Spatial Term Variety Reflected in Eye Movements on Visual Scenes

Cengiz Acartürk (cengiz.acarturk@uj.edu.pl)
Department of Cognitive Science, Jagiellonian University, Krakow, Poland

Şeyma Nur Ertekin (s.n.ertekin@uva.nl)
Department of Psychology, University of Amsterdam, Amsterdam, The Netherlands

Abstract

Verbal descriptions of spatial configurations open a window to a specific aspect of visual cognition relevant to the interpretation of topological relations in the visual world. The present study reports an experimental investigation of the production of spatial prepositions by human participants while they verbally described visual stimuli within a stimuli battery commonly utilized in relevant research. The analysis of participants' eye movements revealed a relationship between the variety of spatial terms in the given language and native speakers' fixation patterns on the stimuli. A broader spectrum of spatial expressions, describing the same visual scene, is related to longer and more frequent fixations on the stimuli. The findings reflect cognitive processes, as indicated by oculomotor control variables, related to the verbal expression of spatial relationships.

Keywords: spatial cognition; spatial language; eye movements

Introduction

Cognitive processes related to objects in space and their interrelationships play a pivotal role in natural cognitive systems. Linguistic articulations of these spatial relationships offer valuable insights for examining the mapping between spatial language and objects in visual world. Spatial prepositions, a specific category within spatial terms, have attracted research interest due to their limited number, compared to the large repository of content words in natural languages (Landau & Jackendoff, 1993). Furthermore, the limited set of spatial prepositions is tasked with covering a wide range of possibilities, contrasting sharply with the richness of spatial relations that exist between objects. This leads to a phenomenon known as polysemy. Polysemy, a complex concept in lexical semantics, refers to a single word having a range of distinct yet related meanings. Given their limited number in natural languages, spatial prepositions are inherently polysemous (Coventry & Garrod, 2004).

The study of the mapping between language and the perceptual system often focuses on the situation-specific meanings of spatial terms (e.g., Coventry & Garrod, 2004). This approach is crucial as defining words solely in terms of other words can be inadequate for mapping meanings to entities in the visual world, as Glenberg & Robertson (2000) note. Previous research analyzed verbal descriptions of spatial relations from various aspects and across languages, aiming to uncover universal semantics in spatial language. This involved empirical studies that employed established assessment tools, such as the Topological Relations Picture Series (TRPS in Bowerman & Pederson, 1992; Bowerman, 1996; Carstensen et al. 2019; Levinson et al., 2003; Majid, Jordan & Dunn, 2015, among others). For example, Levinson et al. (2003) examined closed-class adpositions to determine their ability to reflect universal spatial semantics. They noted the lack of consensus on spatial terms, such as 'in' and 'on' in English, alongside counterparts in other natural languages. In contrast, Carstensen et al. (2019) explored spatial topological concepts akin to 'in' and 'on', and analogous terms across diverse languages, including Dutch, English, French, Japanese, Korean, Mandarin, and Spanish, utilizing the TPRS battery (Bowerman & Pederson, 1992). Their findings indicated that speakers of these languages exhibit similar patterns in articulating core spatial relationships, providing empirical, cross-linguistic substantiation of universal tendencies in spatial relational expressions. Consequently, previous research on the use of spatial terms presents varied outcomes in the context of universal versus language-specific characterizations. This indicates a need for further analysis for the generalizability of the findings.

The investigation of universal tendencies in spatial terms extends to the study of preferred *Frames of Reference* (*FoR*). A significant contribution to this area was made by Majid et al. (2004), showing that speakers tend to utilize the same *FoR* in non-linguistic tasks as they do in linguistic tasks that employ an absolute *FoR*. This finding suggests that language may influence the frame of reference used in non-linguistic cognitive processes, hinting at a potential interdependence between linguistic and spatial strategies. On the other hand, Le Guen (2011) documented a contrasting scenario among the Yucatec Maya speakers, who predominantly employed a geocentric frame of reference in nonlinguistic tasks, despite not utilizing it in their verbal expressions. This challenges the idea of a straightforward correlation between linguistic preferences and the nonlinguistic cognitive processing of spatial relationships, introducing a nuanced perspective on the interaction between language and spatial cognition.

