8-2017

Spatial Relations and Natural-Language Semantics for Indoor Scenes

Stacy A. Doore
University of Maine, stacy.doore@maine.edu

SPATIAL RELATIONS AND NATURAL-LANGUAGE SEMANTICS FOR INDOOR SCENES

BY

Stacy Anne Doore

B.A. University of Maine, 1999

B.S. University of Maine, 1999

M.S. University of Maine, 2010

A DISSERTATION

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

(in Spatial Information Science and Engineering)

The Graduate School

The University of Maine

August, 2017

Advisory Committee:
Kate Beard-Tisdale, Professor of Spatial Informatics, Advisor
Max Egenhofer, Professor of Spatial Informatics
Nicholas Giudice, Associate Professor of Spatial Informatics
Torsten Hahmann, Assistant Professor of Spatial Informatics
Werner Kuhn, Professor for Geographic Information Science, University of California, Santa Barbara

SPATIAL RELATIONS AND NATURAL-LANGUAGE SEMANTICS FOR INDOOR SCENES

By Stacy A. Doore

Dissertation Advisor: Dr. Kate Beard-Tisdale

An Abstract of the Dissertation Presented
in Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy
(in Spatial Information Science and Engineering)
August, 2017

Over the past 15 years, there have been increased efforts to represent and communicate spatial information about entities within indoor environments. Automated annotation of information about indoor environments is needed for natural-language processing tasks, such as spatially anchoring events, tracking objects in motion, scene descriptions, and interpretation of thematic places in relationship to confirmed locations. Descriptions of indoor scenes often require a fine granularity of spatial information about the meaning of natural-language spatial utterances to improve human-computer interactions and applications for the retrieval of spatial information. The development needs of these systems provide a rationale as to why—despite an extensive body of research in spatial cognition and spatial linguistics—it is still necessary to investigate basic understandings of how humans conceptualize and communicate about objects and structures in indoor space.

This thesis investigates the alignment of conceptual spatial relations and natural-language (NL) semantics in the representation of indoor space. The foundation of this work is grounded in spatial information theory as well as spatial cognition and spatial

linguistics. In order to better understand how to align computational models and NL expressions about indoor space, this dissertation used an existing dataset of indoor scene descriptions to investigate patterns in entity identification, spatial relations, and spatial preposition use within vista-scale indoor settings. Three human-subject experiments were designed and conducted within virtual indoor environments. These experiments investigate alignment of human-subject NL expressions for a sub-set of conceptual spatial relations (*contact*, *disjoint*, and *partof*) within a controlled virtual environment. Each scene was designed to focus participant attention on a single relation depicted in the scene and elicit a spatial preposition term(s) to describe the focal relationship.

The major results of this study are the identification of object and structure categories, spatial relationships, and patterns of spatial preposition use in the indoor scene descriptions that were consistent across both open response, closed response and ranking type items. There appeared to be a strong preference for describing scene objects in relation to the structural objects that bound the room depicted in the indoor scenes. Furthermore, for each of the three relations (*contact, disjoint,* and *partof*), a small set of spatial prepositions emerged that were strongly preferred by participants at statistically significant levels based on the overall frequency of response, image sorting, and ranking judgments. The use of certain spatial prepositions to describe relations between room structures suggests there may be differences in how indoor vista-scale space is understood in relation to tabletop and geographic scales. Finally, an indoor scene description corpus was developed as a product of this work, which should provide researchers with new human-subject based datasets for training NL algorithms used to generate more accurate and intuitive NL descriptions of indoor scenes.

ACKNOWLEDGEMENTS

This dissertation is dedicated to my husband, Dr. Brian Doore and our seven children who have supported me in its successful completion. Without their understanding, patience, and encouragement during every stage of the process, this life goal achievement would not have been possible. I would like to also recognize the significant contribution of Dr. Lisa Walton, whose gentle mentoring during my early graduate school experience provided the friendship and guidance I needed at key points in the journey.

I would like to extend my sincere gratitude to my advisor, Dr. Kate Beard who has seen me through a masters and a doctoral degree over the past ten years. Her generous spirit, her approach to collaborative scholarship, and devoted mentoring have provided more opportunities than I can count to grow and learn, both an academic and as a person. I cannot thank her enough for her patience and persistence in helping me achieve, what at times, seemed impossible. I would like to thank the members of my dissertation committee, Dr. Nicholas Giudice, Dr. Torsten Hahmann, Dr. Max Egenhofer, and Dr. Werner Kuhn whose comments and guidance helped to shape the initial research and finished product. This dissertation would not have been completed without the support from many other academic role models and friends, including Dr. Susan Gardner and the leadership team from UMaine’s ADVANCE Rising Tide program, Dr. Matthew Dube, and my friends at the VEMI lab who provided support in creating the virtual environments used in the experiments. Finally, thank you to all of my friends who have supported this effort through providing meals, childcare, and cheerleading just when it was needed the most.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Automated scene description is a challenging problem that requires a combination of vision and language tasks. Conceptually, this type of intelligent system analyzes an image of a scene for its visual content and generates a text or audio description that conveys salient aspects of the image. Systems for describing scenes have been designed to assist in region analysis, pattern recognition and object identification in both indoor and outdoor settings. These systems identify scene objects and their attributes to produce descriptions in the form of short phrases of nouns and adjectives (e.g., wooden bench, a large tree, a red couch, a blue chair). If the objective is to describe a scene for someone who is visually impaired, scene descriptions consisting of unstructured lists of objects are not particularly useful. A question that emerges then is, what constitutes a good scene description? Bernardi et al. (2016) in a recent review article suggest that a "good" image description has competing requirements to be both comprehensive and concise (include all salient entities and their relations to one another), and to be linguistically correct (i.e., have grammatically, well-formed sentences). Bernardi et al. (2016) also state that the automatic generation of image descriptions requires an expert level understanding of how people describe images. Gapp (1994) adds a requirement that scene descriptions must attend to the correct natural-language treatment of spatial relations in order to be considered accurate and effective. In combination, these requirements point to the need for a correct and concise phrasing of spatial relations in natural language. Research in geographic information science (GIScience) has formally identified sets of qualitative spatial relations between objects of different dimensions and within embedding spaces of different dimensions. This thesis investigates the problem of translating qualitative spatial relations (topological, containment, and proximal) identified between objects in symbolic indoor scene representations to appropriate linguistic terms for generating descriptions of indoor scenes.

1.1. Motivation

The following two scenarios provide a motivation for this work.

Scenario 1: Allison, who has a vision impairment, is using a social media platform to share information about her busy life with friends, who also post stories and images about events in their lives. Allison recently increased her use of the social media site because of a new automated alternative text feature that embeds captions read by her speech-access program. She wants to apply it to images her friend just posted of her new apartment. An example of a caption that Allison receives is "the image may contain: table, living room, and indoor." (Figure 1.1). This approach treats the indoor scene as a container that has a list of objects with a binary context of either an indoor or outdoor setting. While it is useful to know what is in the scene, there is no other information available about the relationships between the objects to provide a mental image of the interior scene for someone who cannot directly see it.

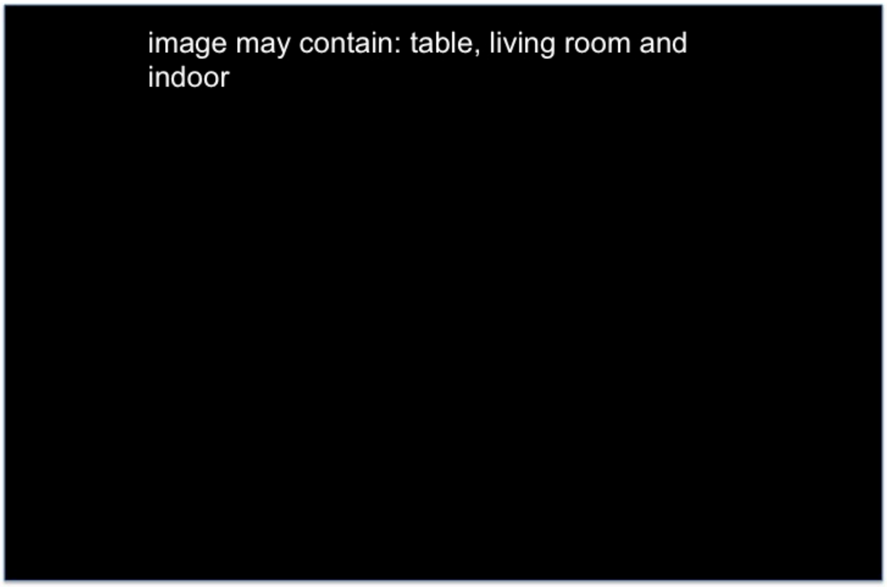


image may contain: table, living room and indoor

Figure 1.1: *Problem Scenario 1 – Description of indoor scene.*

Scenario 2: Imagine Allison, attending a conference in a large hotel, is planning to meet a friend for coffee in the first floor reception area of the conference hotel. Her friend says to meet near the central sculpture (Figure 1.2). Her phone contains images of the reception area and she asks her digital assistant to describe the scene of the reception area surrounding the sculpture. The application on her phone processes the images, identifying objects in the image and spatial relations between them. The result of this initial processing is a set of plausibly identified objects, interior structures, and some geometric and topological relationships between them. Next, the digital assistant translates this information into a concise and correct NL scene description for Allison from her preferred frame of reference.

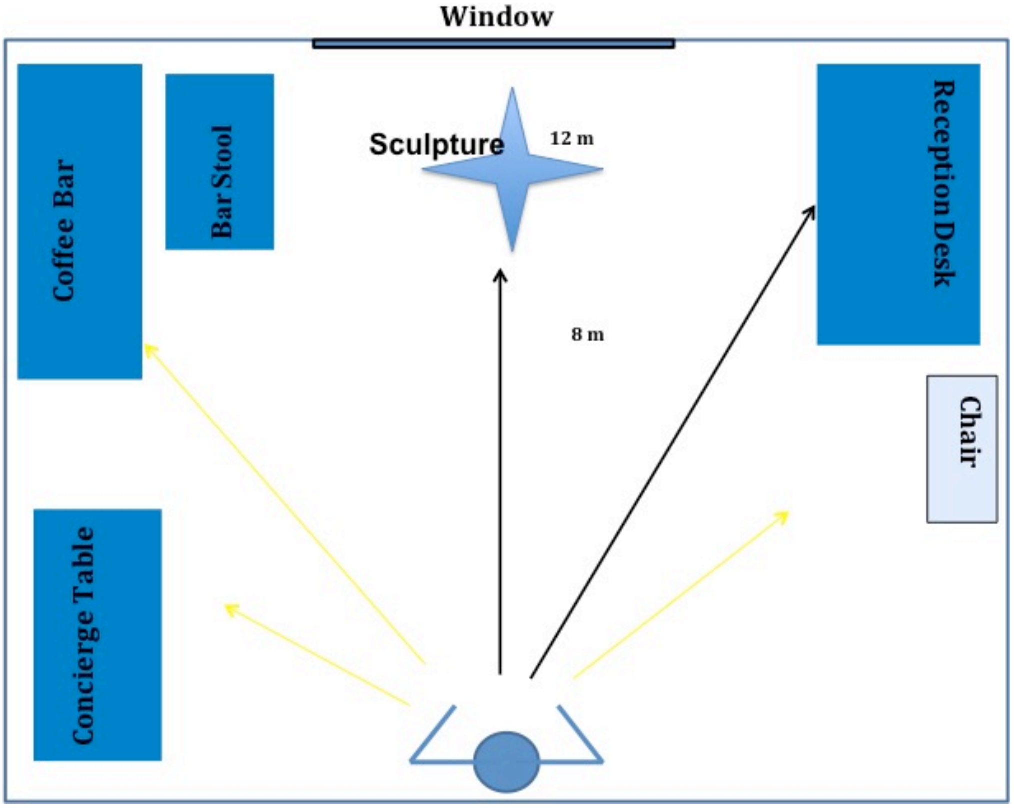


Figure 1.2: *Problem Scenario 2 –Hotel lobby representation.*

In these scenarios, the settings are indoor spaces and we assume that these indoor spaces are represented by one or more images. A first major challenge in the above scenarios, is the computer vision task of converting the images into identifiable objects with some appropriate attributes and appropriately specified relations between objects. A second major challenge, is the natural-language production task of describing the scene. The goal is to move beyond the simple captions of the first scenario to produce scene descriptions with sufficient detail for a person with a sensory constraint (i.e., vision impairment) to understand the composition of the indoor scene.

This thesis addresses the natural-language component of the problem for scenes set in indoor spaces and particularly *vista*-scale spaces (Montello, 1993). A vista-scale space is defined as a space larger than the human body that can be perceived from a single perspective. Contextual factors, landmarks, and scene boundaries are different in such indoor spaces as compared to larger, outdoor settings. Given differences between indoor and outdoor settings we might expect difference in how such scenes are described.

The question becomes what is an appropriate level of detail of spatial-information for describing indoor spaces? What are the guiding structures for providing a concise and accurate description of spatial information that supports the nonvisual interpretation of a scene without the risk of cognitive overload?

The goal of this thesis is to align NL specifications to effectively describe spatial concepts and relations within a simple indoor environment*.* *In particular, the thesis focuses on identifying a controlled vocabulary of spatial prepositions for a small set of spatial concepts to convey spatial relations between objects in indoor environments and used in automated scene descriptions.*

1.2. Research Questions and Experiments

To meet the requirements of the motivational scenarios, there needs to be an alignment of conceptual and linguistic structures to allow a system to generate automated descriptions of NL indoor scenes. Human use of spatial prepositions is influenced by various factors including object categories and functions, as well as topological properties of the objects. Research on the use of spatial prepositions undertaken at spatial scales other than indoor vista-scale space has found object function, expression length, and setting context all contribute to spatial preposition choice.

This thesis investigates factors influencing spatial preposition use in indoor vista scale spaces through the following set of research questions::

1. How do people conceptualize and communicate spatial relations when they describe an indoor scene in natural-language?
2. What spatial prepositions do people use to describe topological and conceptual relations between objects in a room?
3. What are preferred spatial prepositions to express spatial relations between objects in indoor scenes?
4. Do descriptions of indoor scenes differ based on sensory constraints of the intended recipient of the description?
5. What role does object function serve in the choice of spatial prepositions in the description of indoor scenes?
6. Are there differences in the preferences of level of specificity in spatial prepositions used in scene descriptions based on room context factors?

1.3. Scope of Thesis

The scope of this thesis, from a theoretical perspective, builds upon a corpus of knowledge regarding the nature, use, and interpretation of spatial prepositions from spatial information science, spatial cognition, and spatial linguistics but focuses exclusively on object relations in vista-scale indoor scenes. From an application perspective, this thesis focuses on the semantics of spatial prepositions used to describe relations and objects within indoor environments in order to enhance information systems that communicate spatial information.

Although the theoretical framework is based on a Naive Geography (Egenhofer and Mark, 1995) approach that investigates the alignment of spatial cognition and linguistics, the specific focus is on the interpretation of human spatial expressions within

English natural-language discourse. The intent here is not to make generalizations across other languages or cultures. The aim is not to create a computational model, and it is beyond the scope of this thesis to provide the specifications and design of such a model. This thesis does contribute to the existing literature on the semantics of spatial prepositions, specifically in a new environment, indoor vista-scale space, and to using human-subjects testing to inform the design of more effective and accurate systems for the automated descriptions of indoor scenes.

1.4. Approach

This thesis research makes use of virtual 3D indoor spaces in which to investigate human-subject interpretations of indoor scene relationships and their choices of spatial prepositions. In the virtual environment, a room is treated as a container object that is comprised of room structure objects (e.g., wall, ceiling, floor) that enclose a void (Brodaric, Hahmann, Gruninger, 2017). This alternative perspective of the room as a set of objects and a void is adopted to better represent the relations between the objects contained within the room. A scene description framed simply as a list of objects contained within a room provides no information on the spatial arrangement of objects. The description recipient lacks critical information for forming a spatial model for subsequent reasoning and information retrieval. The types of descriptions proposed in the motivating problems require a finer level of information because of the need to describe spatial relations between objects contained in a 3D object (the room as the sum of room structures and the bounded void) that can either be in the void, or part of the room structures (Casati, Varzi, 1999; Hahmann, Brodaric, 2013; Brodaric et al, 2017) (Figure 1.3).

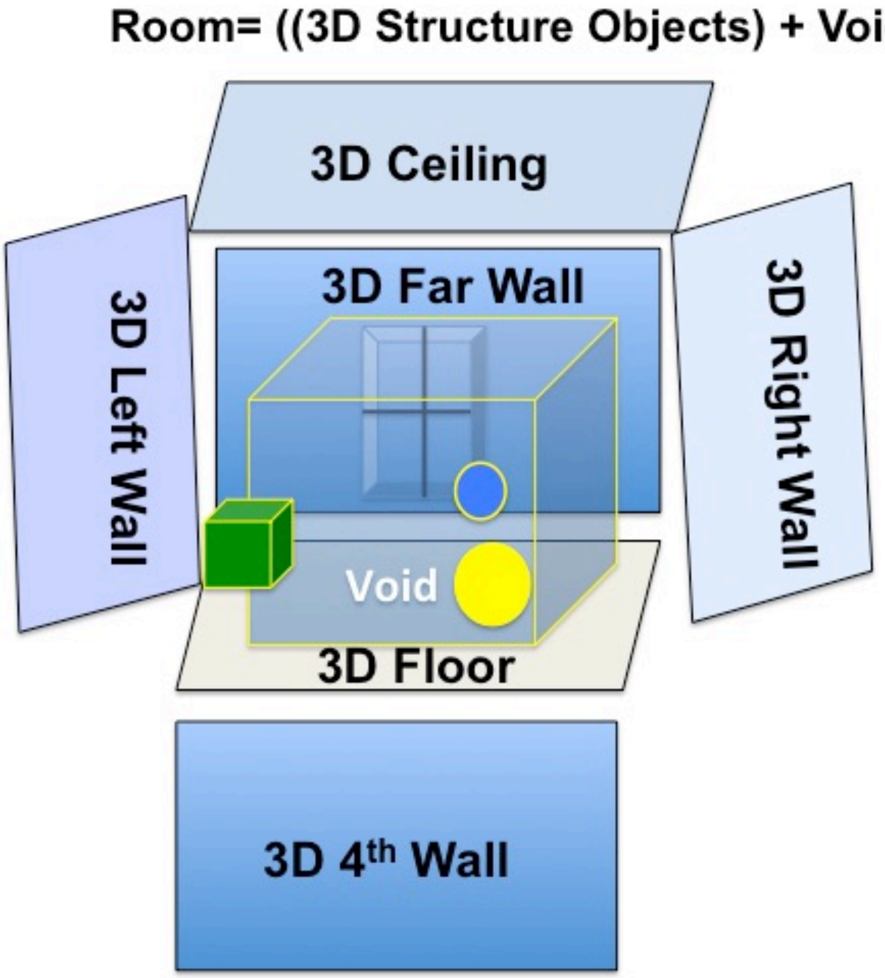


Figure 1.3: *Room represented as set of structure objects and the void.*

This representation uses a set of more conceptual than formal relations such as *contact, disjoint, and partof* to describe configuration of moveable objects (e.g., furniture) and structure objects (e.g., walls, windows, doors) in an indoor scene to build accurate and concise statements from the user preferred frame of reference. A complete rational for this approach is provided in Chapter 2.

In the behavioral experiments, participants are shown an indoor scene and asked to make judgments about spatial configuration and the preferred spatial language terms that most accurately describe that indoor scene.

- *Experimental Environment:* Virtual-reality generated images depict rooms with large, free-standing, regularly-shaped objects such as large pieces of furniture (e.g., bookcases, desks, chairs) and typical room structures (e.g., walls, windows, doors). Objects of interest are much larger than tabletop objects. The experimental room spaces include two sizes*:* a small vista space (10'x12') and a large vista space room size (20'x30'). These room sizes were selected based on findings from previous research that suggested different sizes of indoor vista scale spaces had a significant impact on scanning and search strategy performance (Pingel, Schinazi, 2014).
- *Experimental Image Prompts:* The virtual scenes were designed as simple indoor environments so the participants could easily determine their preferred terms for describing relations between moveable objects (e.g., furniture) and indoor room structural objects (e.g., walls and windows). The experiments used spatial expressions and preposition choices extracted from frequently used terms found in the re-analysis of indoor scene descriptions (Chapter 3). In Experiment-1, participants provided spatial propositions to fill in an open response prompt that matched the relation provided by the image of an indoor scene. In Experiment-2, participants sorted images into groups based on their perceived similarities. In Experiment 3, participants made judgments regarding similarity, clarity and preference of spatial prepositions based on images and text prompts.

1.5. Research Contributions

A major contribution of this research is new information pertaining to human natural-language descriptions of object relations within (real-world and virtual) vista-scale settings in indoor space. This work fills a research gap in understanding conceptual and linguistic structures in a scale space which has until recently, received much less attention than either tabletop or geographic spaces. This research contributes more information in the following areas:

(1) Identification of key object and structure categories and their spatial relations in the descriptions of indoor scenes.

(2) Statistically significant patterns of spatial preposition use as applied to spatial relations between objects and structures in indoor scene descriptions.

(3) Identification of preferred spatial prepositions associated with spatial relations in scene descriptions based on human-subject perceptions of preposition similarity, clarity, and preference.

1.6. Intended Audience

The intended audience of this thesis includes researchers and developers who are interested in the conceptualization of indoor space and the design of systems for the NL description of indoor scenes.

1.7. Organization of Remaining Chapters

The remaining chapters of this thesis are organized as follows:

Chapter 2 reviews relevant research related to spatial linguistic concepts and spatial relations in the context of automated NL descriptions as applied to indoor scenes.

Chapter 3 describes the results of a re-analysis of an existing dataset of scene descriptions that investigates patterns in entity identification, spatial relations, and the use of spatial preposition. This analysis also focuses on identifying contextual information and use preferences for spatial preposition in scene descriptions.

Chapter 4 describes the human-subjects experimental protocol and procedures within a virtual indoor environment. **Chapter 5** presents the results of the behavioral experiments. **Chapter 6** discusses collective findings from the analysis of scene descriptions (Chapter 3) and the human-subjects virtual-scene experiments (Chapter 5). It summarizes the major results and contributions of the dissertation, identifies the limitations, and postulates new questions and directions for future research.

CHAPTER 2

BACKGROUND AND RELATED WORK

This chapter reviews the background and research related to the description of indoor scenes. While the central problem is that of determining the essential properties and terms to generate accurate and concise Natural-Language (NL) scene descriptions, related topics include the properties of indoor space, conceptualization of spatial relations, and principles of spatial linguistics, as applied to descriptions of spatial configurations.

2.1. The Indoor Space Setting

Over the past 30 years, there have been increased efforts to represent and communicate spatial information about entities within indoor environments (DiManzo, Adorni, Giunchiglia, 1986; Riehle, Lichter, Giudice, 2008; Falomir, 2012; Li, Lee, 2013). As people in industrial societies spend an estimated 90% of their lives indoors (American Physical Society, 2008), the efficient representation of and communication about indoor space has become an active area of investigation for geographic information science. Automated annotation of information about indoor environments is needed for natural-language (NL) processing tasks, such as spatially anchoring events, describing objects in motion, scene descriptions, and interpretation of thematic places in relationship to confirmed geolocations. These efforts have also been driven by the demands of industries developing emerging technologies, such as NL scene descriptions for use in robotic automation, indoor navigation, and retail location-based services in indoor public spaces (Aditya et al., 2015; Bernardi et al., 2016).

The description of indoor scenes requires a fine granularity of spatial information about the meaning of NL spatial utterances to improve human-computer interactions and the retrieval of spatial information. Despite an extensive body of research in spatial cognition and spatial linguistics, understanding how people conceptualize and communicate about object relations in indoor space is still a difficult problem, and the focus of this dissertation. Specifically, this research investigates the roles that physical structure objects (e.g., walls, windows, doors, etc.) play in indoor scene descriptions, and how spatial relations are perceived and described in these spaces. The adopted frame of reference for the work conceptualizes a room, as a bounded space in which the boundaries (e.g., walls, floor, ceiling) are represented as objects and participate in relationships with other room objects (e.g., furniture).

This chapter provides the background and discussion of related work on the representation of indoor space based on conceptual and linguistic models as context for the dissertation work. First, it gives examples of current systems designed for automated NL descriptions of indoor scenes and highlights the areas where these systems have difficulty generating effective nonvisual descriptions for perceiving a spatial scene. Next, the key differences between indoor space relative to outdoor space are discussed to illustrate the impact of cognitive spatial concepts and sensory constraints that are associated with different spatial scales. Finally, prior approaches aligning spatial prepositions to spatial relations are discussed, as this motivates subsequent design choices and helps to explain findings elucidated in this dissertation.

2.2. Systems for the Description of Indoor Scenes

Emerging technologies including natural-language (NL) assistants are driving new applications for indoor environments that support robot automation, indoor navigation, and retail location-based services. Included in these developments are systems for automated descriptions of indoor scenes (Lin et al., 2015). Much of the recent work on representing and communicating about indoor space has focused on transitions from outdoor spaces to indoor spaces or on generating indoor route descriptions (Allen, 2000; Nothegger, Winter, Raubal, 2004; MacMahon, Stankiewicz, Kuipers, 2006). However, neither of the motivating scenarios (Chapter 1) involve these types of locomotion of spatial tasks but instead are focused on generating a concise description of an indoor scene. In this work, an *indoor scene* is defined as what objects can be perceived without significant locomotion as a cohesive and obvious entity set within a large-scale indoor space (Ruetschi, 2007).

An agent designed to generate automated descriptions of indoor scenes needs a way to collect spatial data from a variety of sources through multi-sensory channels, such as computer vision, localization sensor networks, and human question and answer input. Increasingly, existing spatial data of public indoor environments can be accessed as graph based representations of building information systems (e.g., Google Indoor Maps, Bing). The research on these types of systems for image description has largely focused on improving the capacity of image captioning systems to describe location-based objects and resources using a variety of neural-network models (Tran, et al., 2016; Vinyals, Toshev, Bengio, Ehran, 2015).

In contrast to these approaches, this dissertation research focuses on the use of spatial prepositions to convey specific information about relations between objects in indoor space. In an example approach to scene description, accessibility researchers at Facebook (www.facebook.com/accessibility) developed a system to automate image descriptions to specifically address the needs of social media users who are blind and vision impaired (BVI). The Automatic Alt-Text algorithm (Wu et al., 2017) used in this approach does not use a typical free-form sentence approach but instead restricts the scene description sentence to begin with the phrase: “Image may contain” followed by a list of general entity tags ordered into categories (people, objects, and setting; Figure 2.1).

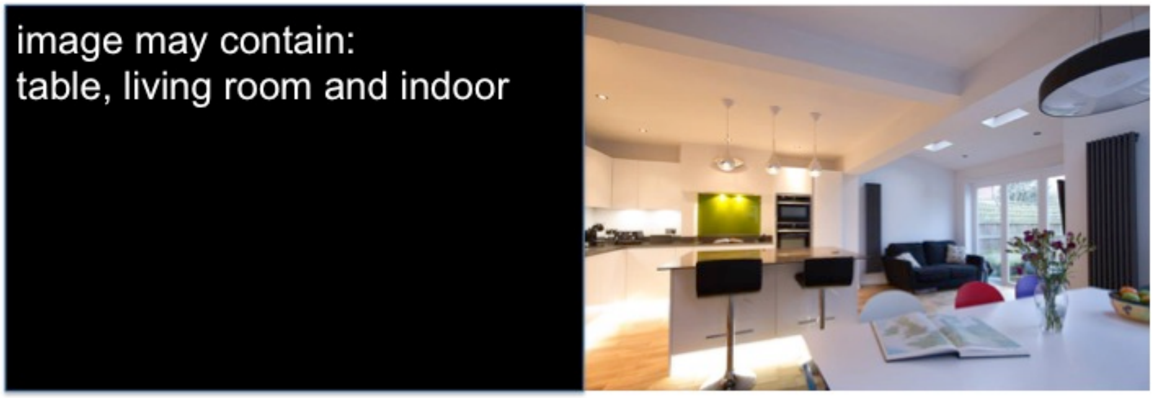


image may contain:
table, living room and indoor

Figure 2.1: *Automatic Alt-Text scene description for interior space (Wu et al., 2017).*

The design was chosen to reduce the level of uncertainty of scene objects and improve object identification accuracy, but the description provides no information on the spatial arrangement of objects. Evaluation of the scene description model by BVI users’ indicated the scene description was helpful but lacking information on object relations and spatial context within the scenes.

