Quantifying and extending the coverage of spatial categorization data sets

Wanchun Li[1], **Alexandra Carstensen**[2], **Yang Xu**[3], **Terry Regier**[4] **& Charles Kemp**[5]
[1]Faculty of Engineering and Information Technology, The University of Melbourne
[2]Psychology, Arizona State University
[3]Department of Computer Science, Cognitive Science Program, University of Toronto
[4]Department of Linguistics, University of California, Berkeley
[5]Melbourne School of Psychological Sciences, The University of Melbourne

Abstract

Variation in spatial categorization across languages is often studied by eliciting human labels for the relations depicted in a set of scenes known as the Topological Relations Picture Series (TRPS). We demonstrate that labels generated by multimodal large language models (MLLMs) align relatively well with human labels, and show how MLLM-generated labels can help to decide which scenes and languages to add to existing spatial data sets. To illustrate our approach we extend the TRPS by adding 42 new scenes, and show that this extension achieves better coverage of the space of possible scenes than two previous extensions of the TRPS. Our results provide a foundation for scaling towards spatial data sets with dozens of languages and hundreds of scenes.

Keywords: spatial semantics; large language models; cross-linguistic comparison; semantic typology; categorization

Introduction

Languages vary in how they carve the world into categories, and this variation has been studied most extensively in the domains of kinship (e.g. Hallam et al., 2025; Jones, 2010; Kemp & Regier, 2012) and color (e.g. Josserand et al., 2021; Kay et al., 1997; Zaslavsky et al., 2018). Progress in these domains has been enabled by the existence of standard representations, such as genealogical grids and perceptual color spaces, that serve as a universal substrate for cross-linguistic comparison. The availability of these representations has allowed the development of data sets such as Kinbank (Passmore et al., 2023) and the World Color Survey (Cook et al., 2005) that include category systems from more than a hundred languages.

Cross-linguistic variation in spatial categorization has also been widely studied, but this domain has been more resistant to formalization than either kinship or color. The greatest challenge is that there is no standard representation of the space of spatial relations, which makes it difficult to develop a spatial data set comparable to Kinbank or the World Color Survey. The most widely used set of spatial stimuli is a collection of 71 images known as the Topological Relations Picture series (TRPS; Figure 1a.i; Bowerman and Pederson, 1992). As Levinson and Wilkins (2006, p 9) note, the TRPS was "specifically designed to investigate the maximal range of scenes that may be assimilated to canonical IN- and ON-relations (and thus includes a number of scenes unlikely to be so assimilated)." The TRPS therefore makes no attempt to cover the space of possible spatial relations, which leaves room for extended stimulus sets that achieve greater coverage of this space.

Previous researchers have recognized the need to extend the TRPS, and Zhang (2013) and Landau et al. (2017) have developed new stimuli that are useful for investigating canonical IN- and ON-relations in even more detail (see Figures 1a.ii and 1a.iii). We pursue a different goal and prioritize the notion of coverage, which we formalize as the extent to which a stimulus set is representative of the full universe of possible scenes. Carstensen et al. (2015) suggest that greater coverage can be achieved by representing relations as lists of spatial features (e.g. CONTACT, SUPPORT, and PROXIMITY) and sampling scenes that provide good coverage of the space of possible feature lists. This feature-based approach is appealing because it provides an explicit characterization of the space of possible relations, but is relatively challenging to implement. Here we adopt a simpler approach which uses spatial terms as a guide to configurations that are missing from the TRPS. We first identify terms (e.g. English *outside* and Chinese 中间, 'among') that are not represented by the TRPS, and then add scenes that can be labeled with these terms (Figure 1a.iv).

Our ultimate goal is to develop data sets that include dozens of languages and hundreds of scenes, but scaling up in this way presents a significant challenge. Previous researchers have used foundation models to scale up psychological data sets including semantic feature norms (Suresh et al., 2025) and psycholinguistic databases (Martínez et al., 2024). The class of foundation models includes both large language models that are trained exclusively on text, and multimodal large language models (MLLMs) that handle multiple kinds of data including text and images. We argue that foundation models can also help to identify scenes and languages that are useful to add to existing spatial relations sets (Figure 1b) and illustrate our approach using MLLM-generated labels for 220 scenes in 23 different languages. To justify using MLLMs in this way, we first evaluate MLLM labels for the original TRPS against human labels collected by Carstensen et al. (2019) and Xu and Kemp (2010), and find that MLLMs achieve a relatively high level of accuracy. We do not propose that MLLMs can or should replace human experimental participants, but argue that MLLMs can contribute to an approach that ultimately includes human data collection.