Yun and Choi (2018) propose that the relationship between language and spatial cognition is more complex than previously thought, marked by both cross-linguistic similarities and differences. In their study, native Korean and English speakers were asked to verbally describe dynamic spatial scenes. The researchers observed notable semantic differences in how speakers of these two languages conceptualize spatial relations, especially concerning containment, tight-fit, and verticality. This linguistic analysis was complemented by non-linguistic data from Choi and Hattrup (2012), in which participants identified spatial relations in similar dynamic scenes. By applying language-specific logistic regression models, Yun and Choi found that linguistic interpretations significantly predicted the non-

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

linguistic data, suggesting a profound influence of spatial semantics on spatial cognition. Yun and Choi argue that the similarities and differences observed between Korean and English speakers may be rooted in a universal perceptual foundation for spatial cognition. Nonetheless, they suggest that each language builds its own semantic framework on this universal base, resulting in distinct linguistic identities. This perspective highlights the intricate interplay between universal perceptual experiences and the unique semantic structures inherent in individual languages.

Johannes et al. (2015) addressed the concept of universal spatial language semantics by exploring the diversity of spatial expressions across different languages. Their research aimed to identify parallels within specific spatial relation categories across languages of varied typologies. A central premise of their study was that analyzing the variation of spatial expressions within a single language might uncover patterns similar to cross-linguistic commonalities. One of their key findings was the discovery of fundamental similarities in expressions, especially within the subcategories of Containment and Support. However, the study also revealed a range of language-specific expressions, highlighting the complex and diverse nature of spatial language semantics.

In this study, we investigate the spatial terms in the Turkish language, with a specific focus on the variability of expressions provided in verbal descriptions of the TPRS stimuli (Topological Relations Picture Series, Bowerman & Pederson, 1992). A key aspect of our analysis involves examining the eye movements of participants as they relate to the diverse range of spatial expressions. To this end, we assess the spatial terms used in response to the stimuli, particularly looking at their variance — that is, the breadth of different spatial prepositions employed. Additionally, we explore the potential relationship between the variety of spatial expressions and oculomotor variables, such as fixation duration. This investigation aims to contribute to our understanding of Turkish spatial language comprehension, and how the diversity of spatial language correlates with the visual processing indicators measured through eye movement analyses.

Gaze Analyses for Studying Spatial Language

Eye movements serve as a dynamic measure of visual processes and offer insight into the cognitive processes (Liversedge & Findlay, 2000). Although more of an operational assumption than an incontrovertible fact, oculomotor variables are often interpreted as indicators of attention allocation. Therefore, ongoing perceptual, cognitive, and behavioral activities are assumed to be closely related to spatial and temporal characteristics of fixations on regions in a visual scene (Duchowski, 2017). The previous research on language production reveals a systematic relationship between eye movements and verbal descriptions of a visual scene (Griffin & Bock, 2000). For example, Meyer, Sleiderink & Levelt (1998) examined speech planning processes in object naming by analyzing eye movements on the objects. They claim that conceptual and most linguistic processing is completed before a gaze shift. Accordingly, time and patterns of eye movements reflect the interaction between visual and linguistic processes, and the underlying mechanisms of language processing and visual attention are tightly structured (Mayberry, Crocker & Knoeferle, 2009). In spatial language research, researchers utilized eye movements to analyze the spatial language comprehension in children. For instance, Lakusta, Hussein, Wodzinski, and Landau (2021) employed eye tracking and reported that 20-month-old infants were able to discriminate between different dynamic support configurations by analyzing the differences between gaze durations. These findings suggest that oculomotor control variables may provide data for studying spatial language semantics.

Recently, there has been limited research focusing on spatial relations in visual scenes from the perspective of adult language comprehension. To date, only a few studies, such as Johannes et al. (2015), have explored the variety of spatial expression encodings used to describe specific spatial relation scenes. A common methodology in this area of research involves selecting the most frequently used spatial expression and associating it with a spatial scene as represented by experimental stimuli. Cross-linguistic assessments are conducted to explore variations in categories of spatial relation types, such as Containment and Occlusion.

In the present study, we explore the relationship between spatial term variety (i.e., the use of alternative expressions to describe a spatial relation) and the time course of describing spatial scenes, as revealed by the characteristics of oculomotor control variables. To this end, we aim to investigate how spatial configurations, represented by a static scene, influence the duration of the eye fixations on specific, linguistically related areas of interest, such as *Figure* and *Ground*, in visual scenes. Figure and Ground are the terms used in the previous studies (e.g., Langacker, 1986) to refer to *located object* and *reference object*, respectively. As noted by Coventry and Garrod (2004), other alternatives include *primary object* and *secondary object* (Talmy, 1983), *trajectory* and *landmark* (Lakoff, 1986), and *theme* and *reference object* (Jackendoff, 1985) for English (p. 10).