This example illustrates the limitations of using the *room as a container* model to describe objects in an indoor scene. If all objects are described as only a collection of entities contained in the room, the description recipient is missing critical information about the relations between the objects, and that information about the spatial configuration is unavailable for additional reasoning and information retrieval tasks. To overcome this barrier, a spatial model should include spatial relations not only between objects and the room as a whole, but also between objects and commonly identified parts of a room, such as individual wall surfaces.

Other recently developed systems for scene descriptions have begun to pay more attention to spatial relations. A system developed by Kulkarni et al. (2011) processes images to detect objects (person, chair, table), physical stuff (e.g., sand, water, grass), object and stuff modifiers (adjectives), and spatial relations in an image and generates text descriptions. An example description from their system is: “This is a photograph of one person and one brown sofa and one dog. The person is against the brown sofa and the dog is near the person and beside the brown sofa.”. This system captures pairwise spatial relations between objects but does not place these objects within a room context. Lin et al. (2015) developed a system particularly for indoor scenes. Their system processes RGB-Depth images, generates a scene graph that represent objects, attributes, and relations between objects, and then uses the scene graph and a sequence of semantic trees to generate multi-sentence descriptions through a learned grammar. The grammar is learned from a training set of RGB-D images annotated with descriptions provided by humans. The scene graph uses nodes to represent objects and defines three types of edges: attribute edges that link attributes to nodes, position edges that represent positions of objects relative to the scene (e.g. corner of room), and pairwise edges that describe relative positions between objects. Their position edge values are restricted to: *corner of room, front of camera, far-away from camera, center of room, left- of room,* and *right of room* and their pairwise object relations are the fixed set: *next-to, near, top-of, above, in-front-of, behind, to-left of,* and *to right of*. The authors do not specify how they arrived at this particular subset of relations. An example description from their system is: “In the kitchen there is a chair. A cabinet is behind the sofa. The sofa is near the chair”. While this system recognizes room parts, it is interesting to note that these are not included in the generated descriptions. Inclusion of room parts in a description is a notable difference between these descriptions and the human generated descriptions analyzed in Chapter 3. The work carried out in this dissertation represents a critical next step toward enhancing indoor scene descriptions by better understanding how humans perceive and describe indoor room objects and structures and the types of spatial expressions they employ in descriptions. Chapters 4 and 5 of the thesis take up the question of what constitutes reasonable linguistic expression for relationships between object pairs in indoor scene by asking human subjects to supply preferred linguistic expressions to relate them.

As illustrated in the examples above, information available to an intelligent agent may include various data structures that provide links between the metric, topological and network information to a linguistic model used for grounding linguistic descriptions of 3D spatial entities (Mozos et al., 2007). In order to present the desired level of spatial information, both the conceptual and linguistic models need to accommodate the user’s preferred frame of reference (e.g., room as a container of objects vs. relationship of objects to one another) and sensory constraints (e.g., emphasis on visual vs. nonvisual interface; Choi et al., 2013). Descriptions of indoor scenes must also convey an appropriate level of contextual information about the indoor space. Contextual information within spatial models is defined as “information gathered and used to enrich the knowledge about the user’s state, physical surroundings and capabilities of any mobile or assistive device” (Afouni, Ray, Claramunt, 2012 p.85). The dynamic nature of indoor settings makes this particularly challenging. In outdoor spaces, buildings and road networks do transform and move over time, however this rarely occurs within a short timespan, except in cases of natural or man-made disasters. In contrast, indoor spaces are inherently dynamic and change within a short temporal scale, and the context for their usage and function can vary greatly based on often competing user needs and tasks. Moveable objects such as furniture, and to some extent, architectural elements such as walls and hallways, can be reconfigured quickly within a span of minutes to days. The dynamic nature of indoor objects and spaces makes it difficult to create the same tools and NL query phrases for the retrieval of spatial information available to consumers in outdoor space.

2.3. Naive Geography

Despite an extensive body of research in spatial cognition and spatial linguistics, automated scene description of spatial configurations still requires basic understandings of how humans conceptualize and communicate about objects and structures within different spaces. The developed systems described above that incorporate spatial relations all point to a greater need to incorporate research about how people generate scene descriptions. Basic questions about how people receive and communicate knowledge about the physical world around them is foundational to the field of geographic information science. An important benchmark in the discipline's evolution was the development of Egenhofer and Mark's theoretical framework of Naive Geography (1995). Based on Naïve Physics (Hayes, 1978), the principles of Naive Geography have been used in geographic information science to model knowledge from a *common-sense perspective*. *Common sense spatial knowledge* is defined as "knowledge about the physical environment that is acquired and used, generally without concentrated effort, to find and follow routes from one place to another, and to store and use the relative position of places." (Kuipers, 1978, p.129). Naive Geography is defined simply as "the body of knowledge people have about the surrounding geographic world." (Egenhofer, Mark, 1995 p.6). Naive Geography takes into account the fact that people perceive, reason, and communicate about space and time in both conscious and unconscious ways. This may include reasoning that is based on high levels of uncertainty (i.e., incomplete information, biases, and errors) and that these factors must be accounted for in computational applications (i.e. GIS) to support human spatial cognition and spatial tasks. Finally, Naive Geography asserts that people often conceptualize and communicate about space using multiple perspectives, shifting levels of spatial detail and perceptions of spatial boundaries are context dependent.

Naive Geography provided a set of theories to guide the emerging field of geographic information science helping to create applications that could reason on space and time in ways that would help humans navigate and investigate changes in the physical world. The set of theories became the basis for the development of formalisms of space and time for current intelligent spatial systems and evaluated the effectiveness of system performance against human conceptualizations with empirical human subjects testing. This ‘human to machine to human’ feedback loop is critical for understanding space from the human user perspective, and provides a rationale for this dissertation using human-subjects to better understand how people conceptualize and communicate object relations in indoor scenes.

In the present study, Naive Geography principles drive the examination of models of indoor space and how spatial relationships of objects are communicated through natural-language spatial expressions. Although Naive Geography was originally situated in large scale, geographic space, this work investigates how the same principles may inform understanding about human conceptualization, representation and communication within smaller scale spaces.

2.4. Distinct Properties of Indoor Space

When thinking about differences in scale of indoor and outdoor space, even a very large building, such as an airport terminal or a mall, is considerably smaller than the outdoor environment around it. Indoor environments limit observers’ field of view, line of sight, and add movement constraints that are not typically present or differ from those in outdoor settings, due to the built environment’s physical structure such as walls, doors, and ceilings (Richter, Winter, Santosa, 2011). Outdoor space is typically represented in symbolic 2D spaces, while, indoor environments are often represented as 3D multi-level models (Figure 2.2; Winter, 2012). Vertical features such as staircases, elevators, and ramps can interfere with cognitive map development and accurate orientation when navigating (Li, Giudice, 2012). Indoor spaces such as buildings are typically (but not always) organized in regular, and predictable patterns, where the connectivity of rooms is often considered more important than metrics of direction, angles, or distances (Giudice, Walton, Worboys, 2010). In outdoor space, people use geographic features such as the sun, geographic features (e.g., mountains, water bodies) as global landmarks, as well as local landmarks consisting of natural or man-made features (e.g., large trees, cell towers) for orienting themselves and locating objects within the environment. Many indoor environments do not usually have the same level of visual access to global landmarks and thus rely more heavily on local landmarks for the same spatial tasks.

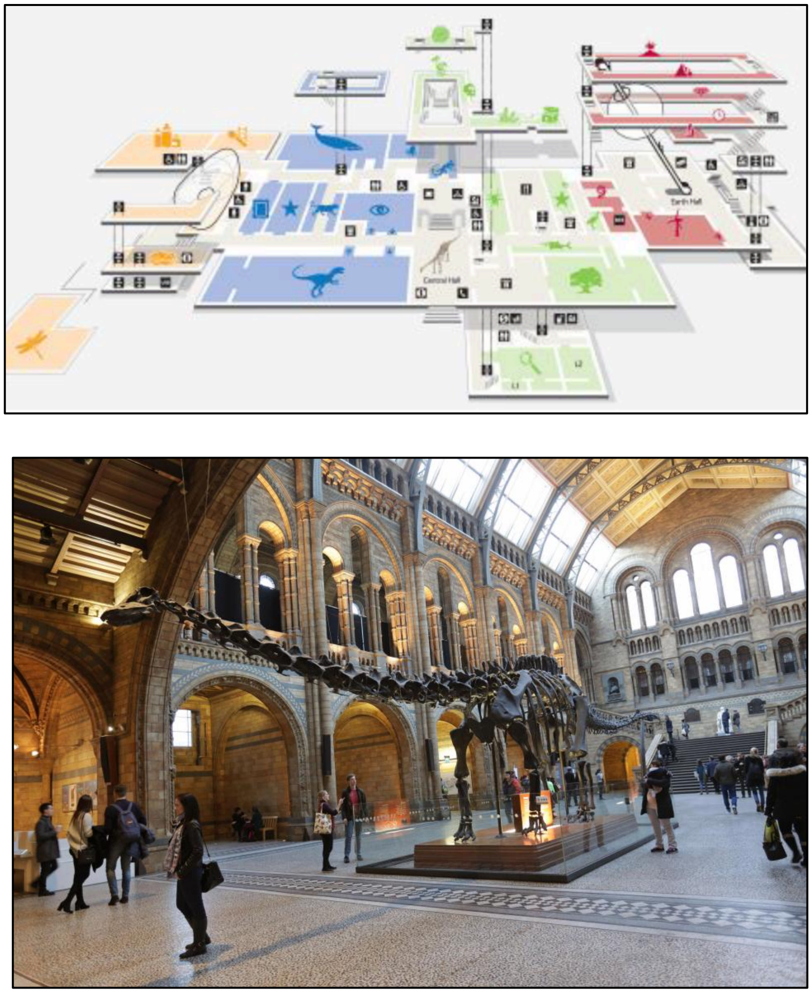


Figure 2.2: *Architectural details and objects.as landmarks*

2.4.1. Scales of Space

Many of the early models of space (Ittleson, 1973; Downs, Stea, 1977; Kuipers, 1978) broadly defined the characteristics of different spatial scales. However, as the field of geographic information science evolved, researchers created new classifications of space that explicitly represented smaller scales including indoor space, thus allowing for a greater level of spatial scale granularity (Zubin, 1989; Montello, 1993; Freundschuh, Egenhofer, 1997).

Zubin (1989) presents a model of space based on scales that people encounter in the real world. It identifies four types of spaces. Type A spaces, often referred to as tabletop space, are those spaces that include objects small enough to manipulate, are less than or equal to the size of the human body, and are contained in a static field. Type B spaces are characterized by objects which are larger than the human body and are typically not moveable, and are able to be perceived from a single perspective. Type C spaces (e.g., scenes) are constructed in components or objects that can be perceived by sensory scanning. Finally, Type D spaces are also constructed because they can not be perceived as a unit, as there is no single perspective.

Classifying aspects of indoor space according to Zubin's model would require the specification of user purpose, as the model could focus on small tabletop objects (Type A), an elevator or a set of bookshelves (Type B), a small room with furniture or the center court of a mall (Type C). Zubin spaces are vague with respect to the characterization of indoor spaces with the category Type C, being most closely matched as a model for indoor scenes due to the necessity of perceiving a scene as a set of objects.

Montello (1993) classifies space based on functional properties and projective size, rather than absolute size. In his classification, *figural* space is defined as smaller than the human body, able to be perceived without motion and with subclasses of pictorial space (small, flat 2D) and object space (small 3D). *Vista space* is defined as larger than the human body and able to be perceived from a single perspective without the need for movement to conceptualize the space (Montello, 1993). Vista-scale space includes a variety of size settings from a single indoor room, to a town square, and up to an entire horizon. Moving into larger spaces, *environmental* space is defined as larger than the human body, and requires motion and time to be able to directly perceive it. This includes indoor spaces such as entire buildings as well as outdoor spaces such as cities. Finally, *geographic* space is defined as much larger than the human body. It is a space that cannot be perceived through time and motion effectively because of its extent, and can only be perceived through symbolic models (i.e., maps). The typical indoor room scale space falls into Montello's' vista space category as it can be perceived from a single location without motion.

Freundschuh and Egenhofer's (1997) framework for experiential categorization of environmental space covers a large indoor room in a similar manner as both Zubin's C space and Montello's vista space. However, the framework classifies spaces based on the ability to manipulate objects, the amount of locomotion required to directly observe the space, and the size of the space. Due to these distinctions, the framework breaks down what might be an indoor space with larger objects into two categories-environmental space (the indoor room) and non-manipulatable space (the larger objects within the indoor room). The review of spatial scale classifications conducted by

Freundschuh and Egenhofer (1997) is helpful in determining the overlap of properties of each model. It also helps to identify a gap in the research on locative understanding and natural-language communication of spatial information at different scales that specifically focuses exclusively on the indoor environment. This identified gap, combined with additional evidence of differences in cognitive representation of spatial properties at different spatial scales (Franklin, Tversky, 1990; Montello, 1993; Tversky, 1981; Freundschuh, 1992) provides the rationale for this thesis.

For purpose of this dissertation, the focus will be on the range of objects and structures characterized by Montello's vista-scale space because, for most people, perception of this scale of a spatial scene depends almost completely on vision and small head and eye movement (Montello, Raubal, 2012), The scene description for this spatial scale should be able to convey a minimum amount of information about the following spatial properties: object configuration, connections, containments, as well as estimated distance and directional information. The open descriptions of indoor scenes (Chapter 3) and the structured spatial expression prompts (Chapter 4 and 5) all convey these basic spatial properties as they apply to a single indoor room that can be perceived from a single location without motion. All observations collected within the real-world and virtual scenes occur with the human subject situated within the room itself. Subjects are given instructions to (1) provide a description of the indoor scene without moving, and (2) to only describe what they can directly perceive from their single viewpoint.

2.4.2. Perceptions of Indoor Space

Behavioral and computational studies suggest there are differences in the visual and semantic information perceived when viewing indoor and outdoor scenes (Vailaya, Figueiredo, Jain, Zhang, 1998; Olivia, Schyns, 2000; Olivia, Torralba, 2006; Greene et al., 2016). Neuroscience studies have confirmed there are differences in the functioning of the posterior posthippocampal area of the brain when these sub-categories of real-world scenes are viewed by subjects while inside a functional MRI (Henderson, Larson, Zhu, 2007; Henderson, Zhu, Larson, 2011). The transition between indoor and outdoor spaces has been shown to cause confusion in orientation and wayfinding, suggesting different perceptions of these spaces (Kray et al. 2013). Cardinal directions are relied on heavily in outdoor settings, however, these systems are not typically used in indoor settings, where body referenced frameworks are favored (Tversky, 1993; 2009).

More recent theories of indoor space build on Gibson's (1976) *affordances* principle (Greeno, 1994; Norman, 2002; Giudice, Walton, Worboys, 2010; Yang, Worboys, 2011). In this approach, *affordances* refer to interaction possibilities that are perceived by an actor, depending on both the capabilities and the experiences of the actor. Indoor and outdoor spaces share many of the same affordance types including *passage, container, portal*, and *barrier*. For example, road networks are a common *passage* affordance type and building hallways can function in the same way within built environments. There are also unanticipated *barriers* within road networks (e.g., traffic and accidents) and hallways (e.g., locked doors). However, *containers* (e.g., rooms) and *portals* (e.g., elevators, stairways, windows, and lobbies) within indoor environments often serve as multidimensional affordance opportunities and these affordance types are not typically available in outdoor spaces. While the affordance type, *container*, is often used to represent a room in relation to the resources located within it, this dissertation moves away from this conceptualization.

Instead of a scene description represented as if the observer is describing what is contained in the room from the outside of the room, the approach adopted in this dissertation intentionally situates the observer directly (or virtually) inside the room, at the entrance, describing the room from an embedded perspective. The choice of this frame of reference is based on the NL scene descriptions collected and analyzed from a previous study (Kesavan, Giudice, 2012). It also follows a logic that in a real-world context, as was described in the hotel scenario (Chapter 1), the automated description of an indoor scene will have the most utility when the agent/user is embedded in the actual space, and the description is communicated from a known vantage point. In this way, it allows for a mapping of the linguistic information onto the physical space in which the agent/user is situated. This helps the description system to locate the user not only in the real-world space, but also in the cognitive map they are developing. This perspective can help to support subsequent spatial behaviors, and reduce reference frame misalignment, which may happen if the description is presented as if the agent is located outside of the room or indoor scene.

This dissertation research specifically investigates the ways in which indoor space can be represented as distinct objects (e.g., walls, windows, doors) that operate as local landmarks within indoor settings. These landmarks are used to create predictable patterns of object relations and spatial terms to form a standardized template for the description of indoor scenes. In order to do this, the conceptualization of the indoor space must move beyond thinking of a room as only a container of objects, and instead to representing the room as a collection of relationships that exist between moveable objects and/or structural objects.

2.5. Formal Representations of Indoor Space

Substantial research has been undertaken on qualitative spatial relations that can apply in any scale of space. Qualitative spatial models define relations based on specific characteristics of space, including topology (Cohn et al., 1997; Renz, 2002; Egenhofer, Franzosa, 1991, 1995; Egenhofer, Vasardani, 2007), direction (Frank, 1996), size and distance (Pacheco, Escrig, Toledo, 2002), shape (Museros, Escrig, 2004), orientation (Freska, 1992; Moratz, 2006) and motion (Galton, 2012). These formal relations are based on abstract mathematical concepts rather than human NL use patterns (Hois, 2010).

Topological relations are often considered the most fundamental way to describe object locations in space. Topological models, such as the 9-Intersection (Egenhofer, Herring, 1990) define primitive relations that hold between points, lines, and regions. For two simple regions without holes embedded in R^2, the 9-Intersection (Egenhofer, Herring, 1990) distinguishes eight topological relations based on how the regions' interiors, exteriors, and boundaries relate to one another. This type of formalization has been primarily directed to 2D views of a geographic scale space. Different subsets of relations may be needed to represent physical relations between 3D space filling objects (e.g., furniture) and the objects that form the structure of the room (e.g. such objects cannot physically overlap). Figure 2.3 illustrates 2D and 3D views of 9 intersection relations with room as an abstract container.

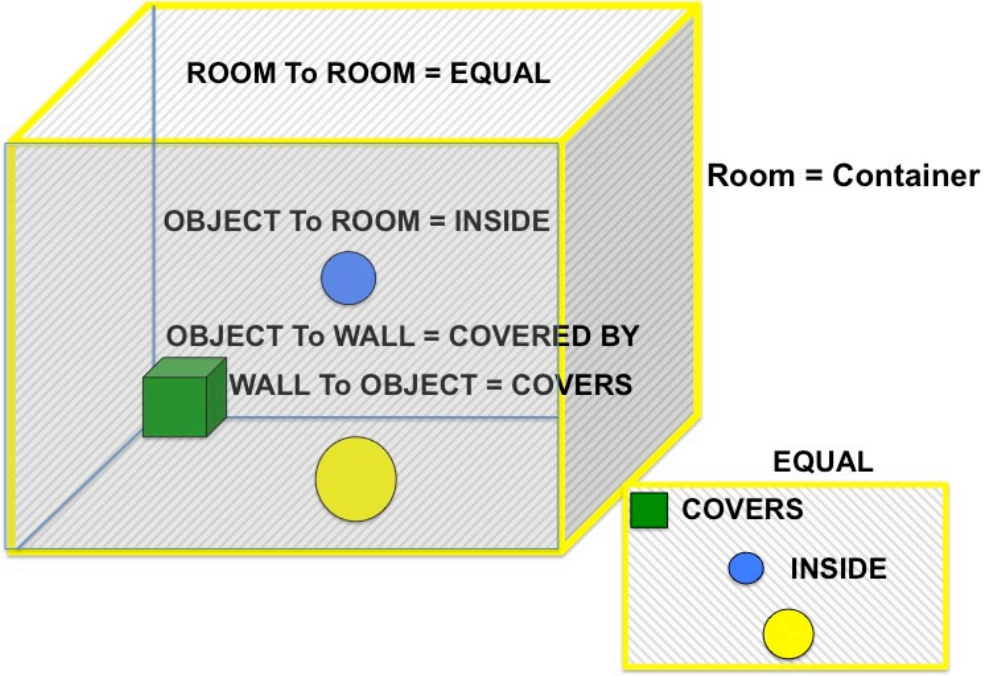


Figure 2.3. *Room and other objects represented in solid 3D and 2D container views*

Other approaches to containment that might be applied to the representation of rooms include using *container* schemata (Lakoff, 1987; Kuhn, 2007; Walton, Worboys, 2009), and formal ontologies (Grenon, Smith, 2004; Masolo et al., 2003; Hahmann, Brodaric, 2013). Hahmann and Brodaric (2013) note that qualitative spatial relations alone may not be the best approach for the conceptual representation of containment when it comes to 3D physical entities. The scenarios described in Chapter 1 require a finer level of information because of the need to describe spatial relations between objects contained in a 3D object (the room as the sum of room structures and the bounded void) that can either be in the void, or part of the room structures (Hahmann, Brodaric, 2013; Brodaric et al., 2017).

2.6. Relevant Aspects of Linguistic Models of Space

Spatial information is found in most classes of words and nearly all prepositions convey some level of spatial and/or temporal information. Yet, spatial concepts expressed in prepositions are often imprecise and non-metric, describing more qualitative than quantitative information about distances and directions (e.g., *near, far, right, left*). Likewise, most spatial terms are dependent on various aspects of context for their interpretation (Montello, 2009). In some cases, spatial prepositions can be characterized strictly as an expression of "spatial configuration", while in other cases, these terms might more accurately be described as a way to express "functional interaction" (Langacker, 2010).

2.6.1. Reference Frames

The ways people communicate about space provides important clues about how the typically functioning brain processes multiple channels of sensory input to create a conceptual model of space (Miller, Johnson-Laird, 1976; Tversky, 1993, 2001, 2009). When people are asked to describe scenes, the amount of precision and the reference frame used in spatial language is just as important as the types of spatial objects employed as landmarks. Some languages, such as English, use egocentric terms to describe spatial locations and relations (e.g., the cup to the right of the pitcher), while other languages, such as Tseltal Mayan, use an allocentric perspective (e.g., the cup to the downhill of the pitcher; Mark, Frank, 1992; Levinson, 2003; Abarbanell and Li, 2015). For the purposes of this dissertation, Levinson's (2003) definitions and distinctions are used to distinguish between three spatial reference frames: (1) *absolute, (2) relative,* and (3) *intrinsic* (Figure 2.4).

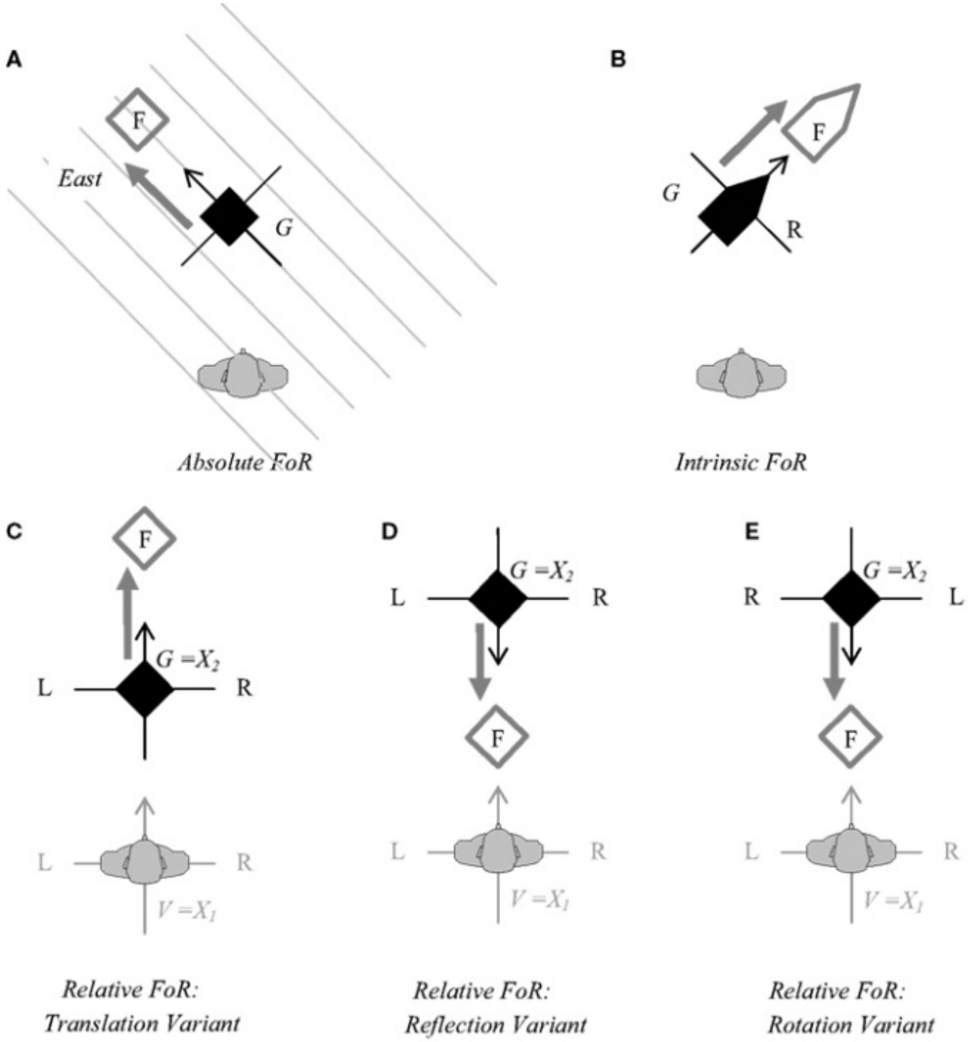


Figure 2.4: *Frames of reference* (Levinson, 2003; Bender, Beller, 2014).

An *intrinsic frame of reference* is an object-centered coordinate system, where the coordinates are determined by inherent features, such as sidedness or facets of the object to be used as the relatum. A spatial expression that illustrates an intrinsic frame of reference would be, “There is a chair in front of the desk.”, where the location of the chair is defined in relation to a part of another object, in this case, the front of the desk. An *absolute frame of reference* refers to the use of a system of coordinates anchored to fixed points and an origin at ground. An expression illustrating an absolute frame of reference would be “The chair is to the north of the desk.”, where a cardinal direction system or degree system might be imposed that is independent of the position of the agent/perceiver or any part of the objects. This reference frame is often used in linguistic descriptions of outdoor scenes but is less frequently observed in descriptions of indoor scenes. The *relative frame of reference,* is viewer-centered. This perspective is expressed through a triangulation of three points from a single viewpoint. The coordinate system is based on imaginary horizontal and vertical planes through the human body (*up/down, back/front, left /right*; Herskovits, 1986). A spatial expression using a relative reference frame would be, “The chair is to the left of the desk.”. In this expression, there are three reference points communicated: the chair, the desk and the agent/perceiver. This dissertation includes an analysis of reference frames used in the descriptions of scenes in order to better understand preferred use patterns of reference frames as they relate to spatial prepositions used in the scene descriptions.

2.6.2. Spatial Prepositions

This dissertation focuses on the use of spatial prepositions to convey specific information about relations between objects in indoor space. Spatial preposition acquisition happens early in language development as most children learn to speak anywhere between the ages of one year to three years (Clark, 1973; Miller, Johnson-Laird, 1976). *In* is most often the first spatial preposition adopted and used as an overgeneralized spatial expression, replaced by more specific locative prepositions *on* and *at* by ages three to five years (Freundschuh, Sharma, 1995; Ursini, Akagi, 2013). Spatial prepositions are often some of the most difficult language structures to use correctly for learners of second languages (Bowerman, 1996; Coventry, Garrod, 2005).