In what follows, we first introduce our coverage-based approach for extending existing spatial data sets and provide a formal definition of coverage. We then present and evaluate our method for eliciting spatial category labels from MLLMs,

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

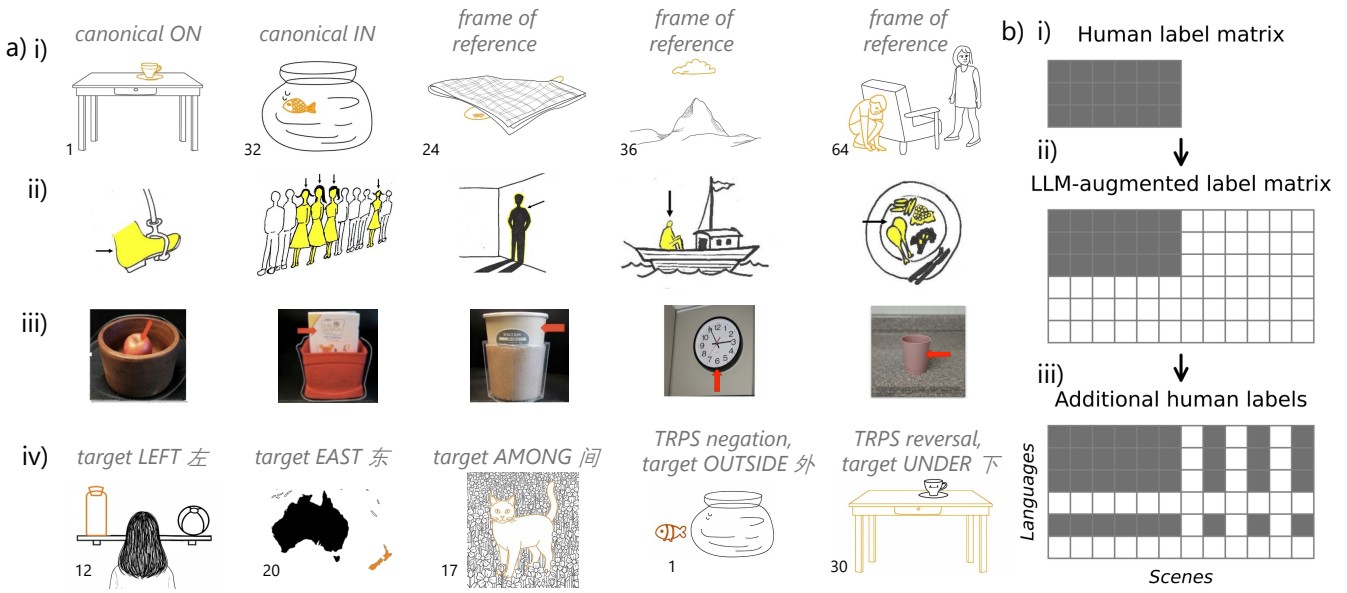

Figure 1: (a) Spatial relations stimulus sets. Each stimulus shows the relation between a focal object (shown in gold, or marked with an arrow) and a background object. Some of the original images have been cropped for this figure. (i) The Topological Relations Picture Series (TRPS; Bowerman and Pederson, 1992) includes 71 scenes that were chosen to explore the boundaries of terms for ON and IN relations. (ii) The Zhang set (Zhang, 2013) includes 63 scenes illustrating configurations that are absent from the TRPS but relevant to the expression of ON and IN in Chinese.[1] The examples here include "girls in line," "shadow on wall" and "food on plate." (iii) The LJSP set (Landau et al., 2017) includes 44 scenes that were designed to span six subtypes of containment and five subtypes of support. The examples here include "paper in box," "cup in sleeve," and "cup on counter." (iv) The LCXRK set developed by us includes 42 images designed to illustrate spatial terms in English and Chinese that are not represented by the TRPS, and to include negations and reversals of TRPS scenes. The examples here include "cat among flowers." (b) Extending a cross-linguistic data set using MLLMs. (i) The original human labels are organized into a dark gray matrix of languages (rows) by scenes (columns). (ii) We add candidate languages (new rows) and candidate scenes (new columns) and generate MLLM labels for all new combinations of languages and scenes. (iii) A coverage measure based on MLLM labels is used to prioritize languages and scenes for subsequent human labeling.

and show how MLLM data can help to choose which scenes and languages to add to existing data sets (Figure 1b). We finish by discussing limitations of our approach and prospects for scaling up spatial data sets to an even greater extent.