We recorded participants' eye movements while they described spatial positions of a set of located objects (Figures) from the TRPS stimuli, consisting of 71 line drawings. Mean fixation duration, total gaze duration, and fixation counts were analyzed as the dependent variables. For the analysis of spatial terms, we investigated verbal descriptions and the variety of spatial terms. Accordingly, the present study has the following basic research question: What is the relationship between oculomotor control variables (e.g., fixation duration, fixation count), behavioral variables (response time), and linguistic stimuli (*Figure* and *Ground*) with the variety of spatial terms? The following section describes the methodology of the present study.

Methodology

Participants

The experiment was conducted in Turkish, the native language of the participants. Thirty-four adult native speakers of Turkish and two adult speakers of Azerbaijani, a Turkic language, participated in the experiment. The mean age of the participants was 21.81 (*SD* = 1.62, 15 females). Participants were undergraduate or graduate university students. Ethics approval and participant content were received before conducting experiment sessions.

Experiment Design, Materials, and Procedure

In this within-subject experiment, participants were exposed to the entire set of experimental stimuli, while their eye movements were recorded using a Tobii T120 desktop eye tracker with 120 Hz sampling rate. For data analysis, eye movements and response times were processed using the manufacturer's software and the open-source statistical software JASP 0.16.4. Participants' verbal responses were captured through a desktop microphone and transcribed manually.

The stimuli were drawn from the Topological Relations Picture Series developed by Bowerman and Pederson (1992). This set consist of 71 line drawings that depict a variety of static spatial configurations representing various topological relationships. These scenes are designed to elicit the use of spatial terms like 'in', 'on', 'under', 'over', 'near', and 'against' in English, and their equivalents in other languages. As of our knowledge, these stimuli were not tested so far, in Turkish.

Each experimental stimulus consisted of three components: Figure, Ground, and Text. The Figure, always colored orange, was positioned against the Ground, depicted in black, on a white canvas background. Accompanying each image was a text prompt asking participants: "Where is the object?", referring to the Figure. Examples of these stimuli are illustrated in Figure 1.

Figure 1: Sample stimuli from the experiment.

Each participant was seated approximately 65 cm in front of the eye tracker. The procedure began with a practice session, which involved a sample stimulus to familiarize participants with the task. In the experiment session, participants were shown the stimuli one by one. As each image was displayed, participants provided verbal descriptions, which were recorded. The order in which the stimuli were presented was randomized to avoid any sequence effects. Participants had control over the pace of the experiment. They proceeded to the next stimulus at their own pace, without any time constraints, by pressing a key on the keyboard.

Analysis

For the analysis, we removed four stimuli scenes (stimulus no. 7, 53, 65, and 22 in TPRS stimuli of 71 stimuli images) as they were significant outliers in statistical box-plot inspections. The dependent variables consisted of the response time (RT) and three eye movement variables: mean fixation duration, total gaze duration, and fixation count. We annotated the transcriptions by labeling the spatial terms, including adpositions and words, with locative case markers.

Results

Eye Movements on Figure, Ground and Text

We report the results for mean fixation durations (i.e., the average duration of single fixations), total gaze durations (i.e., the total duration of visual inspection), and mean fixation count (i.e., the average number of fixations). The results are shown in Table 1.

Table1: Eye movements on the Figure, Ground, and Text.

	Mean Fix. Dur. (ms)	Total Gaze Dur. (ms)	Mean Fix. Count
Figure	251.3 (29.5)	32,705.3 (12,294.3)	131.4 (50.8)
Ground	233.0 (22.2)	24,315.0 (12,497.9)	104.4 (52.0)
Text	219.4 (23.2)	6,897.9 (2,540.9)	31.3 (10.6)

The numbers in parentheses show standard deviation (SD) values.

First, a repeated measures ANOVA was conducted to determine if there were statistically significant differences among the mean fixation durations on Figure, Ground, and Text. Data were normally distributed, measured by the Shapiro-Wilk test (*ps* = 0.06, *ps* = 0.24, and *ps* = 0.21, respectively). The assumption of sphericity was not violated, as measured by the Mauchly test of sphericity, $p = 0.84$. Analysis revealed significant differences, $F(2,134) = 30.70$, $p < .001$, $\eta^2 = 0.31$. Post hoc analysis with Holm adjustment revealed that the mean fixation duration was significantly longer on Figure than on Ground ($M = 17.6$, $p < .001$, $d = 0.52$). The mean fixation duration on Figure was also longer than on Text ($M = 32.1$, $p < .001$, $d = 0.95$). Finally, the mean fixation duration was significantly longer on Ground than on Text ($M = 14.5$, $p < .001$, $d = 0.43$).