A spatial preposition is defined as a term that specifies a relation between a noun or pronoun and another word in the sentence or a noun phrase (prepositional phrase). There are only between 80 and 100 prepositions in the English natural language and far fewer prepositions that explicitly express NL spatial relations (Landau, Jackendoff, 1993). From a linguistic perspective, Coventry and Garrod (2005) classify spatial prepositions broadly by use and meaning (Figure 2.5).

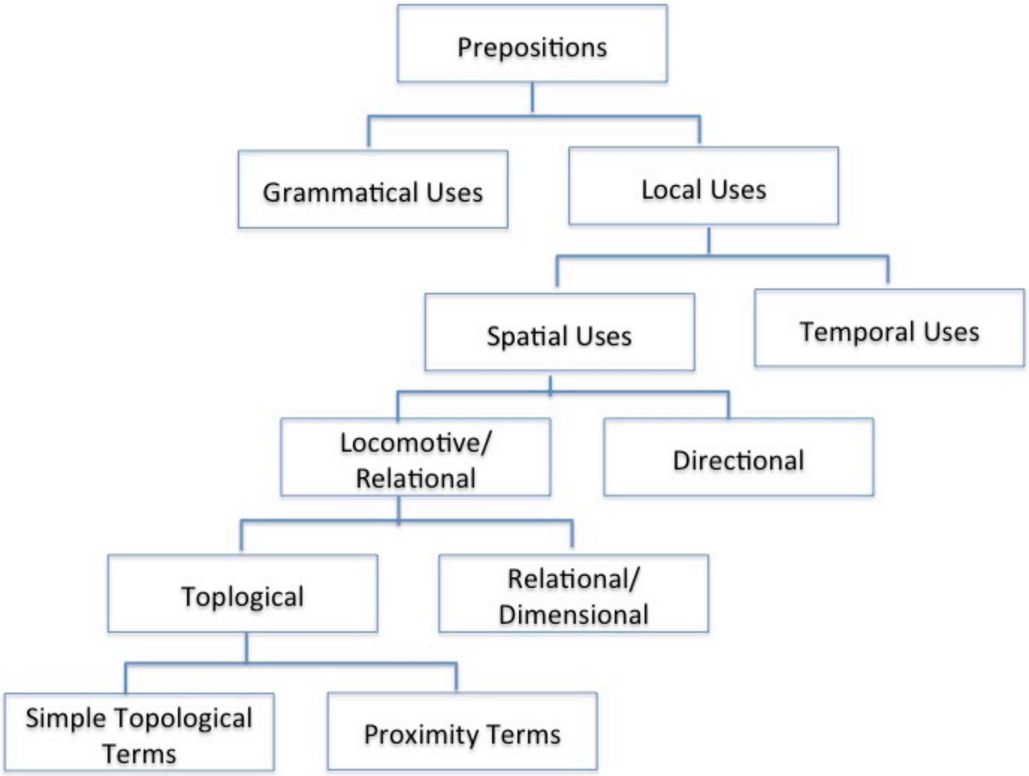


Figure 2.5. *Preposition Classification* (Coventry, Garrod, 2005).

Early work on the semantics of spatial prepositions focused on mapping geometric relations onto lexical entries for spatial prepositions and spatial concepts (Bennett, 1975; Coventry, Carmichael, Garrod, 1994). Herskovits (1980; 1986) outlined a set of object characteristics and contextual factors that impact spatial preposition use and interpretation. These principles revolved around object characteristics, such as shape, function, geometric context, and potential for mobility of objects. The potential for mobility of the reference object (ground) in relation to the located object (figure) impacts the order and use of prepositions, with the more mobile object typically preceding the preposition (e.g., bicycle against the tree) (Talmy, 1978).

Contextual factors of spatial-language use are often interdependent. These factors include the location of the observer as well as an often, imprecise distance threshold, indicating near proximity of the *figure* to the *ground* (Herskovits,1980). Spatial language differences also reveal how a particular object is viewed for a specific purpose, with viewers often ignoring specific characteristics of the object. Herskovits provides an example of a road, which may be communicated as a surface or a line (e.g., *a truck on the road* versus *a town on the road to Bangor*) depending on the viewer's spatial language or the distinction of a path that crosses an object's boundaries (e.g., *walking through town* vs. *walking across town*; Talmy, 1978).

The principle of salience also comes into play when there is an intervening relation between the *figure* and the *ground* (e.g., *The chair is in the room, on the rug.*), distinguishing between a *contain* relation (room) and the *contact* relation (chair; Herskovits, 1980). Some of the additional factors that influence spatial preposition choice and convey contextual spatial information (Feist, 2000) include:(a) *contact* between the figure and ground; (b) use of a *vertical axis;* (c) *inclusion* of the figure by the ground; (d) *support* of the figure by the ground; (e) the *nature of the support*, if any, afforded the figure by the ground; and (f) the *functional relation* between figure and ground. All of these principles can be observed in the patterns of preposition use found in the analysis of scene descriptions (Chapter 3) and provide a rationale for the detailed examination of the relations and prepositions to identify patterns for the construction of concise and accurate automated descriptions of indoor scenes. The hypothesis of this dissertation revolves around the argument that indoor vista-scale space may introduce use patterns for spatial prepositions that are not typically observed and communicated at the other spatial scales.

2.6.3. Spatial Prepositions at Different Spatial Scales

Difference in spatial preposition use has been found across spatial scales. At the figural scale, comprehension of spatial relations and perceptions of relative distance (e.g. nearness and farness) depend on the size of the spatial scale, as well as the presence of distractor objects in between object pairs (Burgio, Coventry, 2010). Likewise, spatial prepositions indicating a flexible ‘boundary’ where something was *near* was found to be heavily dependent on the scale and the context of the scene (Hall, Smart, Jones, 2011). Freundschuh and Blades (2013) found differences in spatial preposition use with a tabletop scale model and a model representation of a large geographic scale. Humans also often combine geometric cues with featural cues (i.e., landmarks) through spatial preposition use (Wang, Spelke, 2002; Wolbers,Wiener, 2014). This research provides additional evidence that different types of prepositions are used in different scale spaces.

2.6.4. Characterization of Spatial Expressions

Traditionally, spatial expressions are classified by concepts of spatial-configuration such as *figure* and *ground* (Talmy 1978) or *locatum* and *relatum* (Bateman et al., 2010). In this dissertation, Langacker’s (2010) conceptual characterizations are used which identify three major functional entities. First, is the *trajector* which functions as the *target* or the entity one might be trying to locate (e.g., box, lamp, and room). Second is the *landmark* which functions as the *reference point* or the entity one uses to find another object (e.g., chair, bookcase, and stairs). Finally, there is *the search domain* or the limited region within which the target can be found (e.g., front, side, and top). This framework has the advantage of conveying more information about interrelated context dependencies, anticipatory motion, and functional properties of the objects than the more commonly used configuration terms. This additional level of information becomes important when developing annotation schema, conceptual models, and spatial ontologies.

2.6.5. Ontologies of Spatial Language

Ontologies have become widely used in the development of information systems. An ontology is typically defined as "an explicit specification of a conceptualization" (Gruber, 1992, p.199) or "a logical theory accounting for the intended meaning of a formal vocabulary" (Guarino, 1998 p.8). An ontology of spatial relations and objects helps describe spatial utterances at a more conceptual level. This dissertation uses the Generalized Upper Model (GUM; Bateman, Henschel, Rinaldi, 1995) and its spatial component, GUM-space, to annotate scene descriptions because it combines both the cognitive and linguistic representations of spatial concepts. GUM provides general task and grammatical semantics for natural language processing. As a linguistically motivated ontology, it specifies semantics expressed in grammatical units (e.g., clauses, nominal groups, phrases) and the semantics of word functioning in a grammatical context. GUM is split into two hierarchies: (1) concepts (top entity: thing) and (2) roles (top entity: relation;).

The spatial extension, GUM-Space (Bateman et al., 2007), formalizes categories that are relevant for the natural language of space (Bateman et al., 2010; Hois, Kutz, Bateman, 2008). As the primary aim of GUM-Space is to provide a basis for the representation of spatial language for NL dialogue systems, it is an appropriate model to use as an annotation schema in the current research. It provides the linguistic components necessary to formally specify spatial language utterances for use within NL dialogue assistants in relation to the formal representations of *spatial scenes* (Tyler, Evans, 2003; Bateman et al., 2007; Hois, Kutz, & Bateman, 2008). GUM-Space provides approximately 70 different types of spatial relations (e.g., SpatialModality) that define how entities can be located in space

GUM-Space has been evaluated for its inter-annotator reliability and its spatial logics using a number of spatial-language corpora (Hois, 2010; Hois, Kutz, 2008; Elahi et al., 2012). These evaluations include testing GUM-Space performance using different spatial corpora such as the Trains 93 Dialogue, the HCRC Map Task, and the CReST corpus (Heeman, Allen, 1995; Anderson et al., 1991; Eberhard et al., 2010).

These spatial corpora are an important component in developing better formal structures because they provide away to test the quality of an ontology in its translation and generation of the inherent uncertainty and inconsistencies of natural-language spatial expressions. Each spatial language corpora focuses on a distinct aspect of spatial behavior and the language associated with that spatial task. For example, the Trains 93 Dialogue corpus (Heeman, Allen, 1995) describes train locations in outdoor environmental space and the spatial prepositions are purposely limited to include only 4 possible relations (*in, to, from,* and *with* (ex: We get a boxcar *from* Avon *to* Bath).

The HCRC Map Task corpus (Anderson et al, 1991) consists of a 128 spatial task-oriented dialogues between two participants with slightly different maps that represent the spatial configuration of approximately 15 landmarks in a fictional outdoor geographic scale space. The spatial task is centered on one participant describing a route printed on her map to the other participant so they can replicate the route based on its description. The Indiana Cooperative Remote Search Task (CReST) corpus (Eberhard et al, 2010) is similar to the MapTask corpus in that it consists of a set of natural-language dialogues of pairs of participants performing a cooperative spatial task. However, it specifically focuses on object search and locating in a variety of timed scenarios (e.g., search and rescue missions in disaster areas). It also consists of discourse between one participant with a map providing instructions to a partner participant about how to interact with physical objects while she is moving through an indoor environment. All of three of these spatial language corpora provide some detail about how people communicate about space, what spatial prepositions they use, and what language structures are common to a variety of spatial scales. However, none of the corpora focus solely on the description of the spatial configuration of objects and structures of simple indoor scenes.

Barclay and Galton (2008) provide a set of requirements for the development of a scene corpus for training and testing grounded spatial communication systems. Similar to a text corpus used for training and testing natural-language processing systems, this type of scene corpus should represent a range of spatial relationships over a variety of spatial scales. Unlike many of the text corpora described above, a scene corpus should ideally move beyond a focus on a single spatial task or element of the problem associated with generating spatial language. The minimal recommended aspects of spatial language built into this proposed scene corpus include: 1) the selection of appropriate reference objects(i.e. trajector and landmark), 2) the selection of appropriate frame of reference, and 3) the selection of appropriate spatial prepositions. If the system was intended to support multimodal forms of communication of spatial information, additional features could be incorporated including the capacity for non-verbal communication (e.g., gestures, intonation, emphasis), listener models that provide information on the presence and location of the listener, as well as multi-phrase and sequential route directions (Barclay, Galton, 2008). This new type of scene corpus should incorporate both traditional 2-dimensional images as well as 3-D images and dynamic scenes (e,g., animations and video clips) to allow for the appropriate mapping of spatial prepositions indicating motion. They also recommend the scene corpus include scenes that range from tabletop through geographic scale space with both indoor and outdoor settings. The size of the corpus that might represent a full range of scale spaces would need to contain at least 1000 scenes to represent the majority of English spatial prepositions and 4 reference frames. This type of scene corpus would have distinct advantages over much larger image captioning datasets currently used for automated image analysis and captioning training, such as UIUC Pascal Sentence dataset (Farhadi et al, 2010) or the Microsoft COCO captions set (Chen et al, 2015). A spatial scene corpus would allow for the incorporation of both the visual information represented in the spatial scene as well as the spatial language structure information that is necessary for testing both the spatial cognition and spatial linguistic aspects of scene descriptions.

The requirements for a specifically designed spatial scene corpus becomes important when testing natural-language motivated ontologies such as GUM-space. For instance, while GUM-Space was found to be adequate for structuring spatial language so that non-experts were able to understand and use the complex annotation schema, there was some confusion when evaluators were faced with similar, but slightly different, annotations. This confusion was particularly apparent when categories were specified hierarchically close together, but needed to be considered in context. This ambiguity in the semantic structure of GUM-Space is problematic for representing indoor environments which often have a high level of contextual uncertainty in the natural-language descriptions of complex indoor scenes. This detailed level of testing of spatial images and natural-language expressions would not be possible with existing large scale image caption datasets because of the lack of control over the specificity of the test data. The research conducted in this thesis aims to clarify this linguistic uncertainty by supplying a preliminary framework for improving specification of scene descriptions using GUM-space annotations for indoor vista scale settings as well as providing a pilot scene description corpus that specifically focuses on spatial information structures found in indoor vista-scale spaces.

2.7. Related Work on Spatial Descriptions of Indoor Scenes

The goal of this programmatic line of dissertation research is to identify patterns of NL spatial expressions that can be used in indoor vista-scale space to provide appropriate

NL descriptions for indoor scenes. There is a large body of work in spatial information science regarding the alignment of NL spatial relations with formal conceptual models in table top and geographic space (Mark, Egenhofer, 1994; Shariff, Egenhofer, Mark, 1998; Schwering, 2007; Klippel, 2012). Linguistic studies have looked at the problem of spatial preposition use for *on* and *in* from a 2D picture perspective (Feist, 2000; Feist, Gentner, 2003; Levinson, 2003). In these studies, the researchers limited the images depicting the conceptual continuum between *support* and *contain* to simple drawings, and did not include images depicting real-world indoor settings for these spatial relations.

Another body of related work uses NL descriptions of space to generate automated scene depictions based on spatial property graphs. Spatial property graphs provide basic spatial information in the form of spatial triples (*trajector, landmark, relation*) that are extracted and parsed from scene descriptions to form spatial networks. The use of spatial property graphs to depict spatial scenes is based on a set of assumptions grounded in the theory of the conceptualization of spatial scenes (Langacker, 1987;1993; Tversky, 1993; Tyler, Evans, 2003; Klippel, 2012; Giudice, Betty, Loomis, 2011; Vardesani et al., 2013). While this related work provides guidance on the methods to be adopted in the current experiments, there are some key differences in these previous studies from the focus and approach adopted in this dissertation. This dissertation research situates itself firmly in a small room setting, in vista-space, rather than figural or environmental space. This is important because although there is a substantial increase in interest in and technology to support indoor information retrieval, there have been traditionally fewer human-subject studies conducted solely at this spatial scale. In addition, many of the related studies have used a 2D line-drawing perspective for test images, rather than real-world or virtual-world scenes when assessing spatial preposition use.

Finally, while the approach of using spatial property graphs to generate automated scene descriptions is a promising avenue for using computer vision to process and interpret the spatial configurations of objects in a scene, this work does not sufficiently address the preferred spatial terms to use for the relations between objects in the brief scene descriptions. Rather, the current work aims to provide guidance about a small set of preferred spatial prepositions that can be used for communicating relations between objects in indoor scenes. The experiments and analyses presented in this dissertation were conducted in controlled indoor environments to create an opportunity to expand the initial indoor scene corpus developed in this dissertation across other types and sizes of indoor environments.

2.8. Chapter Summary

This chapter reviewed the background and related work on the alignment of spatial prepositions and the spatial concepts necessary to support the automated generation of NL descriptions for indoor scenes. The review included foundational work in the fields of spatial information science, spatial cognition, and spatial linguistics in order to better characterize and understand the ways in which people conceptualize and communicate about space. A discussion of the function and use of spatial prepositions was provided in order to motivate the analysis methods used for the scene descriptions described in the next chapter.

CHAPTER 3

ANALYSIS OF INDOOR SCENE DESCRIPTIONS

Given an infinite variety of ways that people could describe a single indoor scene, are there any patterns in the objects and linguistic term choices that might help in creating a model description of an indoor scene? This chapter presents an analysis of a set of indoor scene descriptions collected from ten human-subjects. First, the analysis evaluates if these descriptions match the key characteristics of a ‘good’ scene description, such as complete, correct and concise NL phrasing of spatial relations (Gapp, 1994; Bernardi et al., 2016). To accomplish this evaluation, there is a specific focus on the use of spatial prepositions in the phrasing of spatial relations between objects. The results provide guidance as to the length and structure of a concise and complete description of an indoor scene that might be automatically generated by a scene description system. The analysis also provides guidance on the spatial prepositions and relations to be tested in subsequent experiments designed to control for contextual aspects of indoor scenes, in a way that cannot be accomplished in open scene descriptions. Each scene description was evaluated for: (1) linguistic patterns, (2) functional characteristics, and (3) network structures. The analysis addresses the following research objectives:

(1) Identification of all moveable and structure objects included in the description of indoor scenes.

(2) Identification of all conceptual relations and spatial prepositions used to connect objects in the description of indoor scenes.

(3) Identification of all functional characteristics or relations in the description of indoor scenes.

Findings from this analysis provide the basis for further investigation of user spatial language preferences and the impact of room context characteristics in the experiments presented in Chapter 4 and the results discussion in Chapter 5.

3.1. Scene Descriptions

Data for this analysis was provided by a set of experiments conducted by Kesavan and Giudice (2012). For the experiments, participants were asked to describe small office indoor spaces for someone who might have a visual impairment. Participants were given a specific task to describe what they saw from a standing position at the doorway for someone who could not see the scene themselves. The indoor spaces used in the experiments were arranged to represent an office space (approximately 10' by 12') and included the same types and number of objects arranged in different spatial configurations (Figure 3.1). Specifically, there were two bookshelves, two file cabinets, three chairs, three tables, and one trashcan for a total of eleven objects in each room. Participants were directed to describe the office space as clearly and accurately as possible, include the objects they thought were important and the spatial location of the objects in the room. They were also directed to provide a clear way to address the similar objects (i.e., tables) in a distinct manner but not to focus on the details of all of the objects (e.g., number of books or shelves in a bookcase). The participants were not allowed to move around the space, only to describe it from the door opening.

In the original study conducted by Kesavan and Giudice (2012), two oral scene descriptions were collected from each of twelve participants. One description took place in real-time with the participant standing at the edge of the door opening (Real-World Observation), while a second observation was collected by asking the participant to describe what they saw in a picture of a similar small indoor space (Photo Observation). The observations were recorded and originally analyzed for word frequency, spatial object relations and object frequency patterns but not using formal linguistic methods (Kesavan and Giudice, 2012)

Findings from the original analysis of these experiments suggested that there were no significant differences between spatial information acquired by direct human observations or camera-based observations or in re-creation accuracy based on descriptions generated from these two modes (Kesavan, 2013). In addition, the use of photographs resulted in equivalent performance in the ability to apprehend, remember, and use spatial information in comparison to direct observation of the scene. The findings provide support for the use of photographs or desktop images as an equivalent information source of input in future indoor navigation systems. There is some question about the validity of studies that use desktop simulations of different scale spaces to generalize about spatial learning and the formation of cognitive maps (Montello, 1993). However, Kesavan and Giudice's results (2012) suggest the spatial task performance of participants was not significantly different when physically immersed in the setting (real observations) as compared with performance when viewing an image of the setting (simulation observation). While validity concerns may in fact be valid for simulations of different scale environments, this may not impact spatial task performance when comparing vista scale observations and descriptions and simulated figural scale image descriptions. These experiments also pointed to a 'Round-About' [strategy] description order of the location of spatial objects as the preferred description order for assisting in the acquisition of knowledge about indoor scenes.

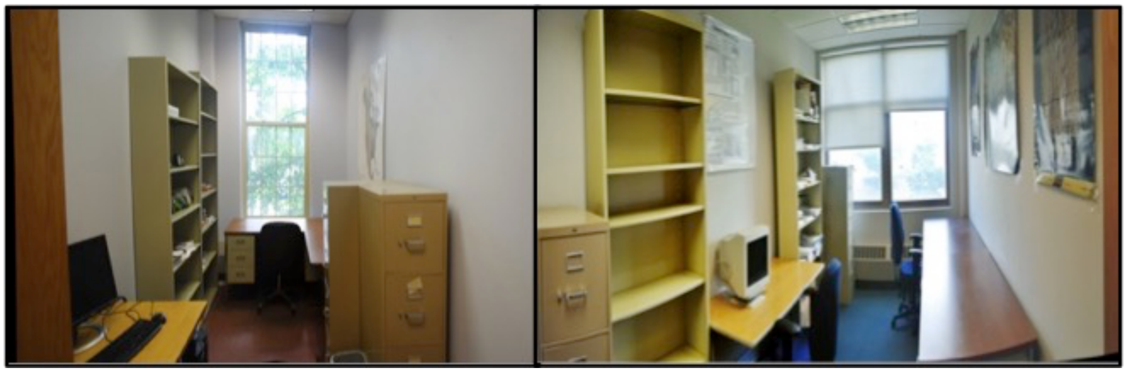


Figure 3.1. *Scene description environments: Room-1 and Room-2.*

3.1.1. Analysis Methodology

The re-analysis in this dissertation applies formal linguistic methods to construct a corpus from the indoor scene descriptions collected in the earlier study. The parsing of the descriptions into utterances, parts of speech tags, and applying a spatial annotation schema provides more details about how people describe indoor scenes based on spatial configuration (i.e., topological) and/or functional (i.e., use) characteristics. The real-world observations from the earlier study were re-transcribed verbatim to specifically include hesitations, false starts, corrections, word replacements, and utterances from each description based on participant intonation. Utterances are small, distinct units of speech with a clear beginning and ending separated by silence or a pause. Utterances make up the conceptual structure of a sentence, and there are typically two or more utterances linking a single spoken sentence together (see Chapter 2). All utterances were tokenized and the positions of tokens were indexed within each utterance using the Stanford Parts of Speech Tagger (Toutanova et al., 2003). Figure 3.2 illustrates an example of an utterance with the parts of speech tags.

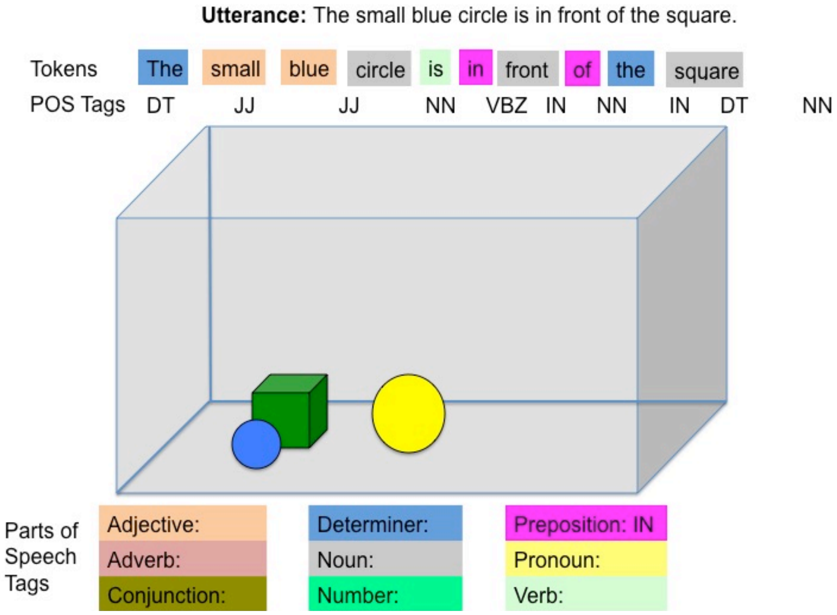


Figure 3.2. *Example utterance with tokens and parts of speech tags* (Stanford POS Tagger, Toutanova et al., 2003).

This tagging process allowed for formal linguistic analysis which included: (1) the raw count of utterances per subject/participant; (2) the average length of total utterances (including words and punctuation marks), (3) indexing and annotation of spatial role assignments of trajectory, landmark and corresponding spatial preposition; (4) spatial relation state (dynamic or static); and (5) GUM-Space annotation modality. The utterances were analyzed for frequency of parts of speech, spatial expressions, spatial roles, and characterization of spatial relations based on GUM-space definitions.

3.2. Linguistic Analysis Results

Once the scene descriptions were parsed into utterances, tokenized, and annotated and formatted into a corpus, descriptive statistics were calculated for utterances and parts of speech terms (Table 3.1). The descriptions were also evaluated for the following characteristics: (1) complete: all moveable objects and structure objects included in description, (2) correct: followed instructions and all moveable and structure objects were identified with an accurate descriptive term; and (3) concise: description did not contain information beyond what was requested in instructions. Overall, the descriptions met all of these basic criteria for a ‘good’ scene description, as specified by Gapp (1994) and Bernardi et al., (2016), although there were several that fell outside of the expected range for being either too long or too short in comparison to the others.

Table 3.1. Descriptive statistics for total scene descriptions

	Mean	**Median**	**Mode**	**Range**
Utterances	17	16	12,18	30
Tokens	400	438	--	238
Nouns	91	86	56,86	143
Prepositions	51	45	30, 45	77
Verbs	43	31	27	74
Adjectives	24	22	---	38
Adverbs	17	14	---	34

There was substantial variance in the number of utterances recorded for each participant, with a mean of 17 utterances per observation and 24 tokens per utterance. There was a mean of five nouns, two verbs and three prepositions used per utterance. The observed tokens and part of speech instances also reflect the wide range of utterance structure found in typical native English speaker’s natural-language descriptions. For example, Participant 9’s (S9) scene description contained the greatest number of nouns (subject/objects), verbs, adverbs and adjectives, while Participant 5’s (S5) scene description used slightly more prepositions (relations) than any of the other observations (Table 3.2).

Table 3.2. Tokens ordered by number of utterance with parts of speech counts

Subject	**Utterances**	**Tokens**	**Nouns**	**Prepositions**	**Verbs**	**Adjectives**	**Adverbs**
S9	37	843	183	95	96	50	42
S5	23	617	111	98	79	34	28
S3	18	469	120	53	42	31	18
S7	18	328	86	30	38	22	12
S8	16	438	106	68	27	29	14
S6	15	243	70	30	27	16	6
S2	12	387	86	45	44	20	26
S1	12	225	56	27	25	13	13
S11	11	250	56	45	31	10	2
S4	7	205	40	17	22	12	8
Total	**169**	**4005**	**914**	**508**	**413**	**237**	**169**
% tokens	______	**100%**	**23%**	**13%**	**10%**	**6%**	**4%**

Frequency counts of parts of speech may provide a preliminary clue as to object and relation focus across observations (Tables 3.3). For example, nouns referencing the indoor scene structural or boundary features such as *walls* (left, right, far) dominated the observations (n = 107). For moveable objects, *desk* was the most frequently referenced noun (n = 43). In terms of prepositions, *of* was the most frequently used preposition (e.g., '*left of the desk'*, or '*on top of the table'*), however*,* it typically functioned as a portion of a larger spatial prepositional phrase. The primitive spatial prepositions *on* (n = 70) and *in* (n = 54) were the most frequently referenced spatial relations in the observation data set.

These basic frequency counts actually point to some important observations about the indoor scene open descriptions. Kesavan and Giudice (2012) focused only on the configuration of the objects in the descriptions, not structural elements of the room (walls, windows, doors). However, the walls of the room were the most frequently used reference objects in the descriptions, more than double any other object referenced in the room (Table 3.3). Next, Kesavan and Giudice (2012) did not analyze the types of relations between the object configurations, only which objects were connected in the descriptions and in what order. Therefore, knowing what types of relation schemas are represented in the descriptions (*support, part of, contact, disjoint)* through the spatial prepositions used helps to better characterize the perception of the spatial scene by the observers (Table 3.3).