Quantifying coverage

Our general approach is summarized by Figure 1b. Suppose that we already have a matrix specifying human labels for a set of scenes across a set of languages (Figure 1b.i). We are considering extending the matrix to include additional scenes and languages, and use MLLMs to compile labels for all of these possibilities (Figure 1b.ii). We then use these MLLM labels to identify the scenes and languages that increase the coverage of the original data to the greatest extent, and ultimately collect additional human labels for these scenes and languages (Figure 1b.iii). Prioritizing scenes and languages based on coverage is consistent in some ways with theory-based sampling (Majid, 2023), or choosing a sample that varies along theoretically relevant dimensions. Prioritizing coverage, however, may also lead to samples that reveal theoretical dimensions that are exposed by the MLLM labels but not yet part of theories entertained by the experimenters.

Coverage can be formalized based on either a space of scenes or a space of languages; we begin with the former possibility. Given a universe U including all scenes in the MLLM-augmented matrix, the coverage of any subset S of U is defined as the average maximum similarity between each scene in U and its closest neighbour in S:

$$\text{Coverage}(S) = \frac{1}{|U|} \sum_{u \in U} \max_{s \in S} \text{sim}(s, u), \qquad (1)$$

where $\text{sim}(\cdot, \cdot)$ is a similarity measure over scenes. $\text{Coverage}(S)$ takes values between 0 and 1, and is high to the extent that each scene in U is similar at least one scene in S. The same equation can be used to compute the extent to which a universe U of languages is covered by a subset S of languages if we use a similarity measure $\text{sim}(\cdot, \cdot)$ over languages.

Later sections explain how similarity measures over scenes and languages can be derived from the MLLM-augmented matrix in Figure 1b.ii. First, however, we describe our method for eliciting spatial labels from MLLMs.

[1] Zhang (2013) mentions that she developed 64 new pictures but 63 were included in the set that she kindly shared with us.

Labeling spatial relations using MLLMs

Previous researchers have noted that AI can help scale up work on spatial categorization, and Meewis et al. (2023) compiled labels for the 71 TRPS scenes in 73 languages using a method that involves machine-translation of scene descriptions provided by English speakers. We developed a different approach in which MLLMs are treated much like human participants, and directly asked to label images.

Carstensen et al. (2019) collected TRPS labels for at least 13 speakers of each of seven languages, and Xu and Kemp (2010) collected labels from a single speaker for each of 21 languages. We aimed to collect MLLM responses for the 23 languages represented in one or both data sets. After some initial experimentation with different MLLMs (including GPT-4o and Grok 3) and prompts, we settled on Gemini 3 Flash (Google DeepMind, 2025) for our final runs. The Gemini series is generally viewed as handling multilingual tasks well, and as of January 2026, Gemini 3 Flash was the top performer on MMMLU (Multilingual Massive Multitask Language Understanding), a key multilingual benchmark. We accessed Gemini 3 Flash through its API and set the temperature parameter to zero with the goal of achieving deterministic and reproducible outputs.

We provided Gemini with a list of scenes, one per page, and each page included a page number along with annotations specifying the focal and background objects. For example, the page for the top-left scene in Figure 1a.i included 'focal object: cup' and 'background object: table.' Different stimuli sets use different visual cues for highlighting focal objects. We used a version of the TRPS that shows the focal object in gold (see Figure 1a.i), and the Zhang and LJSP stimuli use yellow and red arrows respectively to indicate focal objects.

All our MLLM results are based on a single run that included all 220 scenes from the four stimulus sets in Figure 1a. The prompt with Chinese as a target language is as follows:

> You are a native speaker of Chinese and I'd like you to respond in Chinese. Your task is to label the spatial relationships shown in a set of images. Here is a set of images that I'll call scenes.pdf. Each image shows a focal object and a background object. From image 1 to image 113, the focal object is gold and the background object is black. From image 114 to image 176, the focal object is yellow and indicated by an arrow, and the background object is black or blue. From image 177 to image 220, the focal object is indicated by a red arrow. An English speaker used the following spatial terms to describe the relationship between the focal object and the background object in each image: 1) "on"; 2) "in"; ... 220) "on". I'd like you to label the same images in scenes.pdf. For each image in scenes.pdf, please give me the spatial term in Chinese that best describes the relationship between the focal object and the background object. For each image, please respond using a single spatial term instead of a full sentence. And please do not translate the responses I gave you in English! Instead, I'd like you to respond as a native Chinese speaker would. Your responses (one for each image in scenes.pdf) should be organized into a numbered list.