Second, we calculated the total gaze durations to find the total time spent on those constituents in the stimuli. A repeated measures ANOVA was conducted to test for significant differences. Data were normally distributed for Figure and Ground, as measured by the Shapiro-Wilk test (*ps* = 0.83, *ps* =0.33, respectively). The assumption of sphericity was violated, as measured by Mauchly's sphericity test, $p = <$

.001. Therefore, a Greenhouse-Geisser correction was applied (ε = 0.76). Significant differences were obtained in total gaze duration, $F(1.5,102) = 106$, $p < .001$, $\eta^2 = 0.61$. Post hoc analysis with Holm adjustment revealed that the mean gaze duration was significantly longer on Figure than on Ground (M= 8,955, $p < .001$, $d = 0.59$), and on Text (M = 26,436, $p <.001$, $d = 1.73$). Furthermore, the mean gaze duration was significantly longer on Ground than on Text (M = 17,480, $p < .001$, $d = 1.15$).

Finally, we analyzed fixation counts to check for the presence of a pattern similar to the duration results. Repeated measures ANOVA was performed to test the results. Data were normally distributed for Figure and Ground (ps = 0.051, ps =0.46, respectively). The assumption of sphericity was violated, as assessed by the Mauchly test of sphericity, p = <.001. Therefore, a Greenhouse-Geisser correction was applied (ε = 0.74). The analysis returned significant differences, $F(1.5, 99) = 94.7$, $p < .001$, $\eta^2 = 0.59$. Post-hoc analysis with a Holm adjustment revealed that fixation count was significantly higher on the Figure than the Ground (M = 29.2, $p < .001$, $d = 0.46$) and Text (M = 102, $p < .001$, d = 1.62). Moreover, fixation count was significantly higher on Ground than Text (M = 72.9, $p < .001$, $d = 1.16$).

Overall, the participants inspected the Figure more frequently and longer than the Ground, and the Ground than the Text. Nevertheless, the analysis of fixation counts does not contribute much to interpreting the results, since the Text was already expected to attract fewer fixations due to its short size and repetition among the stimuli. On the other hand, a major finding of the fixation analyses is the difference between the mean fixation durations of single fixation, in which the fixations on the Figure were longer on average than the fixations on the Ground. We further investigated participants' eye movements with participants' utterances.

Spatial Terms Variety Analysis

The participants produced 2,412 utterances in total, consisting of 4,940 words to describe the stimuli. The mean number of words used to express each stimulus was 2.05 (SD = 0.34, *range* = 1.37-3.19). Approximately 9% of the utterances (N = 210) included expressions that could not be labeled as static spatial expressions; therefore, they were omitted from the analysis. The remaining 2,202 utterances (approximately 91% of the total data) were included in the analysis.

Further analysis of utterances showed that about one-third (N = 776) included suffixed locative case markers with words, indicating that locative case markers were of notable use in the Turkish spatial language. Therefore, we included words with locative case markers as spatial terms in the analyses. The remaining utterances, which did not include locatives (N = 1,426 of 2,202), had adpositions such as *içinde* 'in' and *üstünde* 'on.'

The TPRS stimuli provided a rich set of spatial configurations such that the participants used a wide range of spatial descriptions in the utterances. In the following, we present example descriptions from the data. Descriptions typically consisted of spatial terms (e.g., adpositions and words with locative case markers LOCs) and reference objects. For instance, the utterance *masanın altında*. 'under the desk' consisted of the spatial term *alt-ın-da* 'under-GEN-LOC' and the reference object *masa-nın* 'the desk-GEN'.

The participants used various spatial terms to express the relationship between the Figure and the Ground. For example, each participant produced a single spatial expression (usually a sentence) to describe the relationship between the envelope and the stamp in Figure 2, while the descriptions of 36 participants consisted of nine different spatial terms (2).

Figure 2: Sample stimulus from the experiment.