Table 3.3. Noun and preposition frequency in scene observations

Noun	**Instances**	**Preposition**	**Instances**
Wall	**107**	Of	86
Side	47	On	**70**
Room	45	In	54
Desk	43	From	27
Cabinet (file)	33	Out	24
Table	28	Against	24
Door	22	With	20
Bookshelf/case	22	By	19
Computer	16	Into	13
Window	13	Across	13

Based on GUM-Space annotation frequencies, each scene description contained approximately 40 spatial triples which consist of a trajector (TR), a spatial preposition (SP), and a landmark (LM). A spatial triple is defined in the descriptions as an "moveable object (TR) + spatial preposition (SP) + structure object (LM)". In many cases, there were multiple spatial triples used to describe a relationship that linked the primary trajector and the landmark feature pair within a single utterance. A spatial triple defined the spatial roles (i.e., TR+SP+LM) between four types of object pairs: an object-object pair (OO), an object-structure object pair (OS), a structure object-structure-object pair, or a structure-object pair (SO).

A sample of Room-1 and Room-2 observations (Tables 3.4 and 3.5) show the frequency patterns in the spatial triples and their corresponding annotations. For example, Subject 1 (S1) described Room-1 using a total of twelve distinct utterances. Those twelve utterances contained 28 spatial triples consisting of 28 trajectors (TR), 28 spatial prepositions (SP), and 26 landmarks (LM). All spatial triples were then classified using the GUM-space annotation category *general type* as a *regional* type relationship (e.g., There is a desk in the room, or The bookcase on the wall), a *distance* type relationship (e.g, The bookcase near the desk), or a *directional* type relationship (e,g., The chair to the left of the table). When examining patterns in the types of relationships in the spatial triples, more than half of all triples were categorized as a *region* type. This is important because it provides an overall characterization of the emphasis on the region type within scene descriptions as opposed to triples that reflected a distance or direction.

Table 3.4. GUM-Space concept annotations (Room-1)

		Spatial Roles				**General Type**		
	Utter	**Sp. Prep.**	**Trajector**	**Landmark**	**Sp. Triples**	**Region**	**Dist.**	**Direct.**
S1	**12**	28	28	26	28	10	7	11
S2	**12**	35	34	33	35	21	2	12
S3	**18**	45	45	45	47	17	9	21
S4	**7**	15	13	9	15	7	1	7
S5	**23**	55	55	51	59	36	8	15
S6	**15**	25	26	25	27	19	4	4
total	**87**	**290**	**288**	**276**	**298**	**110**	**32**	**70**

Table 3.5. GUM-Space concept annotations (Room-2)

		Spatial Roles			**Sp. Triples**	**General Type**		
	Utter.	**Sp. Prep**	**Trajector**	**Landmark**		**Region**	**Dist.**	**Direct.**
S7	**18**	31	36	31	38	24	5	9
S8	**16**	36	36	36	38	27	5	6
S9	**37**	90	90	89	94	56	11	23
S11	**11**	19	18	15	20	15	0	5
total	**72**	**176**	**180**	**171**	**190**	**122**	**21**	**43**

When investigating the patterns in spatial triples, a slightly different picture emerges related to the use of objects as either a trajector or as a landmark (Table 3.6). For example, when looking at frequency of token index position of moveable objects in the triple configuration, although *desk* was the most frequently referenced moveable object (noun), *filing cabinet, a smaller and vertically orientated object* was the object more frequently used in the trajector position, while *desk, a larger and horizontally oriented object,* was the most frequently used in the landmark position.

Table 3.6. Frequencies of moveable object annotation

Moveable Object	**Trajector Role**	**Landmark Role**
File cabinet	56	19
Desk	52	33
Table	39	19
Bookshelf	32	19
	175 (45%)	**80 (20%)**

In terms of structure or boundary spatial features, walls were infrequently used as a trajector but were the most commonly used landmark in the entire observation dataset (Table 3.7). This is consistent with similar studies in both indoor and outdoor spatial settings, where moveable objects (i.e., smaller objects) were more frequently used in the trajector position and immoveable objects that represented structural or boundary features were more frequently used in the landmark position of a spatial triple (Herskovits, 1980). Out of the 12 objects within the room, 7 of the moveable objects (*file cabinet, desk, table, bookshelf*) represented more than half of all trajectors in the observation dataset. Preferences to the room structure *wall* in the landmark position occurred in 37% of the total number of spatial triples. While the smaller objects, such as the chairs and the trashcan, were mentioned as secondary references within longer spatial expressions, the larger, moveable objects were featured in almost all trajector positions. The door of the room was rarely mentioned in any of the observations. When mentioned, it was referenced more frequently in the landmark position, suggesting it may be perceived more as a room structure than a moveable object.

Table 3.7: Frequencies of spatial structure feature annotations

Structure Objects Boundaries	**Trajector Role**	**Landmark Role**
Wall	16	133
Window	14	15
Door	9	18
Corner	3	5
	(10%)	**(42%)**

3.3. Functional Characteristic Analysis

The scene descriptions were also annotated with an additional set of semantic codes of characteristics of human interaction with the world at the physical, perceptual and purposive levels (i.e., *functional characteristics*). The annotation schema coded observations for physical access, the perspective type and then the way in which the observer described the accessible objects/structures. Perceptual access was coded based on if the observer only described what they were able to perceive as accessible based on the stated observation orientation or if they mentioned objects/structures that were behind or otherwise not immediately accessible (e.g., beyond the boundaries, behind the door when opened). Finally, the observations were also coded for the perceived observer direction, and a general vs. a lateral orientation (Tables 3.8 and 3.9).

The annotation for Subject-1 (S1) is interpreted as follows based on this annotation schema: Subject-1 described Room-1 scene's *physical access* from an intrinsic reference frame (I) choosing to describe objects and structures in the room in a *near, right, left* and *far order of potential encounter*. This means that they focused on the items physically nearest to them first, and then moved away from their position to the right of themselves and then to the left. The subject then ended the description by describing objects and structures furthest away from their position on the far wall of the room. The participant only described what they could see in front of them or to the immediate sides. They did not describe anything that might have been outside their field of vision (e.g., door or walls directly behind them).

Table 3.8. Functional characteristics of Room-1 observations

Subject	**Physical Access**	**Perceptual Access**	**Order of Potential Encounter**	**Perceived direction based on General or Lateral Orientation**
S1	I, N, R, L, F	Perceived access only	Near>Right> Left>Far	General
S2	I, F, L, R, N	Perceived access only	Far>Left>Right> Near	General
S3	I, F, L, R, N	Perceived access only	Far>Left>Right> Near	General
S4	Rel, L, R, F, R	Perceived access only	Near> Left> Right> Far	General
S5	I, N, R, F, L	Describes structure object 'behind' perceiver	Near> Right> Far> Left	General
S6	I, N, L, R, F	Perceived access only	Near> Left> Right> Far	General

I: Intrinsic, Rel: Relative R: Right, L: Left, F: Far, N: Near U: Under, B: Behind, A: Above

Table 3.9. Functional characteristics of Room-2 observations

Subject	**Physical Access**	**Perceptual Access**	**Order of Potential Encounter**	**Perceived direction based on General or Lateral Orientation**
S7	Rel, R, F, N, L	Perceived access only	Relative, Right, Far, Near Left	General
S8	Rel, R, F, L, N	Perceived access only	Relative, Right, Far, Left, Near	General
S9	Rel, U, B, R, A, N, L, F	Describes structure 'behind' and 'above' perceiver	Relative, Under, Behind, Right, Above, Near, Left, Far	General
S11	I, R, F, L, N	Perceived access only	Intrinsic, Right, Far, Left, Near	General

I: Intrinsic, Rel: Relative, R: Right, L: Left, F: Far, N: Near, U: Under, B: Behind, A: Above

Although there are only ten annotated observations, there are a few potential patterns that could be further explored in order to guide rule generation for a NL language scene description system. For example, in the observations collected for Room-1 (Table 3.8), most observers began their description from either an *intrinsic-near* or an *intrinsic-far, relative perspective*, meaning they framed the description in terms of “you” or, in one case, “me” and then referenced objects/structures nearest to or furthest away (i.e., directly in front of you… or… furthest away, as you walk in the door…). This may suggest, at least in this room configuration, the forward-oriented starting point was preferred over a lateral start point. However, in Room-2 observations (Table 3.9), most of the subjects began with a *relative perspective* that referenced a lateral-oriented starting point to the right of the observer (i.e., on the right wall…) rather than an intrinsic perspective (i.e., you are…).

Most of the observations began with a reference to the estimated dimensions of the observed rooms. It is unclear if this was a part of the protocol prompt from the study but it does provide some useful information about small indoor space estimation. Subjects who did include dimensions estimated the rooms as anywhere from 14-25 feet long to 6-12 feet wide. Only one observer used the term *paces* rather than an estimated metric, in feet, and only one observer included a vertical estimate of a 9-10 foot tall ceiling. Two out of the ten observations did not include any room dimension estimates. Although there is a wide range of estimates, if the room dimensions are averaged across observations with estimates in feet, the room was observed to be approximately 16’ long and 8’wide which was a very close approximation of the actual dimensions in both Room-1 and Room-2. This level of collective accuracy in estimating room dimensions points to the possible use of crowd sourcing of indoor space descriptions to achieve greater accuracy in scene depiction.

While previous analysis of the description sequence was classified as the 'round-about' description sequence (Kesavan, 2013), when linguistic cues are more closely examined and annotated, a sequential pattern emerges of *nearest to farthest* from (observer-self) across Room-1 observations, and a simple *counter clockwise* description from the observer-self as is evident across Room-2 observations (Table 3.8 and Figure 3.3).

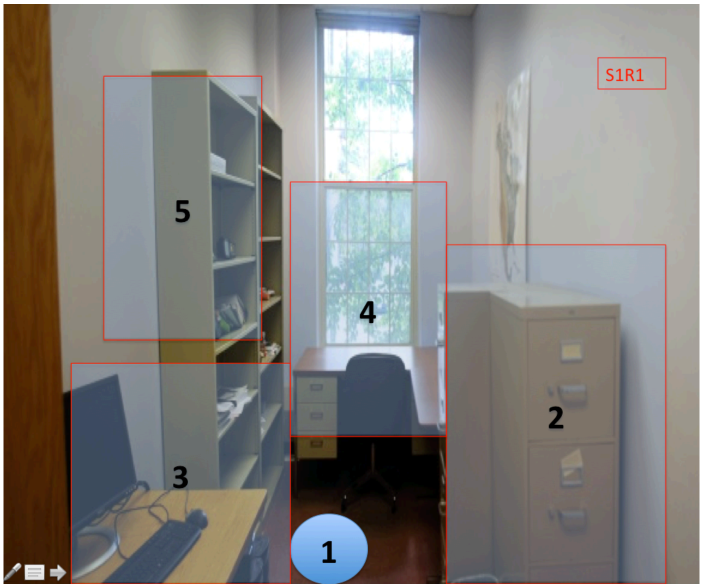


Figure 3.3. *Example of order of perceptual encounter* (Subject-1,Room-1).

Another approach to examine collective characteristics of the observations is through a network analysis of the observations dominant connectivity, object/structure centrality patterns and preferred object/structure/relation order.

3.4. Network Analysis of Scene Descriptions

Although the linguistic analysis is helpful in discovering patterns in frequency of reference of objects and structures within indoor scene descriptions and the prepositions used to describe the spatial relationships, it does not adequately capture the nature of the relationships between groups of objects and structures in the descriptions. It also does not provide much insight into how the description is spatially structured as a whole. A network analysis was conducted in order to look at topological structure of the observations more closely. Each observation was configured as a network and the dataset of networks was explored along multiple dimensions such as connectivity, scale, node association, and node-edge order. The structure and metrics of each observation network were analyzed separately by room configuration, individual nodes for ranking metrics, and identified cohesive clusters of nodes to look for patterns in object/structure groups.

The nodes and edges were based on the spatial triples extracted from the observation utterances. They were classified, analyzed, and visualized based on node clustering, the frequency and order in which the node-edge (spatial triple) was used, and the spatial prepositions that were used as relations between nodes within and across observations (Figures 3.4).

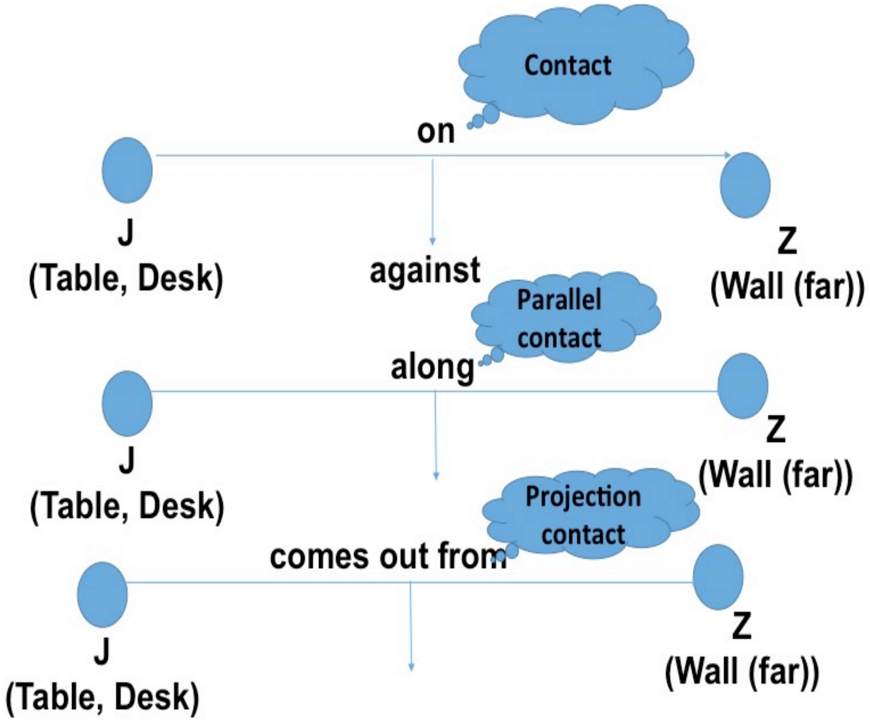


Figure 3.4. *Nodes and edges from a scene description network*

3.4.1 Overall Network Metrics

The network metrics (e.g., number of nodes, number of edges, and network density) for each observation provide insight into the variability of observations in terms of network characteristics (Table 3.10). The analysis was divided by room because the object configurations differed and comparisons between the different room networks would not yield useful information. While the results reported for the network analysis are not statistically significant because of the small number of nodes (objects) in each room, the analysis does provide some insight into the patterns across participant descriptions that serve as the basis for experiment scenes and prompts in Chapter 4.

The number of unique nodes in Room 1 observations ranged from a low of 19 to a high of 43 nodes (m= 29 unique nodes). This was similar for the unique edges with a low of 16 and high of 40 (m=22 unique edges). The networks' densities differed across observations (mean distance= 3.02; m density= 0.04). Subject-5's observation showed the greatest distance and density: a result of more description utterances creating more nodes and edges and a larger, more complex network.

Table 3.10: Room-1 network metrics

	S1	S2	S3	S4	S5	S6
Unique Nodes	23	29	34	19	43	26
Unique Edges	22	30	40	16	40	20
Duplicate Edges	5	6	8	0	19	7
Total Edges	27	36	48	16	59	27
Average Geodesic Distance:	2.52	3.83	3.5	2.95	3.86	1.485
Graph Density	0.049	0.04	0.04	0.046	0.026	0.035
Mean In-Degree	1.087	1.138	1.294	0.842	1.116	0.885
Mean Out-Degree	1.087	1.138	1.294	0.842	1.116	0.885
Mean Betweenness Centrality	22.435	83.241	85.882	24.947	91.674	3.615

Similarly, Room-2 networks had one larger, more descriptive observation (Subject 9) that has a significantly greater number of unique nodes and edges as well as a greater distance and smaller density structure (Table 3.11). The other networks in this observation set were similar in size and dimensions to the majority of Room 1 networks (m=32 unique nodes; m= 34 unique edges; m distance= 2.80; m density= 0.04). These patterns may provide some insight as to the ideal size of a simple scene description using a spatial network. It can also provide guidance about the minimum and maximum amount of spatial information that is necessary for effective indoor scene descriptions.

Table 3.11: Room-2 network metrics

	S7	S8	S9	S11
Vertices – Unique nodes	24	26	55	26
Unique Edges	29	27	66	17
Duplicate Edges	9	11	28	3
Total Edges	38	38	94	20
Maximum Geodesic Distance (Diameter)	5	7	8	4
Average Geodesic Distance:	2.93	3.26	3.5	1.49
Graph Density	0.060	0.047	0.025	0.027
Mean In-Degree	1.375	1.192	1.4	0.692
Mean Out-Degree	1.375	1.192	1.4	0.692
Mean Betweenness Centrality	47.417	59.846	109.345	3.154

The overall connectivity metrics (degree and centrality measures) for the networks do not provide particularly useful information to guide future scene description parameters but they do show the significant differences in the networks with highly connected/central nodes-edge structures (S3, S5 and S9) versus the networks with more isolated node-edge patterns (S6 and S11). Looking at the structures of connectivity and centrality for individual regions of the networks may provide more insight into scene description patterns that might be useful in creating and testing rules for an automated NL scene generator.

3.4.2. Individual Node Ranking

Nodes in the networks were analyzed for out-degree counts (e.g., object as Trajector [TR]) or in-degree counts, (object/structure as a landmark [LM]; Tables 3.12 and 3.13). Of the moveable objects in Room 1, the desk/table (far) had both the highest in-degree and out-degree as well as the highest connectivity (betweenness centrality). This means that this object plays a very important role in the spatial network across Room 1 observations. The observer (you) also served as a highly connected node in the collective network.

Table 3.12: Room-1 moveable objects node metric rankings

	Out-degree (TR)	In-degree (LM)	Betweenness Centrality
Table/Desk-Far	15	10	1313.44
You (Observer)	7	10	996.52
Table/Desk-right	12	9	1046.08
Filing Cabinets (plural)	10	7	821.07

For the structure objects in Room-1 observations, the reference to the bounded space (room) was the most highly ranked node in terms of degree measure and centrality measures (Table 3.13). This was followed by three of the walls (far, right, left) which made up the structure objects of the enclosed space. Although the left wall had the highest in-degree count, it was the far wall that had the highest level of connectivity among the three walls, suggesting that the far wall's role in the network is primary in terms of the description structure.

Table 3.13: Room 1 structure node rankings

	Out-degree (TR)	In-degree (LM)	Betweenness Centrality
Room (space)	3	9	481.57
Wall (left)	1	8	131.97
Wall (far)	1	6	182.36
Wall (right)	1	6	148.28

Of the moveable objects in Room-2, the far desk/table had both the highest in-degree and out-degree as well as the highest moveable object connectivity (betweenness centrality). This means that this object plays a very important role in the spatial network across Room-2 observations (Tables 3.14 and 3.15). The other larger objects that were separated in this room also played more of an important role in the network for Room-2. This difference is likely due to being perceived as separate objects to be accounted for in the description as opposed to being ‘chunked’, as a single object in Room-1 configuration. This perception of object grouping is important because it may provide insights into classification rules about similar adjacent objects in indoor environments.

Table 3.14: Room-2 moveable objects node metric rankings

	Out-degree (TR)	In-degree (LM)	Betweenness Centrality
Table/Desk-left	5	3	149.75
Bookcase (near)	4	2	39.24
File cabinet (far)	4	1	101.00
Tables (plural)	5	0	43.25
Bookcase (far)	3	1	62.41

Table 3.15: Room-2 structure node rankings

	Out-degree (TR)	In-degree (LM)	Betweenness Centrality
Room (space)	4	6	283.17
Wall (left)	1	8	221.46
Wall (far)	1	2	131.33
Door	3	5	128.51
Wall (right)	1	8	88.23

For the structural features in Room-2 observations, the reference to the bounded space (room) was again the most highly ranked node in terms of degree measure and centrality measures (Table 3.15). This was followed again by two of three walls (right, left). The left wall had the highest in-degree count and connectivity measure, and although the right wall had a similarly high in-degree, the far wall had a higher level of connectivity in the network. This again may suggest that the far wall's role in the network is critical in the indoor scene description structure.

Based on the results of the network analysis, there are a number of patterns to consider as a part of any rules created for an intelligent indoor scene description agent. First, the room's structure objects played a central role in the organization of objects within the descriptions as landmarks to "chunk" objects together. Second, there was some evidence of a typical size and density of a network representing a scene description, approximately 30 unique nodes, 30 unique edges with an approximate distance of 3.0 and density of .04.

3.5. Linguistic Analysis of Spatial Prepositions

Beyond the patterns in frequency, position, and association, what does this data suggest about the semantics of the prepositions used to describe the indoor space? An analysis of prepositional semantics must consider both conventional use senses (Lakoff, 1987; Tyler and Evans, 2003; Vandeloise, 2006) as well as other contextual factors including cues that interact with the object's topology characteristics or the object's function. The purpose of this analysis was to determine which prepositions were most frequently used by observers, the manner in which they were used (spatial or functional), as well as spatial synonyms used in place of the complex primitive (Table

3.16). Using the data collected, we analyzed the use of the spatial preposition ‘on’ and its semantically similar spatial relations in the scene descriptions to see to what extent spatial references of objects are favored over their functional roles.

Table 3.16: Frequency of spatial preposition *on* and primary senses

Complex Primitive/ Primary Sense	# Instances	Central Case/ Function	Spatial synonyms	Triple pattern examples
On	70	Spatial-*Support,* *Contact,* Non-Spatial-State *Functional* *Actioning*	Against (23) Across (5) Along (3) verb-touch (5) attached (2) *(none observed)*	“trajector *on* wall” (35), “trajector *on* side”(10)

The spatial preposition *on* has a variety of spatial sense meanings that can be analyzed using a *polysemy approach* (Tyler and Evans, 2003). The following analysis of the use of the spatial preposition *on* in the indoor scene description dataset is based on a semantics theory of lexical concepts and cognitive models (LCCM; Evans 2006; 2009; 2015). An example of a proto-scene and the semantic structure of the spatial sense of the preposition *on* illustrates these concepts (Figure 3.5).

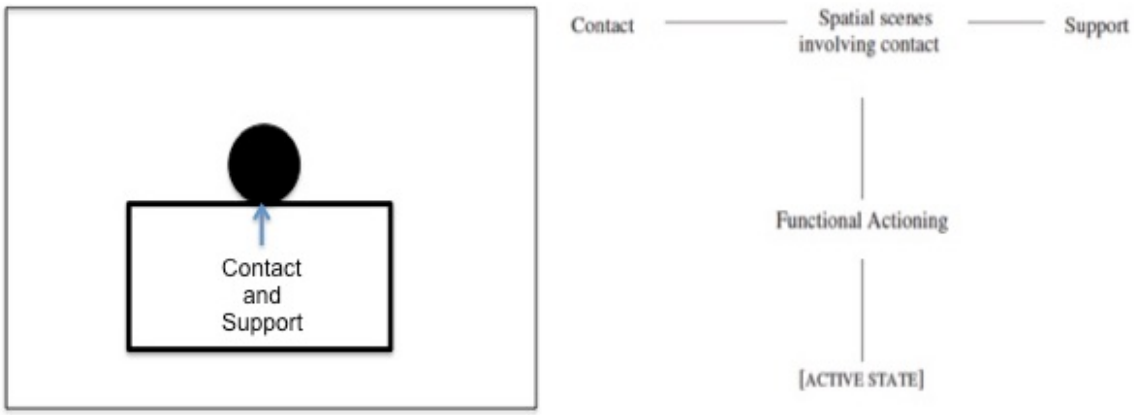


Figure 3.5: *Proto-scene and lexical concepts of on (Evans, 2015).*

Lexical concepts for the spatial preposition *on* involve the use senses of *contact*, *support* or *proximity* to the surface of a landmark (LM). The resulting function of this relation is that the TR is being supported or upheld by the LM or is in close contact with it. An example lexical concept in this case would be:

The computer is on the desk.

The above example illustrates a case where both senses *Contact* and *Support* are encoded by the lexical concept *Contact*. However, based on the utterances observed in the indoor scene descriptions, this encoding of *on* may be too limiting. Evans (2015) suggests that if an object like a computer is held against the wall by someone or something (e.g., *Support*) then the phrase below would be semantically different than *Contact*, unless the computer was attached to the wall by perhaps glue or a shelf, in which case, the phrase would be semantically the same.

The computer is on the wall.

However, analysis of the indoor scene description utterances suggest this may not be the case, as *contact* is the primary sense expressed in the observations over *support*. The spatial triples in the observations that used *on* do not require the use of both senses to appropriately convey the relationship between TR and LM. For example, the most frequent use of *on* in this dataset is in relation to a TR, usually a smaller, moveable object in the space with a structure or boundary as the LM. In most of these cases, there is no other meaning conveyed beyond *contact* or *proximity* (Figure 3.6) (e.g., file cabinet on wall [right]). There were a few cases of dual *support* and *contact* sense but only in spatial expressions of a tabletop space (e.g., knick-knacks on bookshelves), not an indoor vista scale space.

The closest formal spatial relation to the *contact/support* sense of the term *on* is the *contact* (9-Intersection) relation and *connection* (GUM-space) relation. So exactly what is the functional interaction of the wall (LM) with the desk (TR) in this use sense? In most observed utterances of this type, the wall's primary role is as a ground in a spatial configuration where the larger structure locates the smaller, more likely, moveable object. However, because these observations did not require any spatial behavior or task to be performed during the observation that involved the wall or any other object in the room, it is possible that the wall might serve a more active, functional role in spatial task specific scenarios.

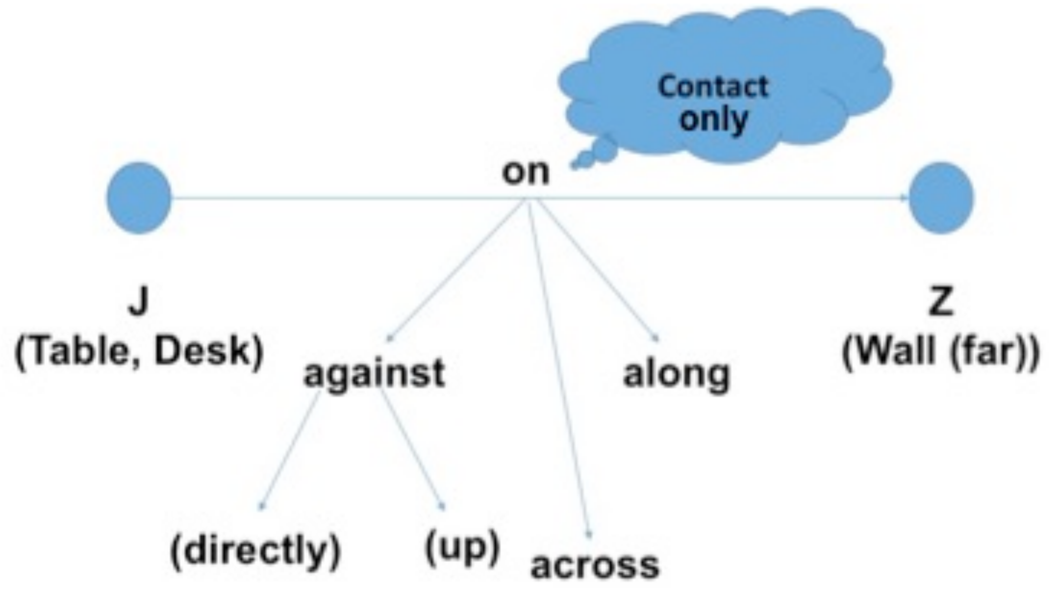


Figure 3.6: *Grounded moveable object on structure with spatial synonyms*.