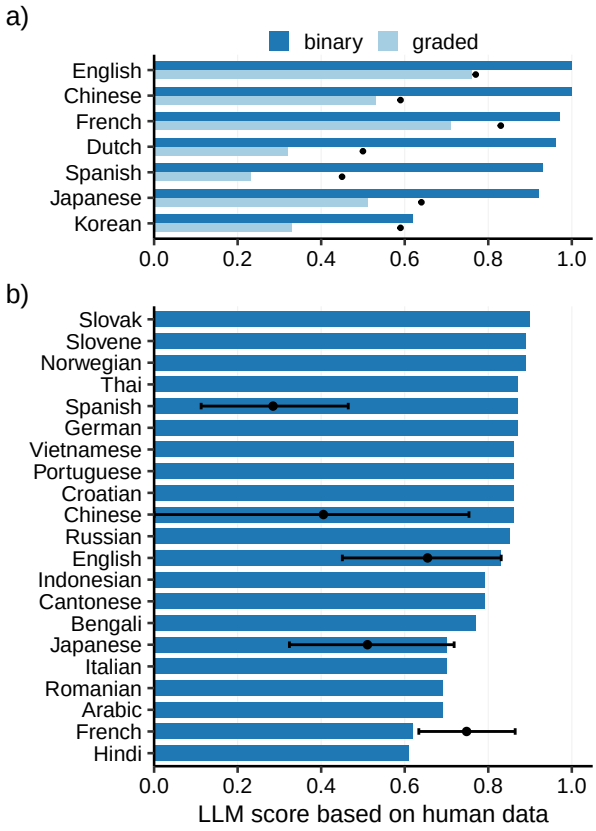

Figure 2: Evaluation of MLLM labels against data collected by (a) Carstensen et al. (2019) and (b) Xu and Kemp (2010). The binary score for an image label is 1 if a human provided the same label, and the graded score is the proportion of humans who provided the label. In (a), black points show the maximum possible graded score for each language. In (b), black points show the average result when a single human from the Carstensen et al. (2019) data is scored with respect to a different speaker of the same language, and error bars show the 95% interquantile range.

All prompts included 220 labels provided in a reference language. Chinese was the reference when English was the target, and English was the reference for all other languages. Choosing English as the reference in most cases induces a bias, but seemed appropriate given that MLLMs are likely to be more proficient in English than any other language. For the TRPS scenes, English reference labels were the modal English responses collected by Carstensen et al. (2019), and for the Zhang and LJSP sets, each English reference label was either *in* or *on*. For the LCXRK set (Figure 1a.iv), the English reference labels were the target spatial terms that the LCXRK scenes were designed to illustrate. For all four data sets the Chinese reference labels were the Chinese responses generated by Gemini 3.

Figure 2a compares the MLLM labels with data collected

Table 1: Coverage scores achieved by four stimulus sets relative to a universe U of 220 scenes.

Stimulus set	Size	Coverage score	95% CI
TRPS	71	0.914	[0.82, 0.95]
TRPS + Zhang	134	0.918	[0.85, 0.95]
TRPS + LJSP	115	0.918	[0.87,0.95]
TRPS + LCXRK	113	**0.964**	[0.96, 0.995]

by Carstensen et al. (2019). The binary score for a label is 1 if at least one human provided the same label, and 0 otherwise. Binary scores for six of the seven languages exceed 0.9, but the score for Korean is lower. For 15 of the 71 TRPS scenes, the MLLM chooses the locative marker -에, which does not appear in isolation in the Carstensen et al. data. Instead, Korean speakers often used verbs such as 끼워져 ('fitted on') for a ring on a finger, or 걸려 ('hanging') for a phone on the wall.

High binary scores indicate that the MLLM's responses tend to match the labels of some humans, but do not reveal whether those responses are broadly typical of human labels. We therefore computed a graded score for each MLLM label defined as the proportion of humans who gave the same response. Maximum graded scores for each language are shown as black points in Figure 2a, and the MLLM averages for English, Chinese, French and Japanese are all within 0.15 of the maximum possible graded score. For Dutch and Spanish the gap between the MLLM graded score and the maximum is greater. The MLLM chooses *en* ('in, on, or at') as the Spanish label for 41 out of 71 scenes, but Spanish speakers often give more specific responses such as *sobre* ('on') for cup on table, or *adentro* ('inside') for fish in bowl.

Figure 2b compares MLLM labels with data collected by Xu and Kemp (2010). Only one speaker was consulted for each language, which means that the binary score for a label now indicates whether or not the MLLM matched the human label for the same image. For comparison, Figure 2b includes black points showing the average alignment score between two speakers of the same language in the Carstensen et al. (2019) data. The average MLLM score exceeds 0.8 for 12 of the 21 languages, and French and Hindi achieve the lowest scores. The MLLM scores relatively well for French in Figure 2a, however, suggesting that the single French speaker in the Xu and Kemp (2010) data set gave somewhat idiosyncratic responses, such as *en haut a droite de* ('on the top right of') instead of *sur* ('on') for a stamp on a letter.