(2)
üzerinde 'on-GEN-LOC'
üstünde 'above/on-GEN-LOC'
sağ üst köşesinde 'in the upper right corner-GEN-LOC'
arka tarafında 'at the back side-GEN-LOC'
arkasında 'behind-GEN-LOC'
sağ köşesinde 'in the upper right corner-GEN-LOC'
sağ üstünde 'in the upper-GEN-LOC right'
sol üst köşesinde 'in the upper left corner-GEN-LOC'
yanında 'beside/in its side-GEN-LOC'

For each stimulus, we transcribed and identified the spatial terms used in the descriptions to obtain the variety of the spatial terms. Then we divided the stimuli into two groups taking the mean number of different spatial terms used for the stimuli as the threshold for the division. Consequently, the stimuli scenes, which were labeled "high variety" (N = 25, M = 10.2, SE = 0.60), were the ones with a larger number of spatial terms than the average in their descriptions. Similarly, the "low variety" (N = 42, M = 4.29, SE = 0.22) were the ones with a lower number of spatial terms than the average, $t(65) = -2.76$, $p < .001$. Figure 3 shows two examples from the high variety and low variety stimuli.

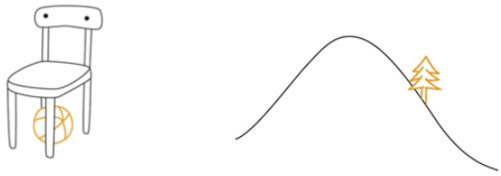

Figure 3: Spatial terms variety conditions: High variety on the right and low variety on the left.

A one-way ANOVA test was conducted to determine whether the mean response time (RT) differed between conditions of spatial terms variety (i.e., high variety and low variety). There was no significant difference in RT between the two groups of utterances. The following section presents analyses of eye movement variables. The mean duration of single fixations, total gaze durations, and fixation counts are presented in Table 2.

Table 2: Eye movement in low and high variety stimuli.

	Mean Fix. Dur.		Total Gaze Dur.		Mean Fix. Count	
	Low	High	Low	High	Low	High
F	246.0 (26.6)	260.0 (32.5)	32,159.5 (13,324.7)	33,622.3 (10,533.3)	131.9 (54.9)	130.6 (44.5)
G	228.9 (21.0)	239.9 (22.7)	19.971 (10,644)	31,611.8 (12,145.8)	87.7 (46.6)	132.3 (49.4)

The numbers in parentheses show standard deviation (SD) values.

A one-way multivariate analysis of variance (MANOVA) was conducted to test the differences between mean fixation durations on the Figure and Ground between the two levels of spatial terms variety (high variety vs. low variety). Using Pillai's trace, the analysis revealed a significant effect of spatial terms variety on the mean fixation duration on the Figure and the Ground (F_{avg} and G_{avg}), $V = 0.093$, $F(1, 65) = 3.27$, $p < 0.05$. Before conducting further follow-up ANOVAs, the homogeneity of the variance assumption was tested. Based on the box test for the equivalence of the covariance matrices, the homogeneity of the variance assumption was considered satisfied ($p = 0.71$). Two one-way ANOVAs were conducted on dependent variables (F_{avg} and G_{avg}), as follow-up tests to MANOVA. Data were normally distributed for each group and evaluated using the Shapiro-Wilk test ($ps = 0.26$, $ps = 0.18$). There was homogeneity of variances, as assessed by Levene's homogeneity test of variances ($p = 0.29$). The tests revealed that the difference in mean fixation duration on the Figure (F_{avg}) between high variety and low variety spatial terms was marginally significant, $F(1, 65) = 3.71$, $p = .059$, $\eta^2 = 0.054$. Tukey's post hoc analysis revealed that the mean increase from low variety condition to high variety condition was marginally significant ($M = -14.1$, $p = 0.059$, $d = -1.93$). Moreover, the mean fixation duration on the Ground (G_{avg}) was significantly different between high variety and low variety conditions of spatial terms, $F(1, 65) = 4.11$, $p < .05$, $\eta^2 = 0.06$. There were no outliers, as assessed by boxplots. Each group was normally distributed and evaluated using the Shapiro-Wilk test. As evaluated by Levene's homogeneity test of variances ($p = 0.86$), there was homogeneity of variances for G_{avg}. Tukey's post hoc analysis revealed that the mean fixation duration in low variety expressions was statistically lower than in high variety expressions ($M = -11.1$, $p < 0.05$, $d = -2.0$).