When we examine other spatial prepositions or prepositional phrases identified in the dataset that might be semantically similar to *on* in the contact/contact or support sense such as *against, along, v. touch(ing)* and *v. attached to*, we can see that again, most of the relationships convey a *contact* sense (e.g., bookshelves against wall [left]) rather than a discrete or distributional *support* sense (e.g., desk along wall [far]) relation or a dual *support* and *contact* relation (e.g., bookshelves sticking out from wall [left]). (Figure 3.7) Other terms with similar semantics to *on* with a *contact* sense such as *against*, *along*, or even *comes out from* could be depicted with the same proto-scene as the primitive ***on*.**

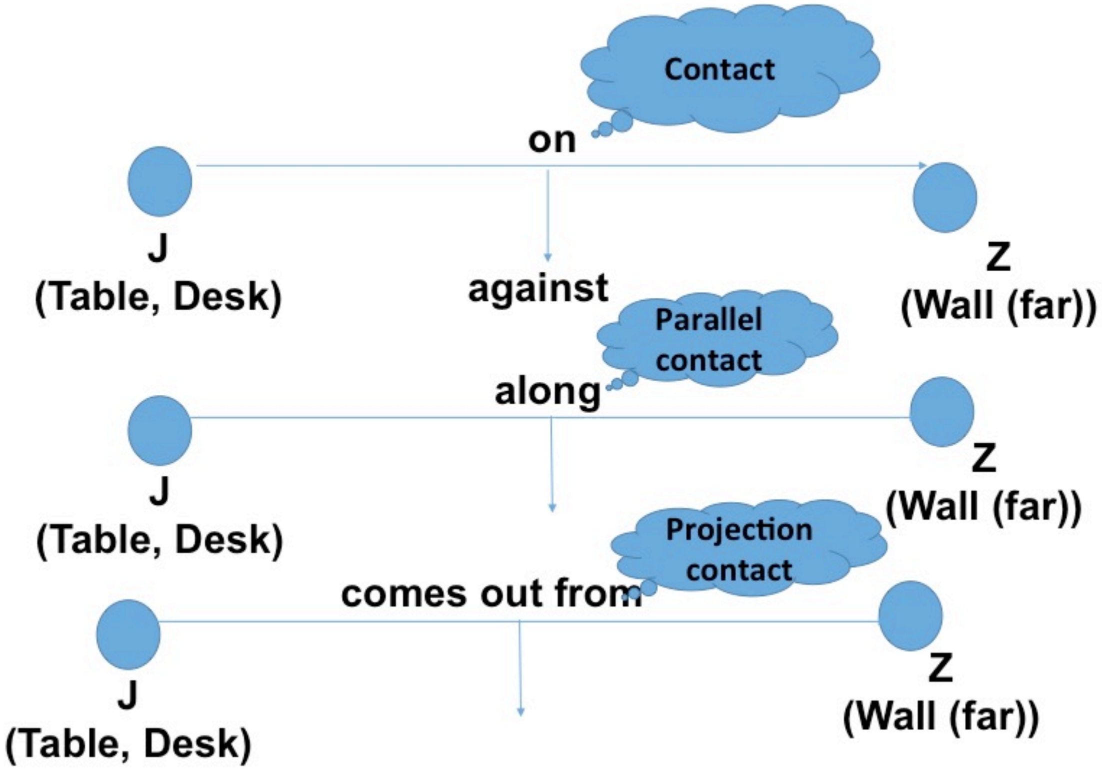


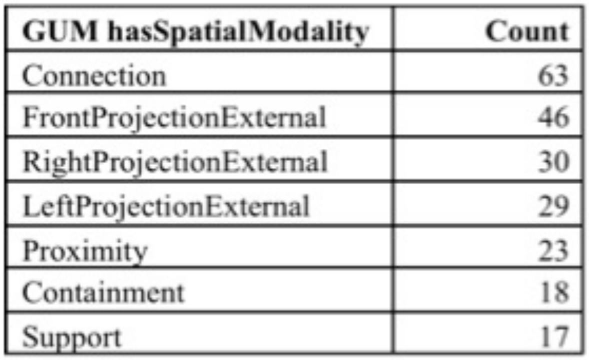
Figure 3.7: *Expanded relations for on.*

These patterns can be seen in other objects through an adjacency graph of all instances of the use of the spatial preposition *on*. In only a few cases is the tabletop space *support* sense used (e.g., knick-knacks on bookcases). In most cases, the use of *on* was a preferred term over other alternative adjacency expressions.

3.6 GUM Concepts Using Spatial Relation *On*

The next analysis maps the spatial relation *on* to the concepts in the spatial linguistic ontology, GUM-space, in order to determine what concepts were dominant according to this more expansive schema. The data suggest that there were seven primary ontology concepts where *on* was used (Table 3.17) starting with the *connection* concept.

Table 3.17: Frequency of Use Sense of *On*



GUM hasSpatialModality	Count
Connection	63
FrontProjectionExternal	46
RightProjectionExternal	30
LeftProjectionExternal	29
Proximity	23
Containment	18
Support	17

The annotation analysis mapped the use of *on* to *connection* concept in 21 instances. Other spatial prepositions used to represent *connection* include instances of *against (25), touching(6), and attached to(2)*. Other high incident uses of *on* associate with the *HorizontalProjection* concept group, specifically *LeftProjection*, and *RightProjection*. The typical use for *on* in these concepts was "*on left/right side*" or "*on left/right of*". Lower incident uses for these concepts were "*to right/left side/of*" or "*facing left/right*". Finally, *on* was infrequently used to represent the *support* concept.

This mapping of GUM-Space concepts represented by the spatial term *on* provides more support for rules placing *on* as a primary preposition to organize the spatial expressions calling for the connection, support or projection concepts with alternative terms used to provide more specificity if required by the user or the task.

3.7. Discussion of Results

The next section synthesizes the results of various analyses of the indoor scene descriptions. Results are discussed in terms of how indoor scene descriptions might help to better classify and describe objects, structures and relations within indoor spaces in relation to existing semantic concepts of indoor space.

3.7.1. Annotation Analysis Results

The annotation analysis provides basic frequencies of syntactic structure and general categories of objects, structures, and prepositions within and across observations. It does not tell us which specific objects or structures were critical in the descriptions nor does it tell us anything about the relationships between entities other than they were a component of a spatial triple. Based on the results, we identified that certain moveable objects (i.e., desk/table, file cabinet and bookcase) were most frequently mentioned in the descriptions along with certain structures (i.e., wall, side, room). Likewise, the most frequently used prepositions were *of, on, about*, and *in*. These results suggest that some types of objects/structures are featured more prominently than others. This analysis demonstrated the variability in description detail in terms of the number of utterances and number of spatial triples used in each description. It also illustrated the dominance of region and direction types of spatial triples over distance type which may be indicative of small scale indoor spaces.

The set of spatial relations used in the descriptions were somewhat limited and did not express formal relations found in models such as the 9-Intersection (Egenhofer, Herring, 1990). Instead the relations reflect more conceptual terms for object relations such as *contact, disjoint,* and for walls with windows, *partof,* may be appropriate for lack of a better term. Also, because the indoor scene descriptions were recorded as open observations given through an unstructured verbal response, we do not know how the types of spatial expressions might differ given a different response format (e.g., typed vs. oral response). Finally, given the directions provided to participants about creating a scene description for someone who could not directly view the scene, we note the unexpected high frequency of the use of underspecified spatial prepositions, such as *on* and *in*, which have broader and more potentially ambiguous spatial semantics.

The frequency of reference to room structures within the room descriptions points to the need to conceptualize a room space as a set of structure objects that happen to bound the void that is the room space (structure objects + void= room/container) (Hahmann and Brodaric, 2013; Brodaric et al, 2017) as illustrated in Figure 3.11 and Table 3.18.

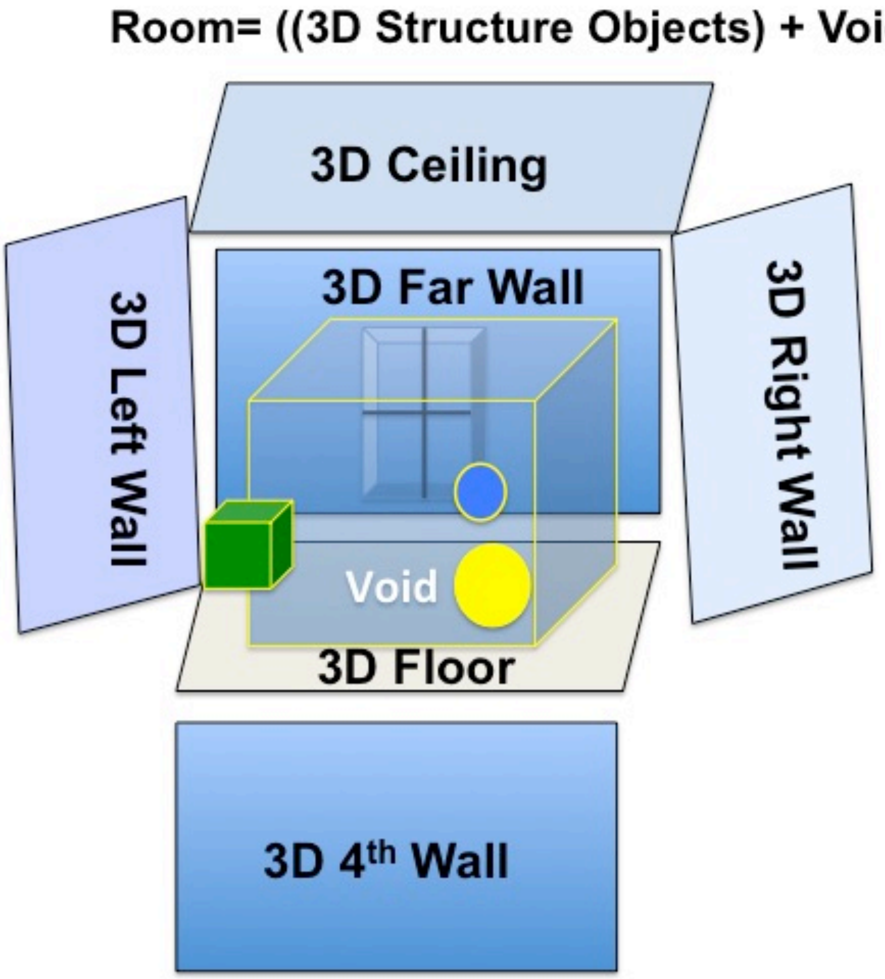


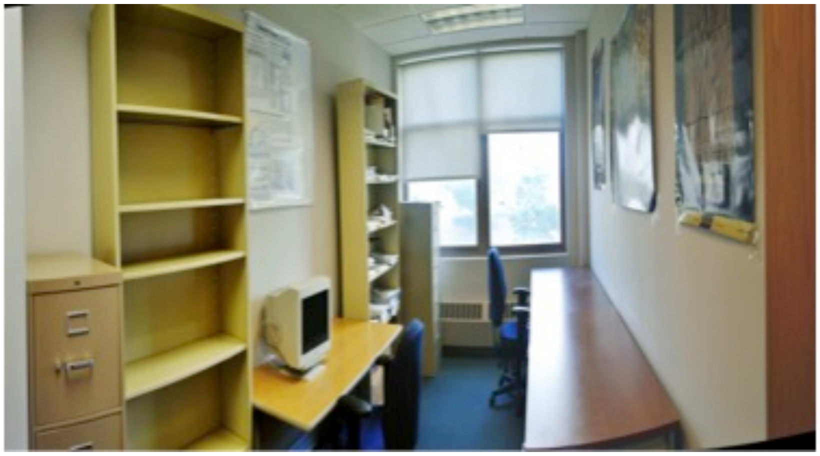
Figure 3.8: *Room represented as set of objects and the void*

Table 3.18. Objects in an Indoor Scene

Objects in Container	**Physical Instantiation Example**
Room structures	Some walls
Room structure	A single wall, ceiling, floor
Room structure	A window or a door
Moveable Objects	Furniture (e.g., bookcase, desk, table, chair)
Room space	Void enclosed by all room structures
Room	Room space and enclosing room structures

Based on evidence from the description analysis, we propose a conceptualization of a room as comprised of a number of different types of objects that conceptually participate in *contact*, *disjoint,* or *partof* relations. This conceptualization aligns with the placement of a person inside the scene. From inside (or at the doorway) the perceptual objects available to the user for description include a set of structural objects and moveable objects. The term moveable object is not used in the literal sense but instead in the broadest sense. These are objects that have the potential to be moved, not based on how heavy they are (e.g., 500 lb desk) or if they are physically attached to something else (e.g., bookcase attached to a wall). They are not a part of an existing structure object which would need to be disassembled in order for one part of the object to be removed from the other (e.g., window in wall).

One of the subject's scene description illustrates this perspective and this conceptual and linguistic pattern is shared among all of the open scene descriptions (Figure 3.12). The observer first situates herself in the room, and then proceeds to describe the walls and windows as individual structure objects, not as a part of a continuous room boundary. These brief description utterances primarily use the relation *on* to relate a moveable object to an individual wall object. Next, the observer describes the windows as “on the far wall” rather than using language signaling some type of containment relation (*surrounds* or *inside*) or parthood relation (*part of* or *intersects*). Most of the moveable objects are in relation to a structure object before they are described in relation to another moveable object. The description contains a collection of conceptual and linguistic features that illustrates the fact that the observer, once situated within the indoor scene, describes the room/container as a set of object types in relation to one another and the structure type objects function, primarily, as landmarks for referencing the location of moveable objects.



> "I am in Room 2 observing in real time. On the right wall… the room is about 14 feet by 8 feet. There are two large windows on the very furthest wall. On the right wall there is … there are two desks that are length wise side by side sticking out about… 3 feet from the wall…and …about…12 feet wide… or 12 feet in length, down the right wall. And there is a chair sticking out about a foot out from the second table. On the left wall, right in front of the window, there is a filing cabinet sticking out about three feet and it's about a foot in width. There is … a cabinet,… bookshelf 1, which is about a foot in … length and sticking out from the left most wall for about a foot. Then there is desk 3… that is sticking out about … 4 feet and there is a chair in front of the desk sticking out about a foot. And continuing on the left most wall, there is another bookshelf, bookshelf 2, that is sticking out about a foot... and it is ... two feet in length. And there is filing cabinet number 2, which is about a foot in width and sticking out from the wall out …2 feet…"

Figure 3.9 *Image of Room 2 with Scene Description*

This observation displays typical conceptual and linguistic patterns for all of the open scene descriptions collected by Kesavan and Giudice (2012) and re-analyzed in this dissertation (Chapter 3). There are several aspects of this description that raise questions about both the scene conceptualization, and the communication regarding the observed real-world scene. It should also be noted that there is a relatively small set of spatial prepositions used to represent all of the different relations between these objects, primarily *on*, *in, in front of, sticking out from.* Given all of the potential terms that could have been used in this description, the questions that immediately arise include: ‘Why these relations?’, ‘Why these terms?’ and ‘Why so little variety of relation terms?’. The next sections consider the results of the functional and linguistic patterns observed in the scene descriptions.

3.7.2. Functional Characteristics Analysis Results

The analysis of anthropomorphic characteristics provides a way to look at observer perceptions of physical and perceptual access or what observers sensed (e.g., visual) in the environment. It also provides some indication of the order of the potential encounter and the perceived directional type (general or lateral). It does not show the relationships between objects, structures and relations but instead provides a way to visualize and describe any spatial configuration or functional role characteristics among them. The observations differed slightly between rooms, in that Room-1 observers were more likely to use an *intrinsic perspective* and move through the description in either a dominant near/far or far/near access pattern with right and left entities following (e.g., *near-right* and *near-left*). Room-2 observers did not start from an intrinsic perspective but instead began from the right side of the observer. Only a single observation explicitly featured vertical access structures (i.e., floor, ceiling, absence of stairs) and only two observations included what was perceived to be behind the observer. Finally, there were few utterances in which entities were described with spatial prepositions denoting functional roles over simple spatial configurations. However, attention to object vocabulary choice points to implicit functional properties of objects and structural relationships (e.g., map/poster and wall [*display/read function*], table/desk and chair [sit/work]). This analysis demonstrated that variability in the start point perspective and sequential descriptions, may be the result of spatial structure of entities within the indoor space, and that under certain circumstances, and for the purpose of basic scene descriptions functional properties of objects/structures may be implied rather than explicit. However, the potential functional role and properties may be stored in the knowledge base in order to provide sufficient detail should that information become important in a task-based scenario or the specification of user need.

3.7.3. Network Analysis Results

The spatial network analysis provides insight into the specific structure of the spatial configurations within and across observations as well as the relationships among individual objects and structures. It also provides quantitative measures of the networks' connectivity, the strength of those connections, and how objects/structures cluster within specific indoor settings. It does not provide any measure on which relationships are critical in providing sufficiently constrained or expanded semantics of relations between spatial entities.

Based on the results of the network analysis, there was a similar number of node-edge relations as well as mean network density and distance. Both rooms were described using networks of a similar size and density which may point to possible minimum/maximum ranges to provide a sufficient amount of spatial detail at smaller scales. We also found specific objects and structures played a more central role in the networks across observations. For example, although the annotation frequency counts tell us how many times the object type "table/desk" was used, only the network analysis could illustrate which specific table/desk was more central to the description and what other objects/structures were most strongly connected to that particular table/desk in the network. In Room-1 observations, the “table/desk (far)” and the “room” were the most highly connected object and structure, whereas, in Room-2 observations, it was the “table/desk (left)” and the “room” that were the most highly connected nodes in the networks. Based on all of the observations in both Room-1 and Room-2, the other pattern discovered was that the wall nodes were ordered in connectivity from wall (left) to wall (far) to wall (right). This pattern may suggest some general rules for structuring scene descriptions and the clustering of objects may provide a way to ‘chunk’ objects and structures within those descriptions.

3.7.4. Linguistic Analysis Results

Finally, the in-depth linguistic analysis of prominent spatial prepositions in the observations provides a way to examine the primary semantic sense of the relations in its contextual use. In the case of *on*, its most frequent use sense was strictly in the Contact or Connection sense (e.g., TR [moveable object] on LM [structure object]) as well as a smaller number of instances using the *support* sense. There were no instances of the use of *on* in the functional active state sense even though, according to GUM-Space, the *support* concept is considered a functional modality. This analysis also provided ways to map out semantically similar spatial prepositions using the *contact* sense such as *against*, providing additional terms to convey a more specific type of *contact*. This mapping of semantically similar prepositions provides the basis for further investigation of similarity, clarity and preference of spatial prepositions based on more structured spatial expression prompts.

3.8. Conclusion

This chapter describes an analysis of indoor scene descriptions that combines methodology from spatial cognition, spatial linguistics, and spatial networks. Findings from this analysis support further examination of the use of NL spatial prepositions for a small subset of spatial concepts. Questions for further investigation related to this analysis include: (1) What set of spatial prepositions are used to describe the specific types of conceptual spatial relations found in the indoor scene descriptions (i.e., *containment, contact, disjoint, partof*)?; (2) How might the description response format (oral versus text) for certain types of user constraints (i.e., vision impairment) impact the types of spatial expressions used in indoor scene descriptions; (3) How similar or how clear are spatial prepositions in comparison to one another for a specific type of indoor scene?; and (4) What context factors impact the use of spatial prepositions in indoor scenes? Chapter 4 presents a series of experiments based on the results of the analysis of indoor scene descriptions that attempt to expand upon the findings and the open questions raised by the analysis described in this chapter.

CHAPTER 4

EXPERIMENTS FOR NATURAL-LANGUAGE TERMS

This chapter outlines a set of human-subject experiments designed to investigate natural-language structures used to describe and interpret spatial relations within indoor scenes. This topic has been examined across several disciplines in both table top and geographic space. In contrast, the experiments described in this chapter are situated explicitly within the vista-scale virtual indoor environment. This chapter describes three experiments that employ virtual indoor scenes to replicate 3D indoor spaces. My contribution to the existing body of research is to extend the understanding of how people conceptualize and communicate conceptual spatial relations through spatial prepositions at the indoor vista-space scale. The following experiments seek to better understand human-generated NL expressions as applied to conceptual spatial relations. The results of the experiments provide more specific knowledge about what information and terms constitute a correct and concise description of an indoor scene that includes both context and spatial references in indoor settings.

4.1. Experimental Stimuli

The virtual environment images used in this study were created in the University of Maine's Virtual Environment Multimodal Interaction (VEMI) Lab using Unity, a virtual reality design program (www.unity3d.com). The objects (i.e., assets) in the virtual environment were purchased through the online Unity asset store and modified for their use in this study by graduate students in the VEMI lab. The set of furniture was purposely chosen to match the same types of large, moveable objects found in the previous indoor scenes (Chapter 3). The moveable objects used in the rooms included bookcases, tables, desks, and office chairs. The rooms were designed to also align with the perspective used in the earlier scene description environment, that is, they present a perspective of a room where the entirety of the room could be perceived from a single location without motion, except the space in back of the participant (Figure 4.1). The specific room sizes (small: 10’x 12’; large: 20’x 30’) were selected, based on previous studies that found changes in the size of vista-scale spaces appear to be a significant factor in exploration search strategies and performance (Pingel, Schinazi, 2014). Context of the virtual rooms was designed to test subject responses to a set of conceptual relations (*contact, disjoint, partof*) identified through the analysis of scene descriptions (Chapter 3). Relations of focus were between indoor structure objects (i.e., walls, windows, doors) and moveable objects (i.e., furniture).

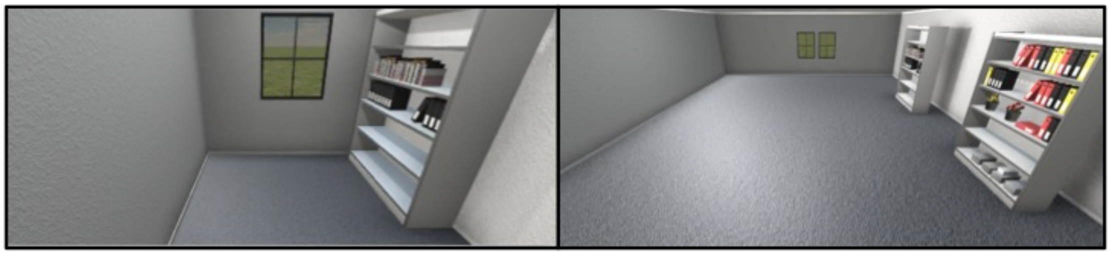


Figure 4.1. *Example images for small room (10’x12’) and large room (20’x30’).*

The experiments start with an open-ended solicitation of participant-supplied NL terms for pairs of objects and structures in the virtual spaces and move on to more constrained questions on term preferences. Each of the experiments attempts to build upon the findings of the analysis of indoor scenes and open questions to extract information about key elements necessary to generate minimally specific indoor scene descriptions for the conceptual relations identified within an indoor vista scale setting.

4.2. Participants

A total of 90 participants were recruited for the experiments. All experiments were approved by University of Maine's Institutional Review Board for Research with Human subjects. The first group consisted of 40 students (n=20 female, 20 male) from the University of Maine with a median age range of 20-24 years old (total range 18-34). All students identified themselves as native English speakers. The majority of participants (92%) reported they had lived in the northeast region of the U. S. from ages three to eighteen years old. Two students reported they had been raised in the southwest and one student reported being raised in the southeast regions of the U.S. from ages three to eighteen years old. The students were enrolled in a wide variety of program majors and were recruited through study opportunity announcements in introductory general education courses (e.g., Biology, Human Sexuality). Most had completed a portion of their college program (82%) and the remainder (18%) had completed at least an Associate degree. The lab participants completed the experiments in under an hour (m = 58 minutes) and they were compensated for their time with a $10.00 gift card to the university bookstore.

The second group of participants consisted of 50 Amazon Mechanical Turk Workers (MTurkers) (n=26 female, 24 male) with a mean age range of 24-34 years old (total range = 20-65). All MTurkers identified themselves as native English speakers. Participants reported a greater variation in where they lived from ages three to eighteen years old. Most reported that they were raised in the midwest (30%), northeast (26%) and southeast (26%) regions of the U.S. but there were participants who reported being raised in the southwest (10%), northwest (6%) and one participant was raised in Alaska.

Only MTurkers who were currently located within the United States were permitted to participate in the study. MTurkers also reported a greater range of educational attainment, ranging from a high school diploma (10%), some college program completion (28%), and the achievement of an Associate degree or higher (62%).

AMT has the ability to limit eligible participants to geographic regions based on MTurker IP addresses. A total of 55 AMT Human Intelligence Task (HIT) responses were originally collected. After a review of each completed AMT survey, five responses were rejected due to incomplete tasks or obvious language confusion indicating a potential non-native English speaker. AMT recruitment methods followed general guidelines for achieving gender-balanced results such as timing of HIT release and study description language (Crowston, 2012; Crump, McDonnell, Gureckis, 2013). AMT participants spent slightly less time to complete the experiments (m = 53 minutes) and were compensated for their completed and approved participation with a $5.00 HIT fee, which is well above the standard rate for similar HIT requests.

Overall, the total group (n = 90) achieved a sufficient distribution of gender, age, education and regional location. Due to the university setting, lab participants were younger as a group, grew up primarily in the northeast and most were in the process of completing a four-year degree (i.e., some college). AMT participants were somewhat older, represented more regional variation in the primary location during their childhood years and were more likely to have completed a post-secondary degree. In some studies, this variation between groups could be problematic, however in this case, the demographic variation of the MTurkers helped to diversify the total participant pool and explore any potential effects due to characteristics of the participants.

Table 4.1: Participants Gender

Setting	Gender		
	F	M	Other
Lab	20	20	0
AMT	26	24	0
Total	46	44	0

Table 4.2: Participant Age Range

	18-19	20-24	25-34	35-44	45+
Lab	17	20	3	0	0
AMT	0	4	21	14	11
Total	17	24	24	14	11

Table 4.3: Participant Region from age three to eighteen

	NE	SE	NW	SW	MW	AK
Lab	37	1	0	2	0	0
AMT	13	13	3	5	15	1
Total	50	14	3	7	15	1

NE: Northeast, SE: Southeast, NW: Northwest, SW: Southwest, MW: Midwest, AK: Alaska

Table 4.4: Participant Educational Attainment

	HS	SC	AS.	BS	AD
Lab	4	29	1	6	0
AMT	5	14	6	19	6
Total	9	43	7	25	6

HS: High School, SC: Some College, AS: Associate, BS: Bachelors, AD: Advanced Degree

4.3. Experimental Survey Instrument

All three experiments were constructed using the web-based survey program, Qualtrics Survey Suite (www.qualtrics.com). Each of the image-prompt items used in the experiment set (n =80) were coded with the following qualitative descriptions: image-prompt spatial relation, image room size, prompt feature pair, trajector (object or structure) orientation, and distance of trajector from observer (Table 4.5). The factors associated with room context were determined based on the findings of the analysis of indoor scene descriptions (Chapter 3) as having the potential to impact the use patterns of spatial prepositions.

*Table 4.5.*Experiment image-prompt variables

Spatial Relation	**Room Size**	**Feature Pair**	**Orientation**	**Distance from Observer**
contact	Small	Moveable object-Structure object	right/left	Far
disjoint	Large	structure object-structure object	Front	mid
partof		moveable object-moveable object	Rear	Near

4.4. Experiment 1: Indoor Image Prompt: Open Response

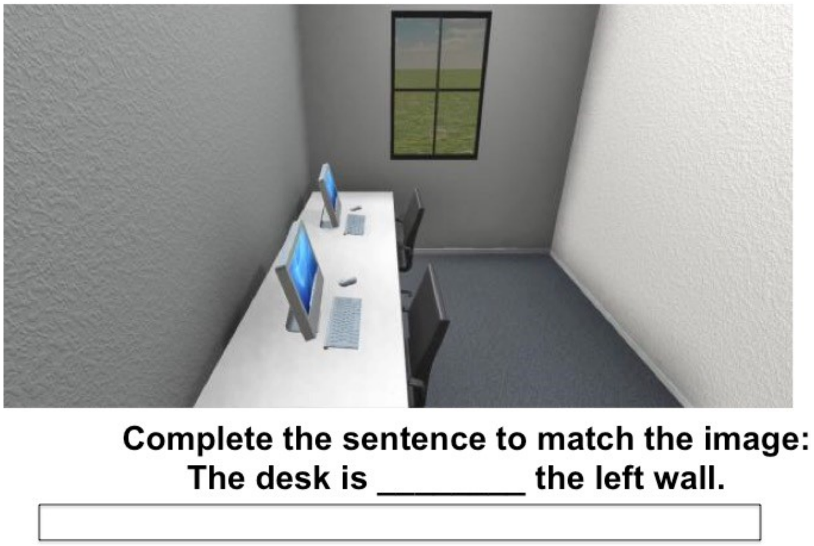
The first experiment investigates use patterns of spatial prepositions observed in the analysis of indoor scene descriptions between objects and structures in indoor vista space. It addresses the following research questions:

- What spatial prepositions are used to describe conceptual relations between objects and room structures in indoor scenes?