For us, knowing whether the MLLM produces relatively reliable labels is important, but understanding how it produces these labels is less essential. We wondered, however, whether the MLLM actually needs images of the TRPS scenes to generate reliable labels. We therefore ran a text-based condition that was almost identical to our original image-based condition except that all images were stripped from the file `scenes.pdf` provided to the MLLM, leaving only specifications of the focal and background objects in each scene. The results were similar to those reported in Figure 2, and average scores across languages were virtually identical: mean binary scores across the 7 languages in Figure 2a were 0.91 (image-based) and 0.90 (text-based), mean graded scores were 0.48 (image-based) and 0.49 (text-based), and mean binary scores across the 21 languages in Figure 2b were 0.80 (image-based) and 0.78 (text-based). These results therefore suggest that image analysis makes little contribution if any to the scores reported in Figure 2, and that text-only models may be able to achieve similar results.

As of January 2026, Google Translate supports 249 languages, including all of the languages in Figure 2, which suggests that 249 is an upper bound on the number of languages to which our MLLM-based approach can be applied. Languages outside of this set account for much of the world's linguistic diversity (Joshi et al., 2020; Skirgård et al., 2023), but are poorly represented in the training data available to MLLMs. For high-resource languages, however, our results so far suggest that MLLM-generated labels of spatial configurations are accurate enough to support research on spatial categorization

MLLMs may be especially valuable when running quick preliminary tests of hypotheses that require many languages to evaluate, or when choosing languages or stimuli to prioritize for future experiments with humans (Figure 1b). The following sections illustrate the latter possibility.

Adding scenes to existing data sets

Levinson and Wilkins (2006) present a taxonomy that organizes spatial relations into three groups: topological relations (e.g. English *in* and *on*), frame-of-reference relations (e.g. *behind*, *left of*, *north of*) and relations involving motion (e.g. *from London to Sydney*). Because we work with static pictures, we do not consider motion-related relations, but we sought to add additional topological relations (e.g. *outside*, *off*) and frame-of-reference relations (*left of*) to the TRPS. Despite its name, the TRPS is not restricted to topological relations. It includes frame-of-reference relations such as "spoon under cloth" and "boy behind chair," which help to characterize how languages treat the boundary between topological and frame-of-reference relations. Adding additional frame-of-reference relations to the TRPS therefore builds on a theme that is already present in the original data set.

We used two distinct strategies to add 42 scenes to the TRPS. 27 scenes were added by identifying spatial terms in some language that are not already represented by the TRPS, and designing new scenes that illustrate these relations. The authors include native speakers of Chinese and English, and we therefore focused on these two languages. We consulted lists of Chinese and English spatial terms provided by Tyler and Evans (2003), 全国科学技术名词审定委员会 (2011) and Landau and Jackendoff (1993), and identified 11 Chinese terms (外, 间, 中, 东, 西, 南, 北, 左, 右, 前, 后) and 15 English terms (*across, after, along, among, at, between, down, east, left, north, off, outside, right, south, west*) that seemed like good candidates for illustration. We chose focal and back-

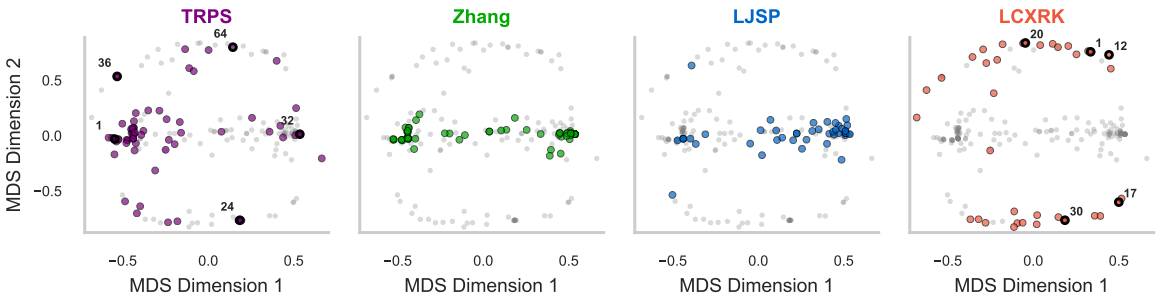

Figure 3: MDS visualization of a space that includes scenes from all stimulus sets considered in this paper. The four panels show the coverage of this space achieved by (a) the TRPS (b) the Zhang set (c) the LJSP set and (d) our new stimulus set. Points labeled in panels (a) and (d) correspond to images shown in Figures 1a.i and 1a.iv.

ground objects for all of these cases, then prompted Gemini to generate illustrations like the examples in Figure 1a.iv.

The remaining 15 scenes were created by using two methods to adjust existing TRPS stimuli. The first method negates the original TRPS relation (e.g. so that a fish is outside a fishbowl instead of in the fishbowl). The second method keeps the original TRPS scene but exchanges the focal and background objects (e.g. so that the table is under the cup instead of the cup being on the table). We did not exhaustively apply these transformations to all scenes, but included 15 cases that seemed sensible to us.