Secondly, differences in total gaze duration on the Figure (F_{sum}) and the Ground (G_{sum}) between high variety and low variety conditions were analyzed by a one-way MANOVA test. A significant difference was obtained, Pillais' trace = 0.22, $F(1, 65) = 9.02$, $p < .001$. The homogeneity of the variance assumption was also tested. Based on the Box test for the equivalence of covariance matrices, the homogeneity of variance assumption was satisfied ($p = 0.42$). However, separate univariate tests revealed that total gaze duration on Figure (F_{sum}) was not significantly different between high variety and low variety conditions of spatial terms variety. A one-way analysis of variance was conducted to assess whether there were differences in total gaze duration on the Ground. The results showed that the total gaze duration on Ground (G_{sum}) was statistically significantly different between high variety and low variety spatial terms, $F(1, 65) = 16.9$, $p < 0.001$, $\eta^2 = 0.21$. Each group was normally distributed, as assessed by the Shapiro-Wilk test ($ps = 0.33$, $ps = 0.22$). As revealed by Levene's test for homogeneity of variances ($p = 0.15$), there was homogeneity of variances in G_{sum}. Tukey post hoc analysis revealed that the mean fixation duration in low-variety conditions was marginally shorter than the high-variety expressions ($M = -11,640$, $p < 0.001$, $d = -4.11$)

Finally, differences in fixation counts on Figure (F_{count}) and Ground (G_{count}) were analyzed with one-way MANOVA. A statistically significant MANOVA effect was obtained, Pillais' Trace = .18, $F(1, 65) = 6.87$, $p < 0.05$. The homogeneity of the variance assumption was satisfied ($p = 0.56$). No significant difference was obtained between high and low variety spatial terms in mean fixation counts on the Figure (F_{count}). On the other hand, total fixation count on the Ground (G_{count}) was statistically significantly different between high variety and low variety, $F(1, 65) = 13.7$, $p < 0.001$, $\eta^2 = 0.17$. Each group was normally distributed, as assessed by the Shapiro-Wilk test ($ps = 0.33$, $ps = 0.74$). Levene's test for homogeneity of variances ($p = 0.51$) showed homogeneity of variances for G_{count}. Tukey's post hoc analysis revealed that fixation counts on low-variety expressions were marginally lower than high-variety expressions ($M = -44.5$, $p < 0.001$, $d = -3.70$)

In summary, the analyses revealed that the participants spent more time (cf. gaze duration) on the stimuli when the descriptions exhibited high variety. We also found that the mean durations of single fixations were longer in high-variety descriptions. Consequently, the findings support the hypothesis that the variety of spatial descriptions used to describe the scenes in the TPRS stimuli is related to oculomotor variables, which may indicate the cognitive processes that take place during the descriptions. Nevertheless, the findings apply to mean differences in gaze durations and fixations counts on the Ground, and overall mean differences in fixation durations (G_{sum}, G_{count}, F_{avg}, G_{avg}). We did not obtain a significant difference between high variety and low variety spatial terms in gaze durations and

mean fixation counts on the Figure (F_{sum}, F_{count}). Furthermore, the average number of spatial terms per participant and stimuli was similar between the stimuli (M = 2.05). In other words, the participants described the scenes in approximately two words. Therefore, the differences in the count of fixations and duration of the stimuli were not caused by the number of spatial terms the participants produced; instead, it was the variety of spatial terms. When the variety of spatial terms was wider, the fixation duration (mean and total) on the stimuli was longer, and the fixation count on the stimuli was larger.

Discussion and Conclusion

The exploration of cross-linguistic similarities and differences in spatial expression use has been a significant topic of research over the past several decades. Various studies, including those utilizing the TPRS battery (Bowerman & Pederson, 1992; Carstensen et al., 2019; Levinson et al., 2003), have provided insights into these aspects. The present study contributes to this field by investigating the use of spatial terms by Turkish native speakers as they verbally describe objects located in visual scenes. While numerous studies have examined spatial expression in a range of languages, Turkish, particularly in its spatial language characteristics, remains relatively understudied. Limited research, such as that by Sumer et al. (2012) and Atak (2018), delved into locative case markers and the usage of spatial terms in Turkish. Some others, such as Karadöller et al (2022) and Özer et al. (2023) studied Turkish spatial language comprehension in the context of gestures and sign language. Our study extends this research by empirically exploring the intersection between spatial language and visual processing, focusing on how the polysemy of spatial terms affects participants' eye movements while they visually inspect and verbally describe spatial scenes.

Additionally, our study touches on a longstanding issue in artificial intelligence, specifically the prediction of appropriate locative expressions for given spatial scenes – a problem known as the decoding and encoding dilemma, still relevant in conversational domains like Human Robot Interaction (Liu, Xiao, & Chen, 2022). While multimodal generative AI approaches are promising engineering solutions, they offer limited insights into human spatial cognition. This gap highlights the need for data from diverse languages to tackle the cross-linguistic diversity challenge.