- How does response format and hypothetical scene recipient sensory constraints (i.e., lack of visual input) impact spatial preposition use?

Findings from the analysis of indoor scene descriptions (Chapter 3) suggested *on* was a statistically significant spatial preposition used to verbally describe the *contact* relation in a small vista-scale room. This experiment tests if the high frequency of the term *on* will be repeated in a more controlled experiment format and if frequency of use depends on modality (oral vs. typed-text). The results of this experiment allow for a better understanding of how spatial prepositions for object relations can account for uncertainty depending on the modality of the dialogue format (oral vs. typed-text).

In the first task, participants were asked to provide open responses to a series of 24 prompts about spatial relationships between objects and room structures in virtual indoor scenes. For each of the 24 images, participants were prompted to fill in missing spatial preposition(s) to describe the spatial relation between the specified moveable object (e.g., desk, chair, and bookcase) and structure object (e.g., wall, door, and window). For the lab participants, twelve of the open responses were collected verbally using a speech to text application and another set of twelve prompts required participants to type in their open response (Figure 4.2).



Complete the sentence to match the image:
The desk is ________ the left wall.

Figure 4.2 *Open response example.*

This first experiment seeks to answer a number questions:

(a) Is there any difference in the use of spatial prepositions to describe relation between moveable object and a structure object in an oral format as compared to a typed-text format? I predict there will be no difference in frequency use of spatial prepositions used to describe relations in oral versus typed-text based descriptions.

(b) Is there any difference in the use of spatial prepositions to describe the relations between object pairs in descriptions intended for sighted versus those descriptions intended for non-sighted individuals? I predict that spatial prepositions used in descriptions of indoor scenes given by sighted individuals (S) will not differ significantly from descriptions given by sighted individuals for non-sighted individuals (NS).

(c) Is there a difference in the use of spatial prepositions to describe relations between feature object pairs (moveable objects and structure objects) in different vista scale indoor spaces? I predict there will be a statistically significant difference in frequency of use of spatial prepositions based on room size.

(d) Is there any difference in the use of spatial prepositions to describe the relations between different types of feature object pairs in vista scale indoor settings? I predict there will be a statistically significant difference in the use frequency of spatial prepositions between feature pairs (OS, SS).

(e) Is there any difference in the likelihood of individuals' use of spatial prepositions to describe the relationship between object-structure pairs based on orientation/alignment of the object? I predict there will be a statistically significant difference in the use of spatial prepositions based on the object's axis alignment/orientation with another room object or structure.

(f) Is there any difference in the likelihood of individuals' use of spatial prepositions to describe relationships between object-structure pairs based on distance between observer and image objects/structures? I predict there will be a difference in frequency of use of spatial prepositions based on virtual observer distance to the feature pair in the image prompt.

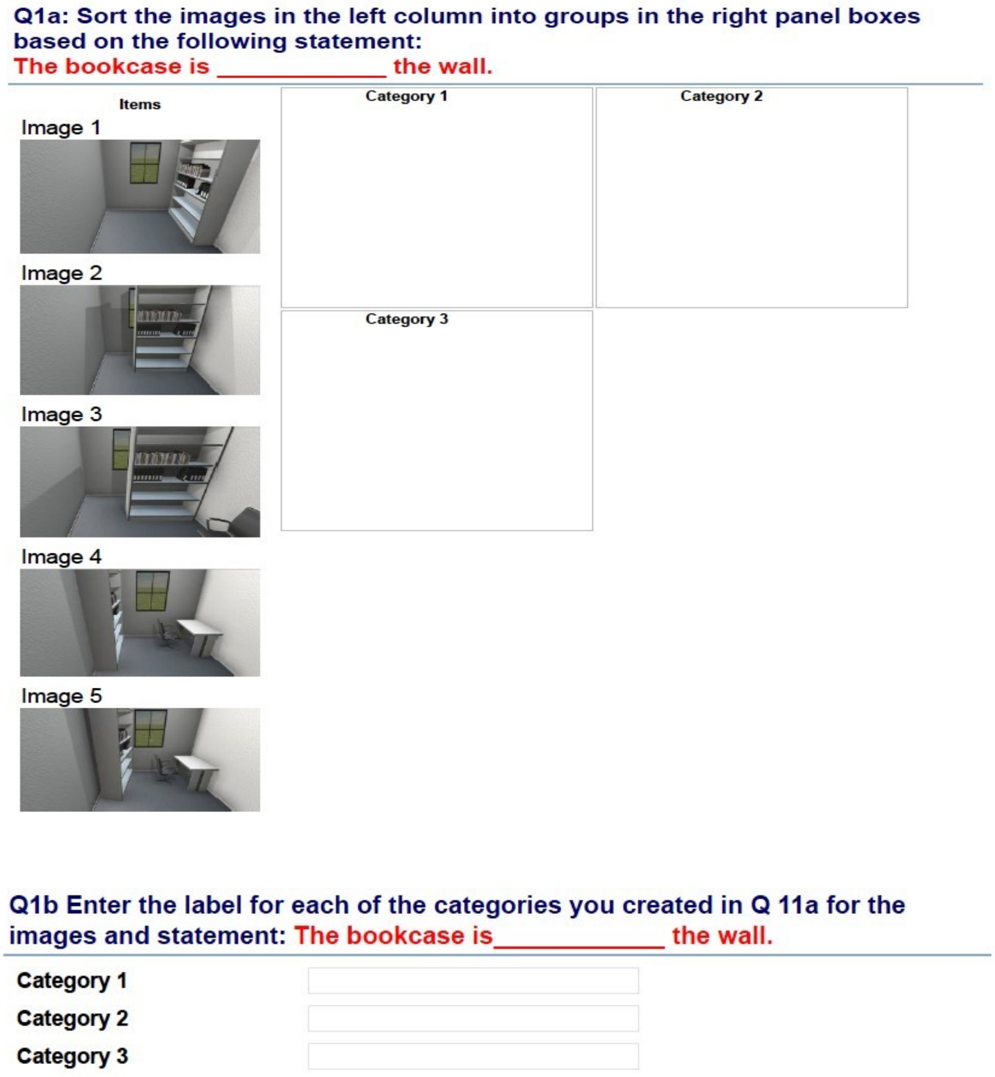
Table 4.6 Experiment 1 Outline

Experiment 1 components	**Question/Hypothesis (number of participants)**
Experiment 1 a	Oral vs. Text Response (Lab group only n = 40)
Experiment 1 b	Sighted vs. Non-sighted protocol (n = 90)
Experiment 1 c	Room size (n = 90)
Experiment 1 d	Feature Pair Type (n = 90)
Experiment 1 e	Object-Orientation (n = 90)
Experiment 1 f	Feature Pair Distance (n = 90)

4.5. Experiment-2: Indoor Image Categorization

Experiment-2 uses a category construction process to determine classification patterns in spatial relations given similar sets of objects and structures in different size indoor spaces. Based on frequency patterns of spatial prepositions found in the analysis of indoor scene descriptions, spatial prepositions were tested in both a free categorization task and in a forced categorization task (Figure 4.3). Participants were asked to classify five sets of five images of similar indoor scenes into three unlabeled groups (n=25 open sort and label items) and five additional sets of images into four pre-determined categories (n=25 closed sort items) based on their evaluation of the most appropriate spatial preposition to represent the *contact*, *disjoint* or *partof* spatial relations. This set of experiments adopts another method for asking two of the primary questions investigated in this dissertation: (1) *What spatial prepositions do people use to describe topological and conceptual relations between objects in a room?; and (2) What are preferred spatial prepositions to express spatial relations between objects in indoor scenes?*

We hypothesize that there will be a statistically significant difference in how images are classified based on the similarity of spatial prepositions used to represent feature pair spatial relations.



Q1a: Sort the images in the left column into groups in the right panel boxes based on the following statement:
The bookcase is _____________ the wall.

Items

Image 1

Image 2

Image 3

Image 4

Image 5

Category 1

Category 2

Category 3

Q1b Enter the label for each of the categories you created in Q 11a for the images and statement: The bookcase is_____________ the wall.

Category 1

Category 2

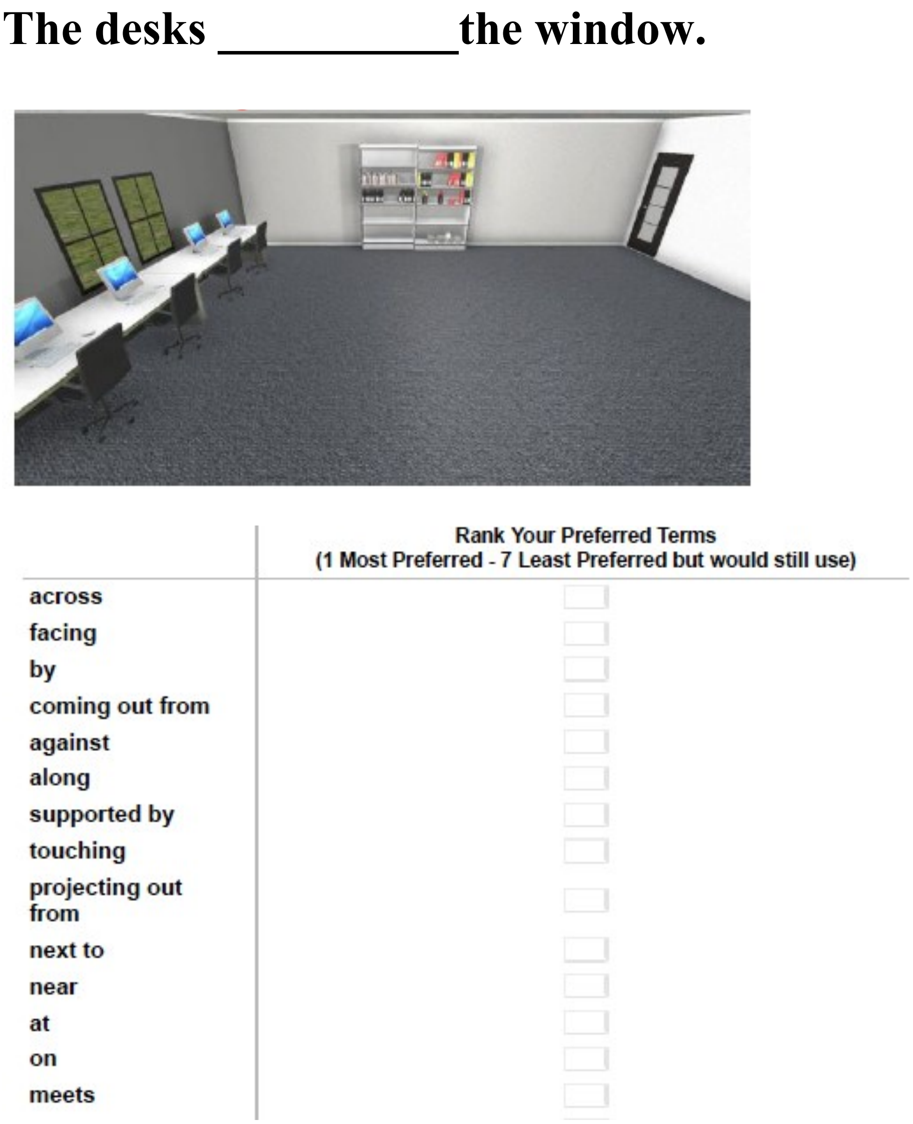
Category 3

Figure 4.3 *Image categorization example.*

4.6. Experiment 3: Indoor Image Ranking

The final experiment required participants to view five virtual scenes and evaluate spatial prepositions used for the same types of relationships based on three scales:

similarity, clarity, and preference. The image comparison and preference ranking experiment builds upon the previous two experiments to investigate the use patterns of spatial preposition for object and structure relations in indoor scenes (Figure 4.4). It is another method for asking the question: *Are there differences in the preference of level of specificity in spatial prepositions used in scene descriptions?* We hypothesize that there will be a statistically significant difference in ranks based on the similarity, clarity and preference of spatial prepositions used to represent feature pair spatial relations.



The desks __________the window.

	Rank Your Preferred Terms (1 Most Preferred - 7 Least Preferred but would still use)
across	□
facing	□
by	□
coming out from	□
against	□
along	□
supported by	□
touching	□
projecting out from	□
next to	□
near	□
at	□
on	□
meets	□

Figure 4.4. *Image Ranking Experiment: Preference Section*

4.7. Analysis

The table below summarizes the questions, stimuli format, and data produced for analysis in each experiment (Table 4.7).

Table 4.7: Summary table for Experiment 1-3

Experiment	**Question**	**Input/Format**	**Variables**	**Analysis**
Experiment 1	Object and relations identification	Images and single text expressions (50 items)	Relation/Prepositions, Oral v. Typed-Text, Intended Recipient, Room size, Feature Pair, Orientation, Object Distance	Descriptive Statistics, Chi Sq.
Experiment 2	Spatial relations classification and preposition identification	Images and three relation categories or four preposition categories (50 sorted images)	Relation/Preposition Classification and Labeling	Descriptive Statistics, Chi Sq., proximity matrices (dissimilarity), Multidimensional Scaling
Experiment 3	NL spatial relation language similarity, clarity and preference	Images and prompts (18 items) with similarity, clarity and preference ranking scales	Relation/Preposition Similarity, Clarity and Preference Ranking	Descriptive Statistics, Chi Sq., proximity matrices (dissimilarity), Multidimensional Scaling (MDS), Friedman test with post hoc (Wilcox signed rank test)

Initial data analysis methods were employed on data collected from each experiment for patterns within each prompt. Analysis included testing results of scalar items for normality of mean distribution and standard deviations. For categorical response items or for scalar data, where the standards for normality are not met, non-parametric approaches for testing associations were used.

4.8. Conclusion

This chapter outlined the study instrument, experimental design and research questions of this dissertation work. The experiments use the findings of the analysis of indoor scene descriptions described in Chapter 3 as the foundation for the selected spatial relations and spatial prepositions investigated and the questions that guide the experiments. Previous approaches regarding the factors influencing spatial preposition use to describe conceptual spatial relations provide the basis for the study design, methods, and procedures employed.

CHAPTER 5

EXPERIMENTAL RESULTS

This chapter presents results of the experiments as described in Chapter 4 regarding the use of spatial prepositions based on different response formats and intended description recipient. It also examines similarity, clarity and preference of spatial terms used to describe spatial relations between moveable objects and structure objects in virtual indoor scenes. The analyses also include use patterns of spatial preposition and room context features such as room size, feature pair type, object orientation/alignment, observer distance, and object-structure distance.

5.1. Experiment 1 Results

5.1.1. Oral vs. Text Response Format

In the indoor scene description protocol, there were explicit instructions for participants to provide an oral description that would represent the indoor scene for someone who was blind or low vision. Given the strong frequency of use of simple spatial prepositions such as *on, at, by,* and *in*, in the scene descriptions (Chapter 3), there was a question as to how the format of the oral response may have impacted the types of spatial prepositions used in indoor scene descriptions. Therefore, the design of Experiment-1 included two sets of similar questions that required two different formats of description response, one oral and the other typed-text based.

Examination of differences in descriptions based on response format used a mean count of words used to fill in each item prompt to create a complete expression that matched the given image. A paired samples t-test was conducted to compare the number of words in verbal response and text response conditions. Based on the mean number of words used to complete the expression, there was not a significant difference in the two conditions (t=1.169, p= .867): oral response format (M=4.06 words, SD= 1.48) and typed-text format (M=4.10, SD=1.77).

5.1.2. Sighted vs. Non-Sighted Audience Description Results

The analysis of indoor scene descriptions (Chapter 3), pointed to a significant under-specification of spatial relationships between feature pairs (i.e., high frequency of *on*) even though participants were told the oral description they were providing was for a non-sighted individual. In order to test how a hypothetical recipient's vision status may impact the spatial prepositions used in scene descriptions, half of the Lab and AMT participants were asked to create these short spatial descriptions for a hypothetical person "*who is sighted and using their phone or mobile navigation device to describe the scene*", while the other half of both groups were asked to create the short descriptions for a hypothetical user "*who is blind or has impaired vision and using their phone or mobile navigation device to describe the scene*". The groups were randomly assigned to each condition. All 90 participants generated a total of 24 open responses to assess differences in spatial preposition choice based on the two different hypothetical recipients.

Differences in spatial prepositions used to describe *contact* relations between objects were assessed by looking at the frequency distributions of spatial prepositions used as well as Mann-Whitney tests for both Lab and AMT participants (Table 5.1). Looking at spatial preposition use frequency for these types of spatial relationships, there was little variation in the terms used across both test groups. Most *contact* relations for object-structure (OS) feature types were described using the terms *on* or *against* in both test groups and for both protocols. A Mann-Whitney test confirmed there were no significant differences in the use of the most frequently used terms *on* or *against* for *contact* OS relationships for the hypothetical sighted and non-sighted users in participant group or in the total participant group. An independent samples t-test showed no significant difference ($p<.05$) in the mean number of words used to describe the spatial relationships in the spatial expressions for the hypothetical sighted recipient ($M= 3.63$, $SD=1.66$) and non-sighted recipient ($M=4.53$, $SD=1.22$) conditions of the experiment ($t=-1.95$, $p=.058$).

Table 5.1: Example spatial prepositions: sighted/non-sighted protocol

	Sighted Protocol		Non-Sighted Protocol	
	Lab	AMT	Lab	AMT
Q18 Against	40%	56%	30%	52%
Q18 On	40%	12%	35%	24%
Q20 Against	40%	56%	30%	40%
Q20 On	30%	16%	30%	28%

5.1.3. Indoor Image Prompt: Open Response Results

Based on the findings in the analysis of scene descriptions (Chapter 3), there was an expectation of significant variation in terms used to describe *disjoint* relations and the high frequency use of *on* to describe *contact* relationships between objects and structures.

The results from this experiment confirmed the high frequency of the use of *on* for *contact* relations. However, there was less variation in spatial terms than in the scene descriptions and *against* was chosen just as frequently to describe *contact* relationships between moveable objects and structure objects (OS) as the term *on* (Figure 5.1, Table 5.2).

Table 5.2: Examples of spatial prepositions: *contact* OS relations

	Term Rank and Percentage of Use				
	1	2	3	4	Other
Q3	On 27%	In 20%	At 12%	Against 12%	29%
Q4	On 45%	To 15%	Against 15%	In 10%	15%
Q16	On 45%	Against 20%	Along 5%	In front of 5%	25%
Q18	Against 45%	On 27%	Along 10%	In front of 5%	13%
Q20	Against 42%	On 26%	In front of 13%	Touching 4%	15%

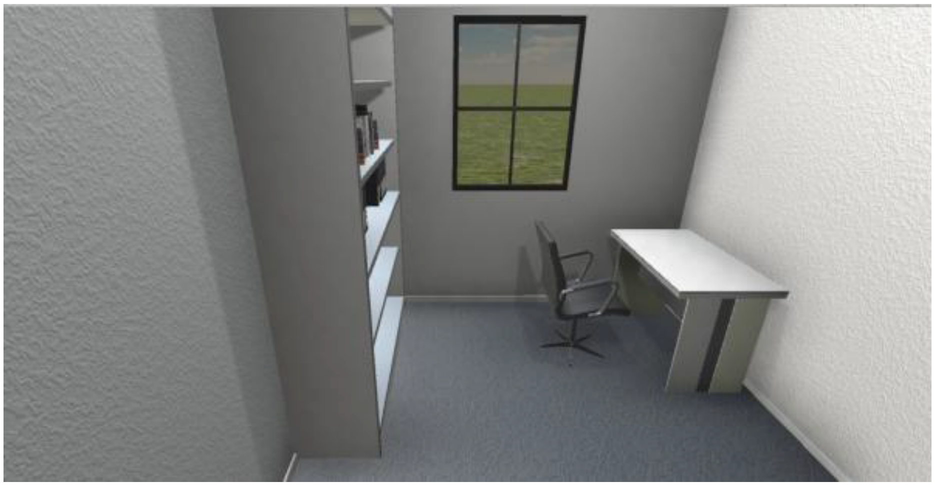


Figure 5.1: *Example contact OS Item: The bookcase is ____ the left wall.*

When describing a *partof* relationship between a window or door and a wall (SS), *on* was the most frequently provided open response term (Table 5.3). Terms supplied for *disjoint* relationships between room object and structures illustrated the greatest variation in spatial preposition use, with *near* and *next to* being chosen most frequently to complete the prompt. (Table 5.4). A chi-square test was performed to determine whether terms were a statistically significant response. No single preposition response in the *contact* and *disjoint* OS relations reached a statistically significant level (p<.05). However, for several *partof* relations, *on* was chosen at a statistically significant level (X^2 range (2, N = 90) = 9.44 to 17.09, *p*<.01). Notably, *on* and *against* are most prevalent in the first two rankings for prepositions in the *contact* relation.

Table 5.3: Examples of spatial prepositions: *partof* SS relations

	Term Rank and Percentage of Use				
	1	2	3	4	Other
Q1	On 71% **	In 8%	To 8%	At 7%	6%
Q2	On 66% **	At 18%	In 6%	To 3%	7%
Q13	On 53%	In 40%	~	~	7%
Q14	On 45%	In 45%	~	~	10%

** p<.01 ; ~ other individual responses < 2%

Table 5.4: Examples of spatial prepositions: *disjoint* OS relations

	Term Rank and Percentage of Use				
	1	2	3	4	Other
Q5	To 53%	Near 10%	Close to 12%	In Front of 10%	15%
Q7	Near 12%	To 11%	Close to 11%	Next to 9%	57%
Q21	Next to 31%	Against 26%	On 10%	Near 6%	27%

Table 5.5: Examples of spatial prepositions: *disjoint* OO relations

	Term Rank and Percentage of Use				
	1	2	3	4	Other
Q12	Next to 22%	In front of 15%	To 11%	Against 6%	45%
Q19	Next to 23%	Behind 23%	To 16%	In front of 9%	29%
Q22	In front of 56%	To 12%	Next to 10%	Behind 3%	19%

5.1.4 Levels of Uncertainty in Open Response Prompts

Based on word counts used to complete the description prompt, there was variance in both the number of words used to complete the prompt and a variety of different terms used as the primary spatial preposition. There were no statistically significant differences between the spatial relations prompts based on mean number of words. Table 5.6 provides a summary of the prompt responses by spatial relation and by feature pair type. Spatial preposition or unique terms represent the most frequently used preposition type among all of the items in that type, and range represents the minimum and maximum number of words participants used to fill in the prompt. Next the mean, median and mode number of words are provided across items for the category type along with the standard deviation and variance across items. Finally, many responses contained additional spatial prepositions, objects and structures that served to triangulate the spatial relationship between the identified objects and structures in the original prompt.

Table 5.6: Word count for open responses based on relation and feature pair type

	Spatial Prep. (unique terms)	Range	M/Mdn/Md	SD/Var.	Add. Ref. Entities
Contact OS	*On/Against (11)*	1-13	3.38/2/1	2.25/5.34	*corner/window/door*
Disjoint OS	*Next to/Near (16)*	1-12	3.46/3/2	2.42/5.99	*room/window*
Partof SS	*On (8)*	1-12	3.28/4/4	1.94/3.83	*(other)window*
Disjoint OO	*Next to/In front (17)*	1-13	3.80/3/2	2.61/6.94	*wall/corner*

A text analysis of the prompt responses also supports the importance of room structures such as windows, walls, and undefined features such as corners in the descriptions as secondary landmarks when a participant used more than one spatial expression to complete the prompt. There were very few cases where objects such as desk, bookcases or chairs were used to anchor or co-locate a trajector to a landmark. Instead, these intermediary spatial landmarks consisted of room structures without clear boundaries. This is consistent with the earlier analysis of indoor scene descriptions where walls were the dominant structure object in all of the participant utterances and were strongly associated with the landmark position in the spatial triples as opposed to the trajector position (see Chapter 3).

Each of these classifications had five items whose response were calculated and averaged to calculate the category descriptive statistics. Based on these data, it would seem that participants had a greater level of certainty as to the *partof* relations between structures objects (e.g., windows, doors, walls) due to the smaller mean number of terms used in the prompt responses (mean = 8) and lower variance (var. = 3.83). Next, it would seem that there was increasing uncertainty moving from *contact* OS to *disjoint* OS to *disjoint* OO pairs. Having some guidelines on the number of words used to complete each of these prompt types is useful. Although prompt types were completed with a minimum of one word to a maximum number of thirteen words, in general, most prompt responses were completed in three to four words. These data are consistent with utterances observed in the earlier analysis of indoor scene descriptions. On average, there were approximately five nouns (e.g., objects/structures), two verbs and three prepositions used per utterance. If the three to four words that form the spatial expression in the open prompts are added to the five to six words that formed head and foot for each of the prompts, it would easily arrive at a similar length of syntactic structure as the open description sentences (Chapter 3). This observation points to a possible optimal length and structure for sentences describing spatial relationships within indoor scenes. Specifically, based on the findings of this dissertation a concise spatial triple should take the following form and length: trajector (≤ 3 words) + spatial preposition or prepositional phrase (≤ 4 words) + landmark (≤ 3 words) = spatial triple (≤ 10 words).

The next section refines the open response experiment with two image sorting experiments. In the first sorting experiment, participants viewed and grouped images of indoor scene and then label the categories based on spatial information in both the image and the prompt. The second experiment was a closed sort task where participants were viewed and grouped images into four named categories (*on, against, along*, and *near*) in order to better understand participant conceptualizations about the underlying relations between the images.

5.2. Experiment 2 Results: Indoor Image Sort: Categorization

5.2.1. Open Sorting/Labeling

The open sorting experiment consisted of five items, each with five images to sort and classify. This task generated a total of 25 individual items for the section. Participants were asked to sort five images into three boxes and then classify the boxes by giving a name that matched the spatial relations of the objects in the images. Both *on* and *against* were the terms used most frequently to label the ten images/prompts with the *contact* relationship between room objects and structures (range = 15%-40%) (Table 5.7). A chi-square test was performed to determine whether *on* or *against* was preferred over other possible choices. Preference for spatial prepositions was equally distributed in the population as neither *on, against* nor any other term was used to label a category at a significant probability level ($p<.01$).

Table 5.7: Frequency of spatial prepositions: *contact* OS relations

	Term Frequency and Percentage of Use				
	1	2	3	4	Other
Q1Image 1	Against 39%	On 35%	To 15%	Touching 2%	9%
Q1Image 3	On 20%	Against 15%	Perpendicular 15%	Next to 9%	41%
Q1Image 4	Against 32%	On 32%	To 11%	Parallel to 5%	20%
Q2Image 1	Against 29%	On 25%	Perpendicular 5%	In front of 5%	36%
Q2Image 2	Against 27%	On 25%	Touching 8%	In front of 7%	39%
Q2Image 5	Against 29%	On 23%	Close to 6%	Touching 7%	35%
Q3Image 1	Against 38%*	On 28%	Along 8%	Touching 6%	20%
Q3Image 2	Against 37%	On 29%	Along 7%	Touching 6%	21%
Q3Image 3	Against 39%	On 28%	Touching 7 %	Along 5%	21%
Q3Image 5	Against 38%	On 26%	Along 8%	Touching 6%	22%

sig. p< .01 level * sig. p< .001

For the five images with a *partof* relation of a structure with another room structure (e.g., window and a wall), *on* was used most frequently (range = 66%-79%) (Table 5.8). A chi-square test was performed to determine whether *on* or *against* was preferred over other possible choices. Preference for *on* was not equally distributed in the population and was found to be significant for four out of five items (X^2 range (2, N = 90) = 8.71 to 17.77, *p*<.01)). Other terms used for this relation were either *in* or *in middle of* or *in center of* (range 5%-13%), however, a chi-square test determined neither of these terms reached a significant level of use (*p*<.01)

Table 5.8: Frequency of spatial prepositions: *partof* SS relations

	Term Frequency and Percentage of Use				
	1	2	3	4	Other
Q4 Image 1	On 66%**	In 9%	In Middle of 9%	In center of 5%	11%
Q4 Image 2	On 69%**	In 9%	In Middle of 9%	In center of 5%	41%
Q4 Image 3	On 59%	In 13%	In Middle of 11%	In center of 8%	20%
Q4 Image 4	On 75%**	In 5%	In Middle of 5%	In center of 5%	36%
Q4 Image 5	On 69%**	In 6%	In Middle of 6%	In center of 5%	39%

sig. p< .01 level * sig. p< .001

There were five *disjoint* relations for objects and structures in the open sort categorization (Table 5.9). Even in images with clear *disjoint* relationships between the targeted object and landmark structure object, spatial prepositions *on* and *against* were still in the top four terms chosen to describe and label the spatial relationship. *Against* was used most frequently to describe three of the five *disjoint* images (range = 27-31%) and *on* and *away from* were used to describe the remaining two spatial relationships between objects and room structure objects. A chi-square test was performed to determine whether *on* or *against* were preferred over other possible choices but they did not reach a significant level of use (*p*<.01).