Thousands if not millions of scenes could be added to the TRPS, and a coverage measure is useful for identifying which scenes should be prioritized when collecting human labels. We used the coverage measure in Equation 1 to compare the coverage of three extensions of the TRPS. The first is the LCXRK set, which consists of our 42 new scenes, and the second and third are the Zhang and LJSP sets shown in Figures 1a.ii and 1a.iii.[2] Both of these projects focused on ON and IN relations, and achieving broad coverage of topological and frame-of-reference relations was a goal of neither project. Comparing three extensions, however, will allow us to illustrate how our coverage measure can be applied.

Our starting point is to compile MLLM labels for all 220 scenes belonging to U, the union of the TRPS, Zhang, LJSP and LCXRK data sets. We collected labels for all 23 languages shown in Figure 2. For each language L, the similarity $\text{sim}_L(i, j)$ between scenes i and j is defined as 1 if the labels for these scenes match and 0 otherwise. The overall similarity $\text{sim}(i, j)$ between scenes i and j is then defined as the average of $\text{sim}_L(i, j)$ across the 23 languages.

We combined this similarity measure with Equation 1 to compute coverage scores for the TRPS along with three extensions of the TRPS. Table 1 shows that the Zhang and LJSP extensions improve the coverage of the TRPS by a relatively small margin, but the LCXRK set achieves a more substantial improvement. All scores in Table 1 are relatively high because the 220 scenes in U are mainly examples of ON and IN, and all four sets include examples of these relations. We computed confidence intervals on the coverage scores by creating 1000 bootstrap samples of the universe U, and recording the middle 95% of scores across these samples. The confidence interval for TRPS + LCXRK does not overlap any of the others, suggesting that our design goal was achieved, and that this data set provides better coverage of the space of possible relations than the other two extensions of the TRPS.

To visualize the space of semantic relations, we applied multidimensional scaling to the 220×220 dissimilarity matrix D, where $D_{ij} = 1 - \text{sim}(i, j)$. The resulting two-dimensional representation of the 220 scenes is shown in Figure 3. The four different panels show how different sets of stimuli are distributed across the space, and are consistent with our quantitative finding that the LCXRK set provides greater coverage than either the Zhang set or the LJSP set. The first MDS dimension can be interpreted as a cline from ON to IN, but the second dimension does not appear to be interpretable, and a stress plot suggests that at least three dimensions are needed to account well for the dissimilarity matrix D. It therefore makes more sense to compute coverage scores using the original dissimilarity D (as we did for Table 1) than using distances based on the MDS solution in Figure 3.

The core idea behind our coverage measure can also be used to quantify the extent to which individual scenes go beyond the TRPS. We ranked the 42 images in the LCXRK set based on the similarity of each image to its closest neighbour in the TRPS. The results suggest that scenes illustrating "under" (including table under cup and head under hat) are the scenes most similar to the TRPS, and in particular to the TRPS scene with "spoon under cloth." Nine of the new scenes have zero similarity to every TRPS scene, and therefore depart maximally from the TRPS. These scenes include scenes illustrating the topological relation "outside" (including fish outside fishbowl, and bird outside birdcage) and scenes showing cardinal relations (including "New Zealand east of Australia"). These rankings suggest which individual scenes are most useful to include in future behavioral experiments.

The LCXRK set, however, is relatively small, which means that it can be easily labeled in its entirety. We collected labels from 10 native Chinese speakers and 7 native English speakers, and have released them at the OSF along with images of

[2]Unfortunately the work of Carstensen et al. (2015) could not be included as a fourth set because those authors no longer have a copy of their stimuli.

the stimuli. The scenes were often but not always successful at eliciting the target relations, and the modal response aligned with the target for 37 (Chinese) and 33 (English) out of 42 scenes. Importantly, the modal human response did not appear among the TRPS labels collected by Carstensen et al. (2019) for 13 (Chinese) and 15 (English) out of 42 scenes. This finding confirms that the LCXRK set does succeed in extending the coverage of the TRPS.

Adding languages to existing data sets

Suppose we would like to decide which of the languages in Figure 2b increases the coverage of the seven-language Carstensen et al. (2019) data by the greatest extent. The similarity measure required by Equation 1 can be defined as $1 - D$, where D is a 23 by 23 matrix of distances between all of the languages in Figure 2. We define the distance between any pair of languages using variation of information (Meilă, 2007), an information-theoretic metric that quantifies the difference between the partitions induced by MLLM labels for these two languages. Figure 4 shows a two-dimensional MDS visualization of distances between these languages. The distance matrix suggests, for example, that Cantonese may be a relatively low priority because its system is relatively similar to the Chinese system already documented in the Carstensen et al. data. If we order languages based on their distance from their nearest neighbour in the Carstensen et al. set, Portuguese and Romanian emerge as the languages most distant from any of the seven Carstensen et al. languages, and the two that increase coverage by the greatest extent.