A notable aspect of our study is its focus on static scenes, contrasting with most of the Turkish research which has emphasized dynamic spatial language (e.g., Johanson & Papafragou, 2014). This approach allowed participants to utilize locative suffixes, a finding not commonly reported in studies of Turkish spatial language.

Our analyses of participant eye movements revealed that the mean durations of single fixations were longer on the Figure than on the Ground. Further analysis showed a significant correlation between the variety of spatial terms used and eye movement variables. Specifically, gaze duration and mean fixation durations were longer for stimuli associated with a wider variety of spatial terms. These findings support the hypothesis that the diversity of spatial terms used in describing scenes is linked to oculomotor variables, potentially indicating underlying cognitive processes during language production and comprehension.

However, linguistic data alone may not sufficiently reveal patterns in spatial cognitive processes, as suggested by previous studies (Le Guen, 2011). Therefore, we utilized complementary methods, including oculomotor variables, to investigate the semantic categorization of spatial terms. This approach helped overcome limitations inherent in behavioral data, such as response times. Notably, we found no significant relationship between the variety of spatial terms or Core vs. Non-Core categorization and participants' response times. This suggests that gaze analysis might provide more robust data for studying spatial expressions in visual scene descriptions than response times alone.

Nevertheless, our study faces limitations. The variety of spatial terms was quantified simply as the number of different instances used by participants. Future research could employ more sophisticated methods, like weighted models, to capture the heterogeneity in spatial term usage. As for the statistical analysis, generalized linear model analyses might be used to address the individual differences among participants and items more effectively than ANOVA methods. Furthermore, while TPRS served as a comprehensive set of topological relations, its focus on static scenes and the unbalanced nature of spatial type subtypes may limit the generalizability of our findings.

Future studies should consider using a broader range of stimuli, including enriched stimuli for Core and Non-Core distinctions. In Turkish, the relational information can be represented in various ways, such as using a locative case marker rather than a postposition. These differences might have an impact on participants' gazing sequence of objects (Figure and Ground). A further investigation of the time-course analysis of eye movements could provide a deeper understanding of the choice and variety of spatial terms used and a more comprehensive picture of perceptual and cognitive processes involved, besides contributing to our understanding of how Turkish encodes spatial relations.

Finally, our findings raise questions about the generalizability to current theories concerning the impact of non-linguistic constraints on spatial prepositions (Coventry, et al., 2023). Addressing these questions could further elucidate the complex interplay between linguistic, visual, and spatial processes in spatial cognition and language use.

Acknowledgments

This study has been partially supported by Jagiellonian University Strategic Programme Excellence Initiative Priority Research Area DigiWorld (ID.UJ) under the project title "Cognitive Aspects of Interaction and Communication in Natural and Artificial Agents". Thanks METU (Middle East Technical University, Turkey) HUMATE (Human Machine Teaming Lab) for supporting data collection.

References

Atak, A. (2018). Türkçede Uzamsal Dilin Konumlanış Açısından İncelenmesi, PhD thesis, Ankara Üniversitesi.

Bowerman, M. (1996). Learning how to structure space for language: A crosslinguistic perspective. In P. Bloom et al. (Eds.), *Language and space*, (pp. 385–436). MIT Press.

Bowerman, M. & Pederson, E. (1992). Topological relations picture series. In Stephen C. Levinson (Ed.), *Space stimuli kit 1.2* (pp. 51). Nijmegen. doi: 10.17617/2.883589.

Carstensen, A., Kachergis, G., Hermalin, N., & Regier, T. (2019). "Natural concepts" revisited in the spatial-topological domain: Universal tendencies in focal spatial relations. In A.K. Goel, C.M. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Conference of the Cognitive Science Society* (pp. 197-203). Montreal, QB.

Choi, S., & Hattrup, K. (2012). Relative contribution of perception/cognition and language on spatial categorization. *Cognitive Science, 36*(1), 102–129. doi: 10.1111/j.1551-6709.2011.01201.x.

Coventry, K.R., & Garrod, S.C. (2004). *Saying, seeing and acting: The psychological semantics of spatial prepositions*. Psychology Press.

Coventry, K.R., Gudde, H.B., Diessel, H. et al. (2023). Spatial communication systems across languages reflect universal action constraints. *Nature Human Behaviour*, 7, 2099–2110. doi: 10.1038/s41562-023-01697-4.