Table 5.9: Frequency of Spatial Prepositions: *disjoint* OS relations

	Term Frequency and Percentage of Use				
	1	2	3	4	Other
Q1Image 2	On 19%	Perpendicular 13%	Against 11%	Next to 10%	11%
Q1Image 5	Against 31%	On29%	To 10%	Along 5%	41%
Q2Image 3	Against 28%	On 22%	Touching 9%	Next to 4%	20%
Q2Image 4	Away from 31%	Not touching 16%	Near 8%	On 8%	36%
Q3Image 4	Against 27%	On 28%	Along 8%	Touching 6%	39%

sig. p< .01 level * sig. p< .001

5.2.2. Open Sort Image Proximity and Spatial Preposition Categories

Further analysis of the data was conducted to evaluate the connections between the images in each of the five sets. A dissimilarity matrix was constructed in XLSTAT (www.xlstat.com) for each set of five questions for the open sort experiment. A dissimilarity matrix displays the distance function showing the dissimilarity between two items. Two items are interpreted to be more dissimilar if the distance between them is high and similar items have a lower distance between each pair. Diagonal elements in the matrix are zero because distance between an item and itself is always zero. For example, in the first set of images (Q1, Images 1-5 with the prompt "The bookcase is _______ the wall"), the dissimilarity matrix (Figure 5.2) suggests the images were categorized as dissimilar from one another, however, Images 2 and 4 are the most dissimilar images in the set (.989 disagreement). This can be interpreted that Images 2 and 4 were almost never grouped and labeled together by any of the 90 participants.

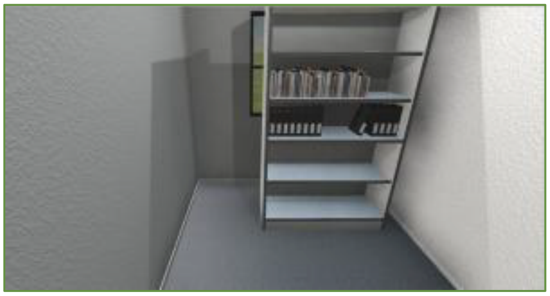


Image 2

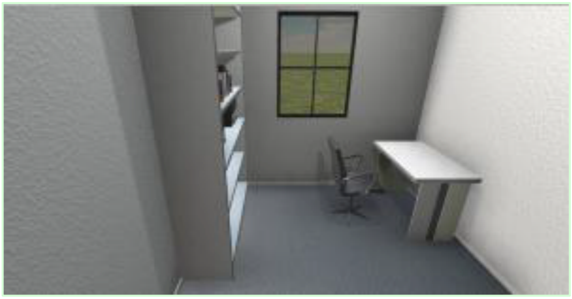
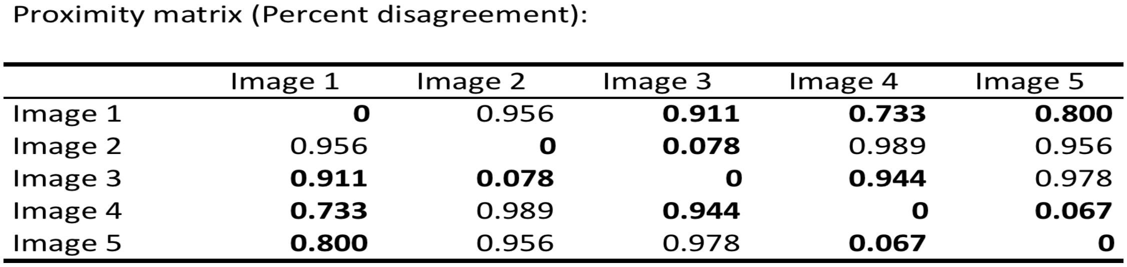


Image 4



Proximity matrix (Percent disagreement):

	Image 1	Image 2	Image 3	Image 4	Image 5
Image 1	**0**	0.956	**0.911**	**0.733**	**0.800**
Image 2	0.956	**0**	**0.078**	0.989	0.956
Image 3	**0.911**	**0.078**	**0**	**0.944**	0.978
Image 4	**0.733**	0.989	**0.944**	**0**	**0.067**
Image 5	**0.800**	0.956	0.978	**0.067**	**0**

Figure 5.2: *Example Dissimilarity Matrix* (Open Sort:Q1, Image 1-5).

In addition to the dissimilarity measures, Multidimensional Scaling (MDS) was conducted for each of the five sets of images. MDS is used in marketing research, user experience, evaluation, and psychometric testing to map responses from a proximity matrix (similarity or dissimilarity). In order to evaluate the quality of the representation, MDS algorithms use a criterion referred to as *stress*. The closer the stress measure is to zero, the better the representation. The goal of the analyses for the image grouping is to show how the images position themselves on a map, given the sorting decisions of the participants. All MDS analyses were conducted using XLSTAT using a Kruskal's stress (1) measure. (Note: MDS maps will be provided in online appendices in final electronic version).

For example, in the MDS results for Question 1, participants discriminated Image-2 and Image 4 (Figure 5.3) from each other (Kruskal's stress (1) = .007). This makes sense as Image-2 scene has a bookcase that is *disjoint* to the right wall and is front projecting in comparison to Image-4, which has a bookcase in a *contact* relationship with the left wall and is left projecting. On the 2D map (Figure 5.4), they are diametrically opposed. In the initial data set, participants significantly grouped/labeled Image 2 as the bookcase has a weak *on contact* relation to the wall and Image-4 was grouped with a stronger *against* contact relation with the wall. In some cases, images may have similar average scores, but are not close in the 2D representation space, signifying that the participants' decisions about the groupings are sometimes opposed even if the data appears to have similar frequency scores. This may be explained by some differences in the room context attributes, which may be used for grouping by some participants and not by some others.

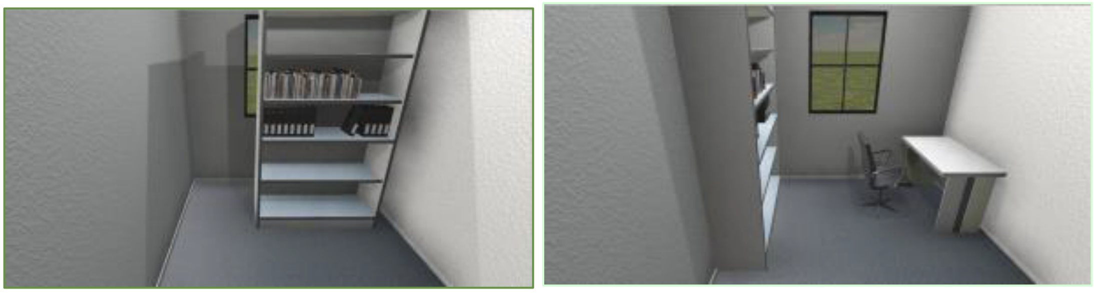


Image 2 Image 4

Figure 5.3 Experiment 2 images: *The bookcase is_____ the wall.*

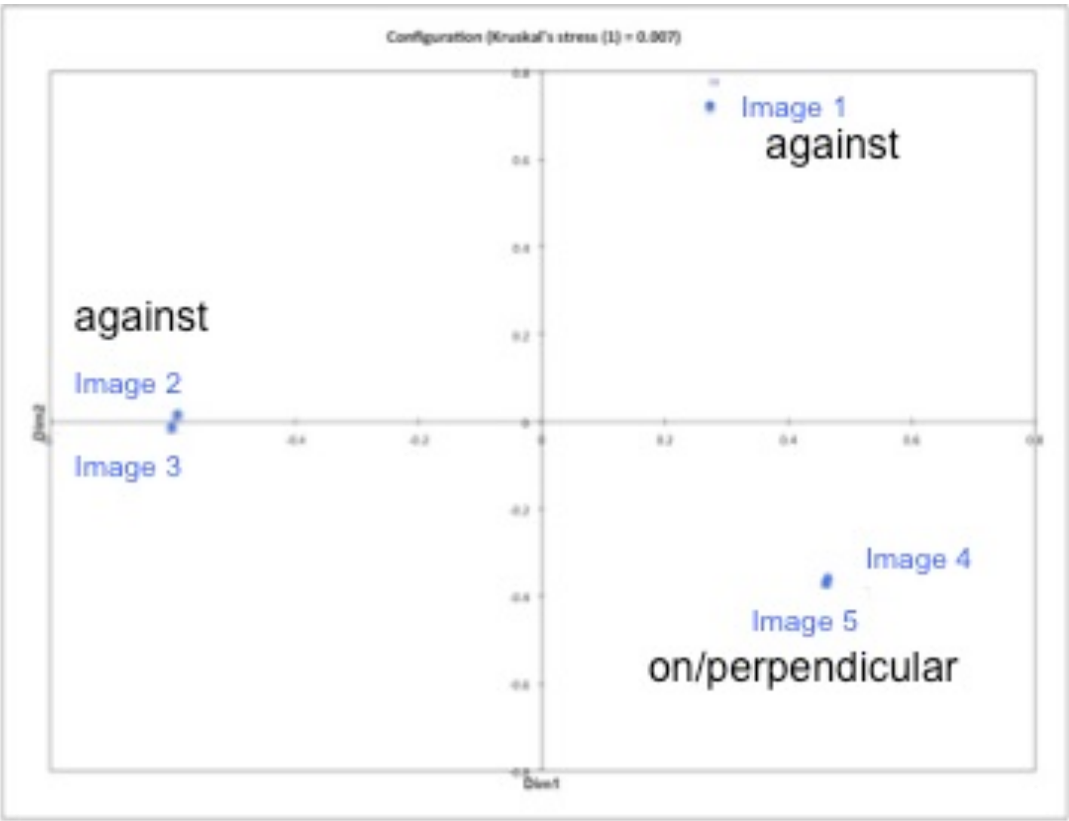


Figure 5.4: *Example of MDS Configuration Map: Q1, Images 1-5.*

The open sort image set results map of the individual sets provide information about how the images were grouped with more than just the category labels they were associated with. The open sort results are consistent with the open response prompt results in that the spatial preposition *against* was chosen as the category label for images with object-structure *contact* relations and *on* was chosen for images with structure-structure relations at statistically significant levels. However, *disjoint* relations between object-structures showed inconsistencies in labeling responses with the spatial preposition *facing* being numerically chosen the most frequently but not at statistically significant levels. This choice of *facing* may indicate some participants' emphasis on orientation over topological properties in *disjoint* relations. In the next version of the sorting experiments, the closed sort method provides the grouping labels in order to isolate factors driving participant grouping strategies even further.

5.2.3. Closed Sorting Classification

In the closed sort classification experiment, participants were asked to sort five images into four boxes with pre-determined classification labels (*on, against, along* ,and *near*). The spatial preposition labels were selected based on high frequency terms emerging from the analysis of relationships between objects and structures (Chapter 3). Images depicted eight items with *contact* relations between objects and structures (Table 5.10). In this task, *against* was chosen most frequently for *contact* OS relations (range = 51%-75%). A chi-square test was performed to determine whether any image was associated with one spatial preposition over other possible choices. Preference for *against* was not equally distributed in the population and was found to be statistically significant for all eight items ([X^2 range 4, N = 90] 48.93 to 120.40, p<.001). While both *on* and *along* were also chosen frequently as labels for the relationships (*on*: 10%-26%, *along*:10%-37%), a chi-square test determined neither of these terms were associated with a single image category at a significant probability level (*p*<.05).

Table 5.10: Frequency of Spatial Prepositions: *contact* OS relations

	Term Frequency and Percentage of Use				
	1	2	3	4	Other
Q6 Image 2	Against 67%**	On 13%	Along 12%	Near 6%	--
Q6 Image 4	Against 72%**	On 15%	Along 10%	Near 3%	--
Q7 Image 2	Against 75%**	Along 13%	On 11%	Near 1%	--
Q7 Image 3	Against 69%**	Along 13%	On 14%	Near 4%	--
Q7 Image 4	Against 67%**	Along 17%	On 12%	Near 2%	--
Q7 Image 5	Against 54%**	On 26%	Along 16%	Near 4%	--
Q8 Image 1	Against 58%**	Along 17%	Near 15%	On 10%	--
Q8 Image 2	Against 51%**	Along 37%	Near 9%	On 2%	--
Q8 Image 3	Against 65%**	Along 27%	On 5%	Near 0	--
Q8 Image 4	Against 62%**	Along 16%	On 16%	Near 6%	--

* sig. p<.05 **sig. p< .01 level *** sig. p< .001

For the five items representing structures with a *disjoint* relation in very close proximity with an object ‘The window ____ the tables.” (Table 5.11), the *near* category was chosen most frequently (range = 53%-89%) and was statistically significant for all five items (X^2 range = 80.88 to 195.95, *p*<.001)). The other three spatial preposition categories (*on, against, along*) for this relation were chosen infrequently by participants (all other terms range = 3%-28%).

Table 5.11: Frequency of Spatial Prepositions: *disjoin*t SO relations

	Term Frequency and Percentage of Use				
	1	2	3	4	Other
Q9 Image 1	Near 70%**	Along 11%	Against 13%	On 6%	--
Q9 Image 2	Near 53%*	Against 28%	Along 12%	On 6%	--
Q9 Image 3	Near 81%**	Against 8%	Along 9%	On 2%	--
Q9 Image 4	Near 82%**	Against 8%	Along 7%	On 2%	--
Q9 Image 5	Near 89%**	Against 3%	Along 3%	On 4%	--

* sig. p<.05 **sig. p< .01 level *** sig. p< .001

There were five items with *disjoint* relations for objects and room structures in the forced sort categorization task (Table 5.12). Unlike the free sort task for *disjoint* relations, participants chose *near* more frequently to label *disjoint* relations for four out of the five items (range = 37%-97%). A chi-square test was performed to determine if any term was more likely to be associated with that image. These results found the use of the term *near* was statistically significant (X^2 range ((4, N = 90) = 23.15 to 238.97, *p*<.001)). The term *along* was associated with the remaining image at a statistically significant level (p<.05). In images with a *disjoint* relation between an object and an object, *near* was the spatial preposition most strongly associated with this type of spatial relation (Table 5.13).

Table 5.12: Frequency of spatial prepositions: *disjoint* OS relations

	Term Frequency and Percentage of Use				
	1	2	3	4	Other
Q6 Image 1	Near 97%**	Along 3%	Against 0	On 0	--
Q6 Image 3	Near 40%*	Along 34%	Against 17%	On 9%	--
Q6 Image 5	Along 42%*	Against 37%	Near 12%	On 7%	--
Q7 Image 1	Near 92%**	Along 5%	Against 2%	On 0	--
Q8 Image 5	Near 67%**	Against 15%	Along 12%	On 5%	--

* sig. p<.05 **sig. p< .01 level *** sig. p< .00

Table 5.13: Frequency of spatial prepositions: *disjoint* OO relations

	Term Frequency and Percentage of Use				
	1	2	3	4	Other
Q10 Image 2	Near 51%*	Against 30%	Along 13%	On 6%	--
Q10 Image 3	Near 70%**	Along 19%	Against 6%	On 6%	--
Q10 Image 5	Near 80%**	Along 10%	Against 9%	On 1%	--

* sig. p<.05 **sig. p< .01 level *** sig. p< .001

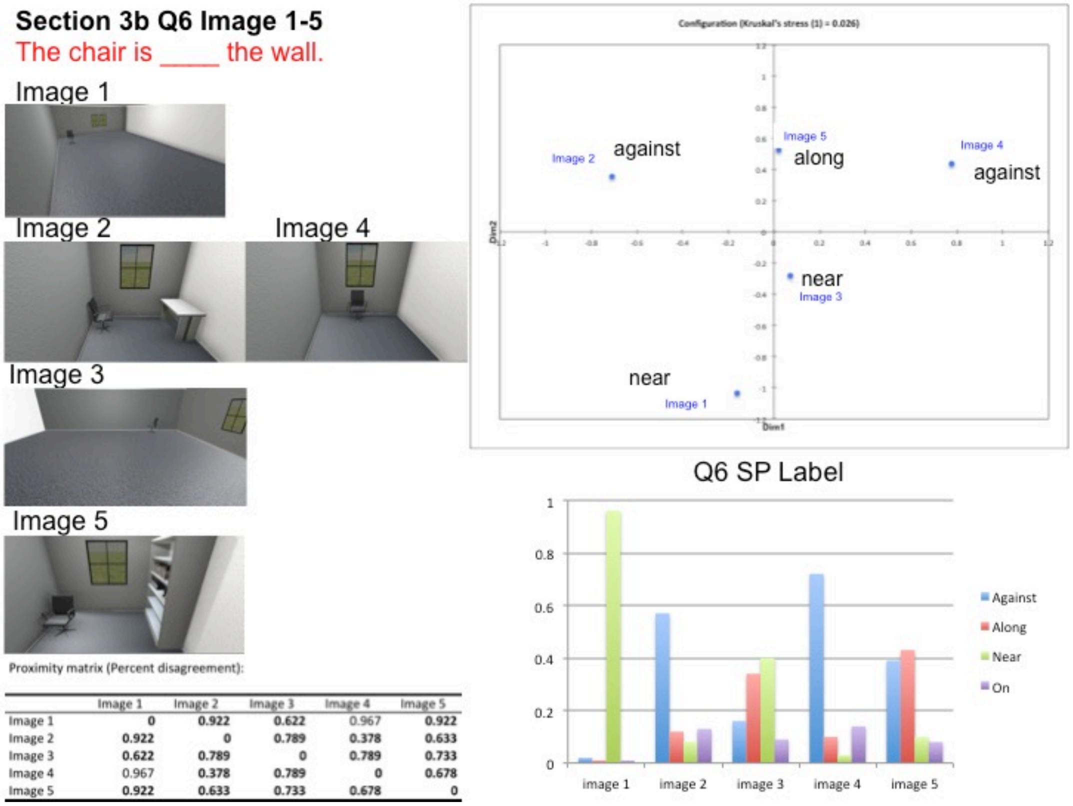
Which spatial relations are associated with which spatial prepositions?

Based on the results of the descriptive statistics for the closed sort task, it would appear that the ten images with a *contact* relation between an object and a room structure are most strongly associated with the term *agains*t. For *disjoint* relationships between objects and structures (OS) and image prompts with structure object (SO) or (OO) relationships, *near* is the spatial preposition most strongly associated with these types of spatial relations. In the few cases of *contact* relations with object-object image prompts *against* was chosen, but these associations did not reach the same levels (p<.05) as the OS image prompts indicating some level of uncertainty. Based on the results, this suggests a strong preference for using very specific terms for contact relations (*against* and *on*) and *disjoint* relations (*near*) between objects in indoor scenes. In addition to providing guidance about the length and format of a concise and correct spatial description, the actual terms that can be used to convey these spatial relations are emerging from these open and closed response experiments even without directly asking participants what terms they prefer.

5.2.4. Closed Sort Image Proximity and Preposition Categories

Similar to the open sort data, the closed set results were evaluated to better understand the connections between the image prompts in each of the five sets and a dissimilarity matrix was constructed in XLSTAT for each of the five sets questions (n = 25 image prompts). Because all the sorting categories were the same, questions in the five different sets could be evaluated for similarity/dissimilarity in sorting patterns relative to each other, across all 90 participants. The five sets with their full dissimilarity matrices and MDS results are provided in Appendix B. Some interesting sorting patterns emerged from individual sets and the total question analysis and are discussed here. For example in Q6 Images 1-5 (Figure 5.5), there is a large disparity in how Images-1, Image -2 and Image-5 are sorted

with Image-1 being classified as 'near' by almost all of the 90 participants in comparison to Image-3 which was also classified as *near* but did not reach a level of statistical significance using a chi-square test.



Proximity matrix (Percent disagreement):

	Image 1	Image 2	Image 3	Image 4	Image 5
Image 1	0	0.922	0.622	0.967	0.922
Image 2	0.922	0	0.789	0.378	0.633
Image 3	0.622	0.789	0	0.789	0.733
Image 4	0.967	0.378	0.789	0	0.678
Image 5	0.922	0.633	0.733	0.678	0

Figure 5.5: *Example of Closed Sort MDS analysis and output*

Furthermore, Q6 Image-5 is sorted into the *along* category. This image is perceived to be very different from any other image in this set and in the whole set of 25 images as demonstrated by the overall MDS configuration map (Figure 5.6).

As expected, the configuration map in 2D space shows images sorted into the *against* category were classified in a similar manner for images with *contact* relations and images with *disjoin*t relations were similarly sorted into the *near* category.

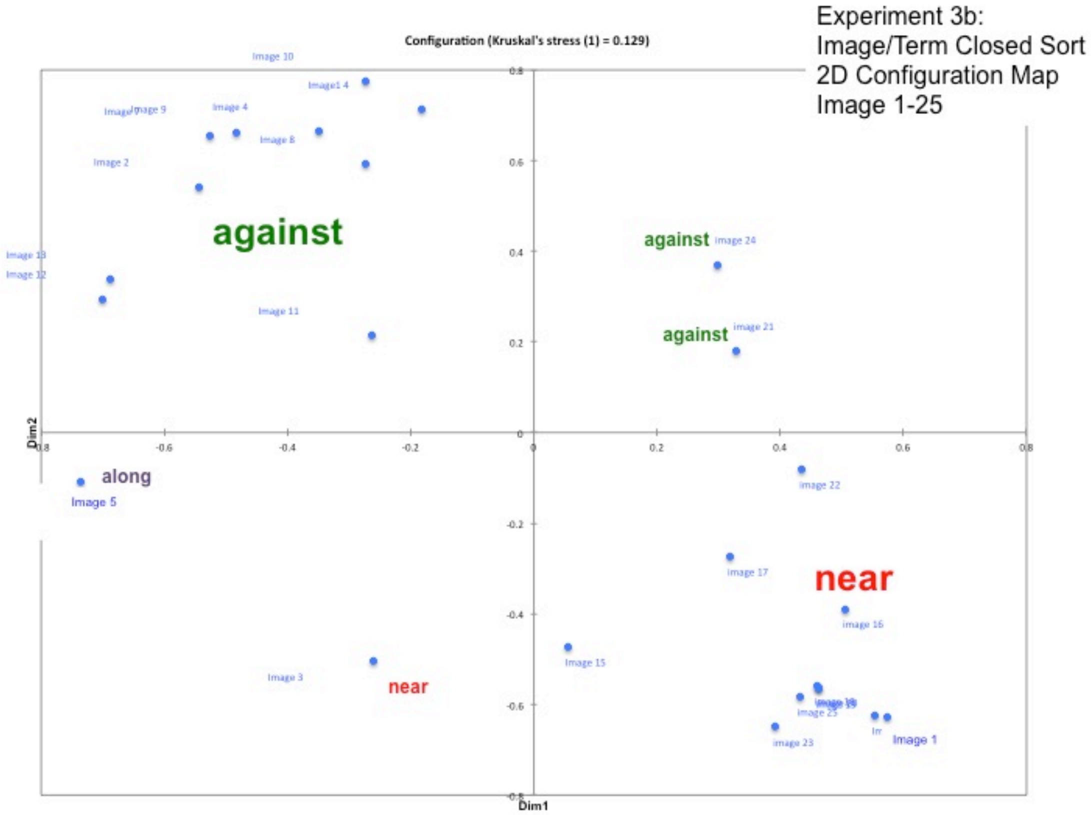


Figure 5.6: *MDS Scale Results Configuration Map*

There were a few other non-typical image results. In Q10, Image-1 and Image-4 (Figure 5.7) were categorized as weak *against* for object-object *contact* relation (bookcase and desk) with almost as many participants classifying this same pair of images as a *near disjoint* relation. This level of disagreement over how to classify the images can signal a strong degree of uncertainty among the entire group conceptualizing the spatial relation between the objects (*contact* or *disjoint*) since the only difference in the two images is the location placement of the bookcase in relationship to the desk. Room size, object orientation and distance from both the observer and the objects remained the same in both Image-1 and Image-4.

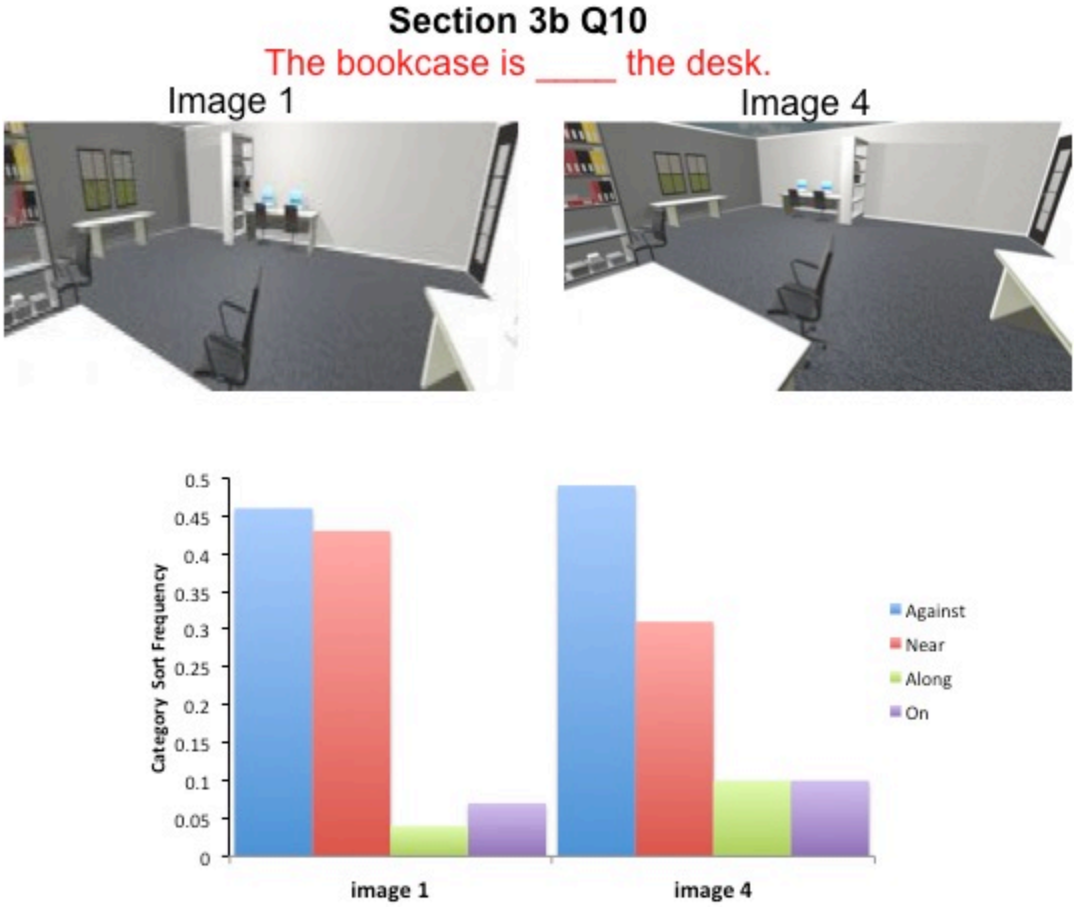


Figure 5.7: *Example of similar images with group sorting uncertainty.*

Which spatial relations had the least/greatest variation (i.e., uncertainty/disagreement)?