To test this prediction we computed distance matrices D separately based on MLLM labels and the human labels for the 21 languages in the Xu and Kemp (2010) data set. Five of the Carstensen et al. languages are included in these matrices, and for each of the remaining 16 languages we extracted its distance to the nearest Carstensen et al. language. Distances based on MLLM and human data had a Pearson correlation of 0.49 (95% CI [0.19, 0.83]), where the confidence interval is based on 1000 bootstrapped samples of the 21 language set. Portuguese and Romanian (the two targets identified using the MLLM data) appear among the four languages most distant from the Carstensen et al. languages according to the Xu and Kemp data. These results provide tentative support for the idea that MLLM data can help identify languages that increase the coverage of existing data sets.

Discussion

We showed that MLLMs align relatively well with humans when asked to label spatial relations in high-resource languages. This alignment suggests that MLLMs can contribute to research on spatial categorization, and we illustrated how MLLM labels can be used to choose scenes and languages that extend the coverage of existing spatial categorization data sets. Our approach can potentially be used to scale up to a data set including the roughly 80 languages represented on `prolific.com` and hundreds of scenes. For this purpose, the

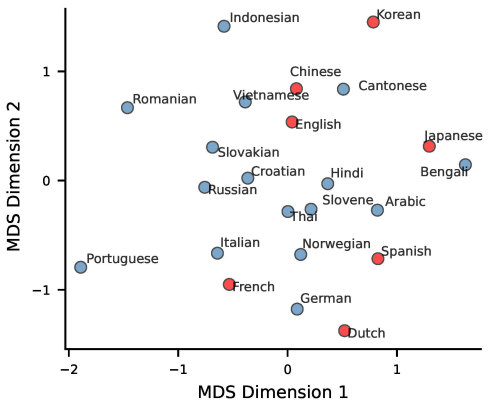

Figure 4: MDS visualization of the similarity between spatial systems from the languages represented in Figure 2. Languages appearing in the Carstensen et al. (2019) data set are shown in red.

set of target languages is already clear, but choosing which scenes to include remains a difficult challenge. One possibility is to generate thousands of scenes using the feature-based approach of Carstensen et al. (2015), then apply our coverage score to MLLM labels for these scenes to identify a smaller subset to include in the actual experiment.

Instead of working with features, we filled gaps in the TRPS by identifying spatial terms in English and Chinese that lacked representation. This approach can easily be extended to include missing terms from other languages, but does not guarantee systematic coverage of the space of possible relations. In contrast, a feature-based approach can enumerate all logically possible configurations of features and then select a subset that provides both broad and even coverage of the entire space. Combining our MLLM-based approach with a feature-based approach is therefore a high priority for future work.

Although we focused on applying our coverage measure to MLLM labels, this measure can usefully be applied to human labels. One possible application might consider the data set of Levinson et al. (2003), which includes TRPS labels in nine typologically diverse languages, four of which are not supported by Google Translate. Our coverage measure could be used to compare the coverage of this set relative to a larger set that includes only high-resource languages. This comparison would provide some information about how much diversity is missed by MLLM-based approaches that work only with high-resource languages.

Our approach is inspired by previous experimental work using the TRPS, but another line of work uses textual corpora to study cross-linguistic variation in spatial categorization (Beekhuizen, 2025; Viechnicki et al., 2025). These two approaches are complementary and MLLMs can contribute to both: for example, MLLMs can be used to extract spatial relation markers from multilingual corpora (Inostroza et al., 2026). As MLLMs continue to improve, we expect that they will prove increasingly useful for both experimental and corpus-based work on spatial categorization across languages.

Acknowledgments

We thank Zhang (2013) for sharing her stimuli. Generative AI was used as described in the paper to label images and to generate images illustrating spatial relations. The LCXRK data set is available at https://doi.org/10.5281/zenodo.18765858

References

Beekhuizen, B. (2025). Spatial relation marking across languages: Extraction, evaluation, analysis. *Proceedings of the 29th Conference on Computational Natural Language Learning*, 571–585.

Bowerman, M., & Pederson, E. (1992). Topological relations picture series. In S. C. Levinson. (Ed.), *Space stimuli kit 1.2* (p. 51). Max Planck Institute for Psycholinguistics.

