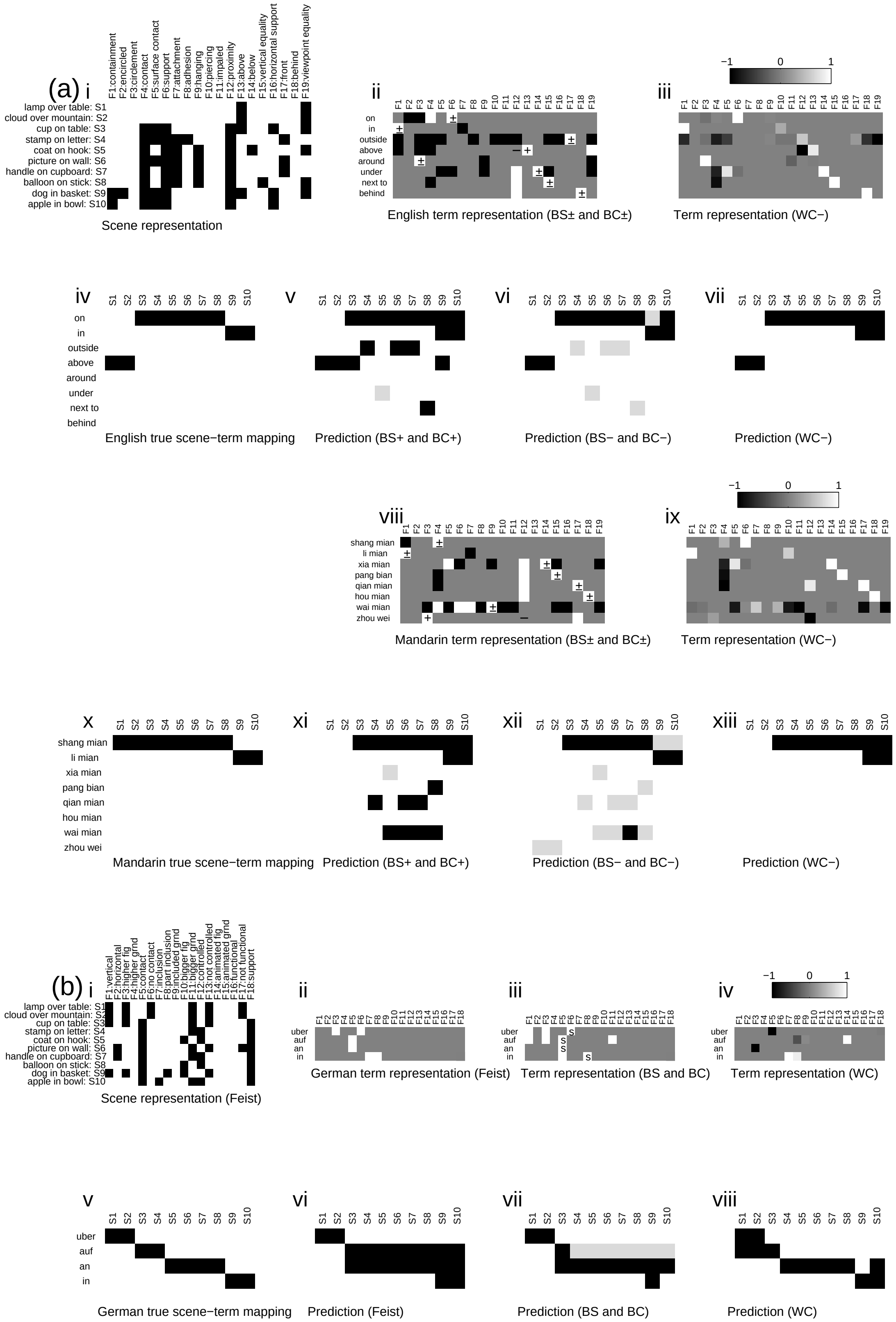

Figure 2: Term vectors and scene-term mappings for (a) English and Mandarin in the data set collected by the authors and (b) German in the data set collected by Feist. (a)(i) Ten scenes coded according to the nineteen primitives in Table 2. (ii) Inferred term vectors for four models: BS+ (indicated by +), BS- (-), BC+ (white cells) and BC- (white and black cells). Model BS- chooses a negated primitive only once (*above* is defined as “not F12”). (iii) Inferred term vectors for model WC-. (iv) True scene-term mappings (v) - (vii) Predicted scene mappings for five models. The predictions of models BC+ and BC- (black cells) are a subset of the predictions of the singleton models (black and gray cells). (viii)-(xiii) Results for Mandarin. (b) Results for German. Feist provided the encoding in (i) and the term vectors in (ii). (iii)-(iv) Term vectors for the singleton model (”S”), the conjunction model (white cells) and the weighted combination model. (v)-(viii) Actual and predicted scene-term mappings. Since the primitives in (b)(i) already include negations, models BS and BC do not allow additional negations.