Duchowski, A.T. (2017). Eye tracking methodology: Theory and practice (3rd ed.). Springer, Nature. doi: 10.1007/978-3-319-57883-5.

Glenberg, A. M., & Robertson, D. A. (2000). Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of Memory and Language, 43*(3), 379-401. doi: 10.1006/jmla.2000.2714.

Griffin, Z.M., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science, 11*(4), 274-279. doi: 10.1111/1467-9280.00255.

Jackendoff, R.S. (1985). *Semantics and cognition*. MIT Press.

Johannes, K., Wang, J., Papafragou, A., & Landau, B. (2015). Similarity and variation in the distribution of spatial expressions across three languages. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, (pp. 997-1002). Pasadena, United States.

Johanson, M., & Papafragou, A. (2014). What does children's spatial language reveal about spatial concepts? Evidence from the use of containment expressions. *Cognitive Science, 38*(5), 881-910.

Karadöller, D.Z., Sümer, B., Ercenur, Ü. & Özyürek, A. (2022). Sign advantage: Both children and adults' spatial expressions in sign are more informative than those in speech and gestures combined. *Journal of Child Language*. doi: 10.1017/S0305000922000642.

Lakoff, G. (1986). A figure of thought. *Metaphor and Symbol, 1*(3), 215-225.

Lakusta, L., Hussein, Y., Wodzinski, A., & Landau, B. (2021). The privileging of 'Support-From-Below'in early spatial language acquisition. *Infant Behavior and Development*, 65, 101616. doi: 10.1016/j.infbeh.2021.101616.

Landau, B., & Jackendoff, R. (1993). Whence and whither in spatial language and spatial cognition? *Behavioral and Brain Sciences, 16*(2), 255-265.

Langacker, R. W. (1986). An introduction to cognitive grammar. *Cognitive Science, 10*(1), 1-40.

Le Guen, O. (2011). Speech and gesture in spatial language and cognition among the Yucatec Mayas. *Cognitive Science, 35*(5), 905-938.

Levinson, S., 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, 79*(3), 485-516.

Liu, M., Xiao, C., & Chen, C. (2022). Perspective-corrected spatial referring expression generation for human–robot interaction. *IEEE Transactions on Systems, Man, and Cybernetics: Systems, 52*(12), 7654-7666.

Liversedge, S. P., & Findlay, J. M. (2000). Saccadic eye movements and cognition. *Trends in Cognitive Sciences, 4*(1), 6-14. doi: 10.1016/S1364-6613(99)01418-7.

Majid, A., Bowerman, M., Kita, S., Haun, D. B., & Levinson, S. C. (2004). Can language restructure cognition? The case for space. *Trends in Cognitive Sciences, 8*(3), 108-114. doi: 10.1016/j.tics.2004.01.003.

Majid, A., Jordan, F., & Dunn, M. (2015). Semantic systems in closely related languages. *Language Sciences*, 49, 1-18. doi: 10.1016/j.langsci.2014.11.002.

Mayberry, M.R., Crocker, M.W., & Knoeferle, P. (2009). Learning to attend: A connectionist model of situated language comprehension. *Cognitive Science, 33*(3), 449–496. doi: j.1551-6709.2009.01019.x.

Meyer, A.S., Sleiderink, A.M., & Levelt, W.J.M. (1998). Viewing and naming objects: eye movements during noun phrase production. *Cognition, 66*(2), B25–B33. doi: 10.1016/S0010-0277(98)00009-2.

Özer, D., Karadöller, D.Z., Özyürek, A., & Göksun, T. (2023). Gestures cued by demonstratives in speech guide listeners' visual attention during spatial language comprehension. J*ournal of Experimental Psychology: General, 152*(9), 2623–2635. doi: 10.1037/xge0001402.

Sumer, B., Zwitserlood, I., Perniss, P. M., & Ozyurek, A. (2012). Development of locative expressions by Turkish deaf and hearing children: Are there modality effects? In A.K. Biller et al. (Eds.), *Proceedings of the 36th Annual Boston University Conference on Language Development* (Vol. 1, pp.568 – 580). Cascadilla Press.

Talmy, L. (1983). How language structures space. In *Spatial orientation: Theory, research, and application* (pp. 225-282). Boston, MA: Springer US.

Yun, H., & Choi, S. (2018). Spatial Semantics, Cognition, and Their Interaction: A Comparative Study of Spatial Categorization in English and Korean. *Cognitive Science, 42*(6), 1736–1776. doi: 10.1126/science.7777863.