Carstensen, A., Kachergis, G., Hermalin, N., & Regier, T. (2019). "Natural concepts" revisited in the spatial-topological domain: Universal tendencies in focal spatial relations. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *41*, 197–203. https://escholarship.org/uc/item/7tq511sd

Carstensen, A., Xu, Y., Kemp, C., & Regier, T. (2015). The space of spatial relations: An extended stimulus set. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *37*, 2864. https://escholarship.org/uc/item/8501t1k5

Cook, R. S., Kay, P., & Regier, T. (2005). The world color survey database. In *Handbook of categorization in cognitive science* (pp. 223–241). Elsevier.

Google DeepMind. (2025). Gemini 3 Flash [Accessed January 2026]. https://ai.google.dev

Hallam, M., Jordan, F. M., Kirby, S., & Smith, K. (2025). Predictive structure emerges during the generalisation of kin terms to new referents. *Open Mind*, *9*, 1431–1466.

Inostroza, D., Vylomova, E., Kemp, C., Carroll, M., Li, W., & Mistica, M. (2026). Where the cat sat: A multilingual benchmark for spatial language understanding. In M. Liakata, V. P. Moreira, J. Zhang, & D. Jurgens (Eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*.

Jones, D. (2010). Human kinship, from conceptual structure to grammar. *Behavioral and Brain Sciences*, *33*, 367–416.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. *arXiv preprint arXiv:2004.09095*.

Josserand, M., Meeussen, E., Majid, A., & Dediu, D. (2021). Environment and culture shape both the colour lexicon and the genetics of colour perception. *Scientific reports*, *11*(1).

Kay, P., Berlin, B., Maffi, L., & Merrifield, W. (1997). Color naming across languages. In *Color categories in thought and language* (pp. 21–56).

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*(6084), 1049–1054.

Landau, B., & Jackendoff, R. (1993). "What" and "where" in spatial language and spatial cognition. *Behavioural and Brain Sciences*, *16*, 217–265.

Landau, B., Johannes, K., Skordos, D., & Papafragou, A. (2017). Containment and support: Core and complexity in spatial language learning. *Cognitive science*, *41*, 748–779.

Levinson, S. C., Meira, S., & The Language and Cognition Group. (2003). 'Natural concepts' in the spatial topological domain—adpositional meanings in crosslinguistic perspective: An exercise in semantic typology. *Language*, (3), 485–516.

Levinson, S. C., & Wilkins, D. P. (2006). The background to the study of the language of space. In S. C. Levinson & D. P. Wilkins (Eds.), *Grammars of space: Explorations in cognitive diversity* (pp. 1–23). Cambridge University Press.

Majid, A. (2023). Establishing psychological universals. *Nature Reviews Psychology*, *2*(4), 199–200.

Martínez, G., Molero, J. D., González, S., Conde, J., Brysbaert, M., & Reviriego, P. (2024). Using large language models to estimate features of multi-word expressions: Concreteness, valence, arousal. *Behavior Research Methods*, *57*(1), 5.

Meewis, F., Fourtassi, A., & Dautriche, I. (2023). Typology of topological relations using machine translation. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*.

Meilă, M. (2007). Comparing clusterings—an information based distance. *Journal of multivariate analysis*, *98*(5), 873–895.

Passmore, S., Barth, W., Greenhill, S. J., Quinn, K., Sheard, C., Argyriou, P., Birchall, J., Bowern, C., Calladine, J., Deb, A., et al. (2023). Kinbank: A global database of kinship terminology. *PLOS One*, *18*(5), e0283218.

Skirgård, H., Haynie, H. J., Blasi, D. E., Hammarström, H., Collins, J., Latarche, J. J., Lesage, J., Weber, T., Witzlack-Makarevich, A., Passmore, S., et al. (2023). Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, *9*(16), eadg6175.

Suresh, S., Mukherjee, K., Giallanza, T., Yu, X., Patil, M., Cohen, J. D., & Rogers, T. T. (2025). AI-enhanced semantic feature norms for 786 concepts. *Topics in Cognitive Science*.

Tyler, A., & Evans, V. (2003). *The semantics of English prepositions: Spatial scenes, embodied meaning, and cognition*. Cambridge University Press.

Viechnicki, P., Eberstadt, Z., & Landau, B. (2025). Spatial terms in English plus twelve languages: Evidence for functional and geometric classes. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *47*.

Xu, Y., & Kemp, C. (2010). Constructing spatial concepts from universal primitives. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *32*, 346–351. https://escholarship.org/uc/item/1kd825gh

Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, *115*(31), 7937–7942.

Zhang, Y. (2013). *Spatial representation of topological concepts IN and ON: A comparative study of English and Mandarin Chinese* [Doctoral dissertation, Concordia University].

全国科学技术名词审定委员会. (2011). 语言学名词 [In Chinese]. 商务印书馆.