Based on the results of both open and closed image prompt sorting experiments, which spatial relations had the least or greatest level of variation in participant classification responses (i.e., collective uncertainty)? The results of the MDS method allows for the mapping of the image prompts that have been sorted and classified by the participants and facilitates a richer interpretation of the sorting tasks than a summary of descriptive statistics and chi. sq. tests provide. The Shepard diagram generated as a part of the MDS analyses, illustrates some patterns that go beyond just which spatial prepositions were used to classify the spatial relationships conveyed in the images. The comparative table contains three sets of measurements that correspond to three rankings for every pair of 25 images (n = 300 pairs) and the Shepard diagram provides a visualization of the quality of the representation (Figure 5.8). The Shepard diagram corresponds to a scatter plot, where the x-coordinates are the observed dissimilarities, and the y-coordinates are the distance on the configuration generated by the MDS. The disparities are also displayed. The more the points are spread, the less reliable the MDS map. If the ranking of the coordinate pairs is respected, then the MDS is considered reliable; the closer the points are on the same line, the more reliable the MDS mapping. For the total set, the Kruskal stress (1) was 0.129, indicating a quality 2D mapping of the images with one another.

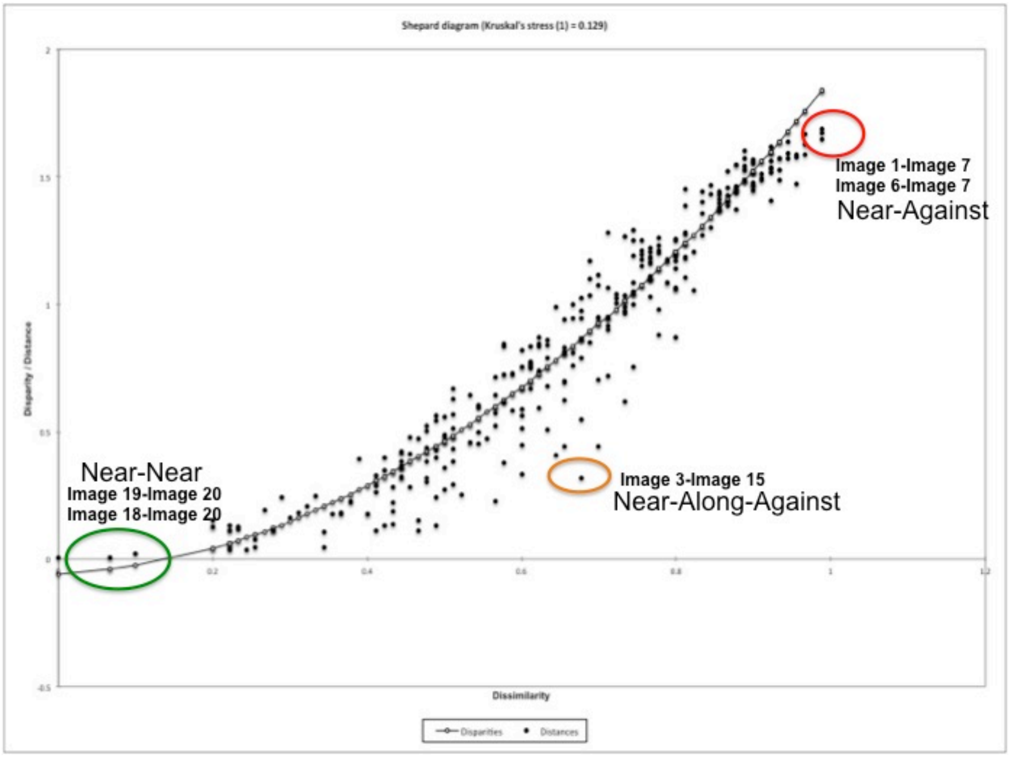


Figure 5.8: *Shepard diagram of dissimilarity coordinates of images pairs*

Examining the image pairs and their spatial relations highlighted on the Shepard diagram (Figure 5.8) helps to illustrate the relationships between participant classification decisions and the consistency in the entire closed sort data set (Figure 5.8). The image pairs at the lower end of the diagram have low levels of dissimilarity in classification and group disagreement (Images 18, 19, 20) as all images mapped to the prompt “The window is ___the table.” describing a structure-object relationship. Although Image-20 has the additional distractor object in the room (bookcase), all image pairs are strongly associated with the spatial preposition *near* by a highly significant proportion of participants. On the opposite end of the diagram, there are images that have a high level of dissimilarity, or participants classified these images in different categories (*near* and *against*) with a high level of participant agreement. Q1 Image-1 and Q2 Image-7 illustrate this type of dissimilarity. Both are classified as strongly either *near*-d*isjoint* (chair and wall) or *against-contact* (bookcase and wall) by a high proportion of participants. This suggests participants are able to distinguish the images with *disjoint* relations as most similar and were able to identify images with clear *contact* and *disjoint* relations as the most dissimilar (Figure 5.9). This provides insight as to the relative accuracy of image classification based on these general relations between objects and the spatial prepositions used to represent them.

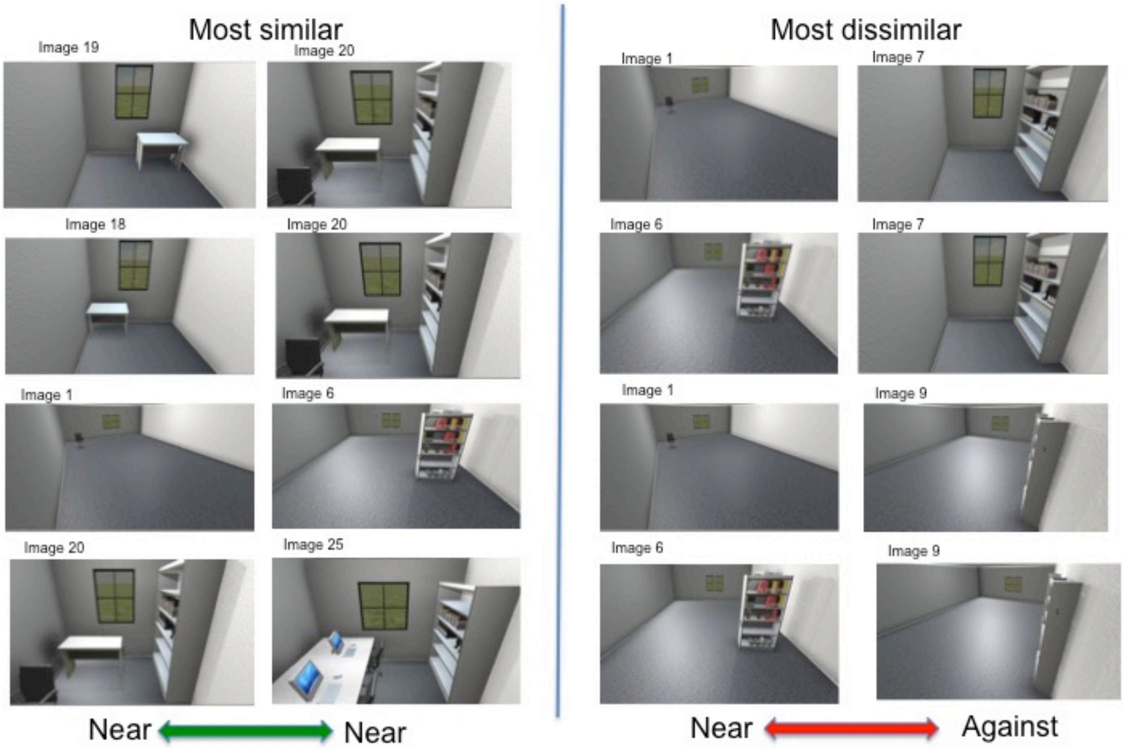


Image Pair	Relation	FP	Near
Image 19 Image 20 Image 18-Image 20	disjoint	SO	Window-table
Image 1 – Image 6	disjoint	OS	Chair-wall/Bookcase-wall
Image 18- Image 25 Image 20- Image 25	disjoint	SO/OO	Window-table/Bookcase-desk

Image Pair	Relation	FP	Near/Against
Image 1- Image 7 Image 1 – Image 9	Disjoint/Meets	OS	Chair-wall/Bookcase-wall
Image 6 – Image 7 Image 6 – Image 9	Disjoint/Meets	OS	Bookcase-wall

Figure 5.9: *Comparison of MDS dissimilarity of closed sort images pairs.*

The next section describes the final experiment that examined spatial preposition similarity, clarity and preference based on a set of six images and 15 spatial terms that could be used to describe the spatial relationships in the indoor scenes.

5.3. Experiment 3 Results: Indoor Image Comparison and Preference Ranking

Participants were presented with six indoor images and prompts. In the first set, they were asked to rank the similarity of a specific spatial preposition to fifteen spatial prepositions, including the preposition used in the prompt given the image context (e.g., 'The bookcase is on the wall.'). Next, they were asked to rank how clearly each spatial preposition described the spatial relation in the image. Finally, they were asked to consider all fifteen spatial prepositions and rank order their preference of these spatial prepositions to describe the image. These data were analyzed separately based on the represented spatial relation and feature type using descriptive statistics, chi-square test, a Friedman test and a Wilcox signed ranks test. The Friedman test is appropriate if the dataset meets four assumptions:

Assumption #1: One group that is measured on repeated measures.

Assumption #2: Group is a random sample from the population.

Assumption #3: Dependent variable is measured at the ordinal/continuous level.

Assumption #4: Samples do not need to be normally distributed.

As all of these assumptions are met with the preference data, the data were recoded from 1 (top preferred choice) to 7 (least preferred choice) instead of 7 (top preferred choice) to 1 (least preferred choice) for improved interpretation of results. In cases of spatial preposition terms in the preference set that were not chosen, they were given values of zeros. There were only six images and prompts evaluated in three different ways (similarity, clarity and preference), the tables below report the similarity and clarity results for the items coded for *contact* OS relations and then report the results of the Friedman and Wilcoxon signed ranks test evaluating if there were significant differences in preference for the fifteen given terms for each image prompt.

5.3.1. Similarity, Clarity and Preference: Object *Contact* Relations

Previous experiments found a number of patterns for *contact* relations amongst object-structure feature pairs. In the open response format in Experiment-1, the terms *on* and *against* were most frequently chosen to describe *contact* relations between objects and room structures in the prompts. In Experiment-2, *against* was also chosen most frequently to group and label these types of *contact* relations. Finally, in Experiment-3, a set of spatial prepositions were evaluated for similarity, clarity and preference in comparison to one another and the same types of patterns were observed in these results as in the earlier experiments (Table 5.14).

Table 5.14: Spatial preposition preference: *contact* OS relations

	Similarity	Clarity	Preference
Q 1 The bookcase __ the wall.	Against/On***	Against*** On**	Against ***
Q4 The table ___ the wall.	Against/Touching	Against *** Touching	Against***
Q 6 The desks ___ the window.	Along/By**	Along***	Along ***

* sig. p<.05 **sig. p< .01 level *** sig. p< .001

In cases of an object in a *contact* relation with a wall, there was a statistically significant difference in perceived similarity and clarity of the terms. For example, in Question 1 both terms *against* and *touching* were evaluated to be statistically significant in similarity to *on* when describing the relationships between the bookcase and the wall (*against*: X^2 [16] = 68.48, p ≤ .001; *touching*: X^2 [16]= 61.55, p ≤ .001). In terms of clarity, however, 'against' and 'on' were the only terms to reach a statistically significant level of clarity over the other terms (*on*: X^2 [16]= 104.22, p ≤. 001; *against*: X^2 [16]= 85.02, p = ≤. .001). In both MDS maps (similarity and clarity) these terms cluster closely together (Figures 5.10 and 5.11). This suggests that participants found *against* and *touching* identical or very similar to *on* as a term to correctly describe the same types of contact relations between objects. However, when it came to clarity, the term *touching,* although similar, was not judged to be as clear a term as were the terms *against* and *on* for these *contact* relations between objects in indoor scenes.

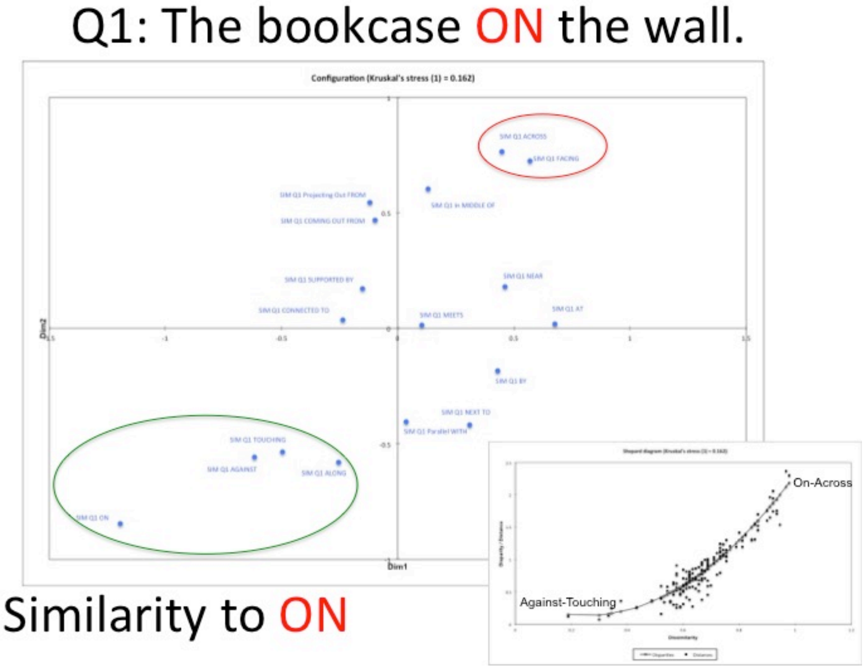


Figure 5.10: *MDS map of similar terms for Question 1*

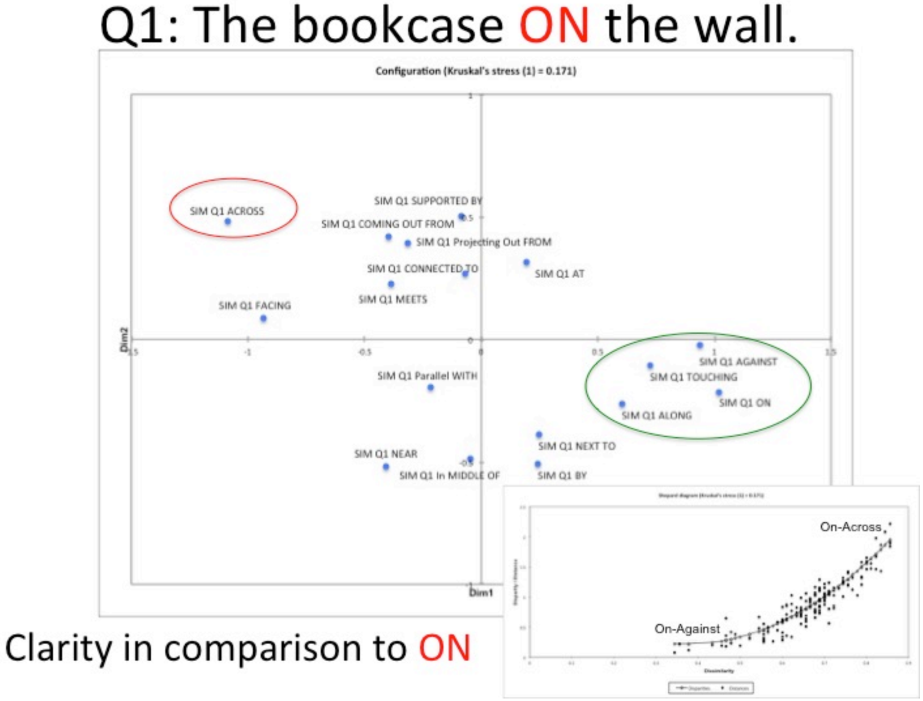


Figure 5.11: *MDS map of clarity terms for Question 1*

There was a significant difference in term preference based on the Freidman test (X^2 = (5) = 37.462, p < .001). A post hoc analysis with Wilcoxon signed rank tests (Bonferroni correction = p<.005) was conducted. There was a statistically significant preference for using *against* instead of *along*, *next to* or *touching*. However, there was no statistically significant difference in preference in using *against* versus *on* to communicate a *contact* relation (Z = -2.773, p = .006). Furthermore, *on* was not preferred at statistically significant difference levels over the other highly rated terms *along, next to*, and *touching*. Therefore, in ranking the preference of spatial terms for the *contact* relation between objects, *against* was the most highly preferred term. Although a similar term, *on*, was evaluated to be highly similar and just as clear a term in comparison to *against*, the term *against* was ranked to be the most preferred term to describe the contact relation between objects in an indoor scene. This would suggest that both *against* and *on* can be thought of as spatial synonyms for *contact* relations, and could be used interchangeably in spatial expressions, conveying similar levels of spatial information about the contact relation between objects in the descriptions of indoor scenes. So in addition to the indirect evidence of use of spatial prepositions for contact relations between objects in indoor scenes, there is more evidence regarding prepositions that are judged to be significantly similar and clear enough to be used interchangeably to represent the same *contact* relation between objects. These results also suggest that the term *against* is the first choice of term that a system for scene descriptions should use for contact relations between objects in a simple indoor setting.

This pattern was also observed for the image prompt: “The table is _______the wall.” In this prompt, *against* was evaluated to be most similar to the terms *touching* and *along* (touching:X^2=[16] =56.31, p = <.001) (along: X^2 =[16] = 22.08, p = <.001) when describing the relation between the table and the wall. In terms of clarity, however, *against* was the only term to reach a statistically significant level of clarity over other terms (*against*: X^2=[16]= 165.82, p = <.001). There was a statistically significant difference in preference mean rank of the spatial preposition term for Q4 image prompt “The table ____the wall.” (X^2 = [16] = 317.532, p <.001) with *against* being the most preferred term to describe this contact relation over all of the other possible terms. Post hoc analysis with Wilcoxon signed rank test (Bonferroni p<.002) confirmed a statistically significant preference of using *against* to describe the relation over along, near, touching, and by to describe the image prompt *contact* relation between the table and the wall.

There was also a statistically significant difference in perceived similarity and clarity of the terms for *contact* relations in Q6, “The desks are ________the window.” The term *along* was evaluated to be most similar to the terms *by* and *next to* (*by*: X^2=(16)= 29.28, p = <.001; *next to*: X^2=(16)= 21.82, p = <0.001) when describing the *contact* relation between the bookcase and the wall. In terms of clarity, however, *against* was the only term to reach a statistically significant level of clarity over the other terms (*against*: X^2=(16)= 41.28, p = <0.001) There was a statistically significant difference in preference mean ranks of the spatial preposition term for Q6 (X^2 = (16) = 220.201, p = <0.001) with *along* being the most preferred term to describe this spatial relationship over all of the other possible terms. Results of the post hoc analysis with Wilcoxon signed rank tests (Bonferroni correction = p<.002) found there was a statistically significant preference of using *along* to describe the relationship over *against, by, near, next to, on,* or *facing.*

5.3.2. Similarity, Clarity, and Preference: *partof* relations for structure-structure feature pairs

Consistent with response patterns in Experiments-1 and Experiment-2, Q5 provided an image prompt with a *partof* relationship between two room structures (e.g., window and wall). The spatial term *on* was ranked at statistically significant levels of similarity to the prompt term *at* and *on* was also ranked as a statistically significant term for clarity in addition to *at* and *along* (Table 5.15).

Table 5.15: Spatial preposition preference: *partof* SS relations

	Similarity	Clarity	Preference
Q 5 The window ___ the wall.	On/At ***	On***	On***

There was also a statistically significant difference in preference mean ranks of the spatial preposition term *on* for this same image prompt (X^2 = (16) = 187.252, p = <.001). Post hoc analysis with Wilcoxon signed rank tests was conducted (Bonferroni correction = p<.005. The term *on* was preferred to describe the relationship of the window and the wall over the highly ranked terms *in the middle of*, *connected to,* and *supported by*. There was also a statistically significant difference in preference for using *in the middle of* over the terms such as *connected to* (Z = -3.093, p = .002), and supported by (Z = -3.051, p = .002)' to describe the *partof* relationship between the window and the wall.

5.3.3. Similarity, Clarity, and Preference: *disjoint* relations

For images with *disjoint* relationships between objects and structures, consistent with the earlier experiment results, *near* and *next to* were the only statistically significant terms for similarity, clarity and preference (p<.01). Both terms were observed to be statistically significant in their similarity unlike the less specific *by* to describe a *disjoint* relation between objects and structures (Table 5.16).

Table 5.16: Spatial Preposition Preference: *disjoint* OS relations

	Similarity	Clarity	Pref. Sig.
Q 2 The desk ___ the wall.	Near/By***	Next to ** Near **	Next to**
Q 3 The chair ____ the wall.	Next to/Near**	Next to Near	Near**

There was a statistically significant difference in preference mean ranks of the spatial preposition term for Question 2 (X^2 = (16) = 186.410, p = <.001) with the term *next to* being the most preferred term to describe a *disjoint* relation over all of the other possible terms. Results of the post hoc analysis with Wilcoxon signed rank tests (Bonferroni correction = p<.003) confirmed there was only one term in which *next to* had a statistically significant difference in preference for describing the relationship, which was the term *touching* (Z = -3.476, p = 0.001). This can be interpreted to mean there was no more preference for *next to* than the other more highly ranked terms. This could signal more uncertainty in the terms applied to *disjoint* relations with these types of objects.

In Question 3, there was a statistically significant difference in preference mean ranks for this image prompt (X^2 = (16) = 474.393, p = <.001) with near being the most preferred term to describe this *disjoint* relation over all of the other possible terms. Results of the post hoc analysis with Wilcoxon signed rank tests (Bonferroni correction = p<.003) found there was a statistically significant preference for *near* to describe the relationship over all other terms including the closest preferred terms, *next to* (Z = -3.447, p = .001) and *by* (Z = -3.602, p = <.001).

5.4. Room Context Impact

Based on the consistency in the use patterns of spatial prepositions observed across the three experiment formats (open response, classification, ranking), a final set of analyses were conducted to investigate a set of dependent variables (Room Size, Feature Pair, Orientation, and Distance) and their impact on scene descriptions. We conducted dependent samples t-tests across Experiment-2 closed sort items to determine if there was a statistically significant difference in the mean of participants' sorting of images into spatial preposition categories based on room size, feature type, orientation, and distance as well as the effect size of any difference.

5.4.1. Room Size

Results of t-test for dependent groups indicate a significant preference for *against* in both room sizes (Table 5.17). The term *against* was used more often in both large rooms (t (89) = -3.254, p<.01)) and small rooms (t (89) = -9.282, p<.01)) for OS *contact* relations with a moderate to large effect size difference (Cohen's d = -.609 (L) d = -1.695(S). Room size had no impact on use patterns of spatial prepositions for *partof* relations in SS settings, with *on* being chosen exclusively over *against* in all cases. Room size also had no impact on spatial preposition use for *disjoint* relations in OS settings, with *next to* and *near* being chosen most frequently, but not at a statistically significant level.

Table 5.17: Spatial preposition mean by room size

	M	SD
Small room *on*	.1670	.1748
Small room *against*	.4967	.2122
Large room *on*	.2431	.2728
Large room *against*	.4257	.3246

5.4.2. Feature Pairs

When comparing the use of *on* and *against* for *contact*, *partof* and *disjoint* relationship between feature pairs (OS, SS), there were mixed results. The term *against* was used more frequently than *on* in contact relation OS settings (Table 5.18). However, across all types of these questions there were no significant differences in the means. That is, although *against* was used more frequently, there was no statistical difference between the use of *on* and *against* to describe *contact* relations across all items in OS settings (t (89) = -1.352, p =0.180)). However, in SS settings *on* was used statistically significantly more than *against* to describe *partof* relations (t (89) = 17.336, p= <.001)). Finally, neither of the terms *on* nor *against* was used frequently to describe *disjoint* relations in OS settings (t (89) = .194, p=.847)).

Table 5.18.: Spatial preposition use mean by feature type

	M	SD
OS *on – contact*	.2514	.2659
OS *against -contact*	.3248	.3152
SS *on – part of*	.6911	.3673
SS *against – part of*	.0056	.0370
OS *on – disjoint*	.2067	.2406
OS *against – disjoint*	.1983	.2463

5.4.3. Orientation

When comparing the use of *on* and *against* for *contact* relations and orientation (Right, Left, or Front), we found that although both terms were used frequently there was no difference in their use in the front orientation condition (Table 5.19). However, there was a significant difference in the use of *against* in the right/left condition (t (89) = 3.590, p.001). As noted previously, *on* was chosen at a statistically significant level in every SS item and there were no statistically significant patterns in any of the *disjoint* relationship images, including by object orientation/alignment.

Table 5.19: Spatial preposition use mean by orientation for *contact* relations

	M	SD
Right/Left *on*	.2672	.3023
Right/Left *against*	.3653	.3703
Front *on*	.2148	.3288
Front *against*	.2407	.3121

5.4.4. Distance

When comparing use of *on* and *against* for *contact* relations and distance conditions (Table 5.20), there was a significant difference in the use of *against* in images with objects in the far distance condition as compared to the mid-distance condition (t (89) = 2.816, p.006). Distance did not have an impact on SS *partof* images nor *disjoint* OS conditions.

Table 5.20: Spatial Preposition use mean by distance type

	M	SD
Mid *on - contact*	.2630	.2773
Mid *against – contact*	.2907	.3014
Far *on - contact*	.2417	.2913
Far *against – contact*	.3537	.3531

5.5 Discussion

The purpose of this dissertation was to investigate if there were factors that may influence preposition choice in NL descriptions used to convey different types of spatial relations within indoor scenes. The overall hypothesis was that underspecified spatial prepositions such as *on* are used frequently in indoor scene descriptions and serve as oral short cuts for describing spatial relations between objects in indoor scenes. There were several major findings of the research. First, results across question types (i.e., open response, categorization, and ranking) provide strong evidence of preference for the use of *against* for the *contact* relation in almost all room context conditions (room size, orientation, distance) featuring object and structure feature pair relationships (OS). Even in the open response format, where there was a much wider variation of terms used to describe the *contact* relations, *against* and *on* were the most frequently chosen terms. Second, in the forced choice categorization task, against was strongly chosen as the preferred term over on (p<.01) for every item with a *contact* relation. Along with the similarity, clarity and preference rankings, these results suggest that while these two terms can be used interchangeably to represent *contact* relations between objects and structures within virtual indoor scenes, *against* is clearly the most preferred term. This finding held across room sizes (small and large), right/left object orientation and to some extent when objects and structures were a further distance from observer than in mid-distance images.

Therefore, the hypothesis that underspecified terms such as *on* may serve as a minimum specificity term for this relation is supported by the frequency with which *on* was chosen and the strength of its similarity, clarity and preference ranking in comparison to the most preferred term *against*. However, *on* was not confirmed as a statistically significant preferred term for contact relation. Instead, there was a statistically significant preference for *against* to describe these spatial relations. The implications of these findings are that in designing a flexible system for NL scene descriptions, *on* may be used as the minimum specificity term for *contact* relation between objects and structures, however, *against* appears to be the strongly preferred term to describe these spatial relations.

Another major finding is that the use of *on* was significantly preferred in all room context conditions featuring structure-structure (SS) *partof* relations (e.g., window, door, and wall). While this is consistent with the patterns observed within the analysis discussed in Chapter 3, the results imply a disconnect in how structures such as windows and doors were classified in this study as being *partof* within the wall structure. The strong preference of the use of *on* to describe these relations suggests some alternate interpretation of these relationships such as a *supports* relation rather than a *partof* or a *contact* relation. These results confirm our hypothesis that in descriptions with structures in a *partof* relation to other room boundary structures, *on* is the term with the minimum specificity (as opposed to *in*). Likewise, as the statistically significant preferred term, *on* should be used as the strongly preferred term to use in a NL indoor scene description framework to describe these types of structure-structure relations.

In settings with object-structure *disjoint* relationships, our hypothesis on the more frequent use of minimum specificity terms such as *on, at,* and *by* was not supported. Although both *on* and *by* were frequently used to describe OS *disjoint* relationships, this did not occur at statistically significant levels. Instead, terms such as *next to* and *near* were preferred at statistically significant levels (p<.01) for these types of *disjoint* relationships across all question formats. These results suggest in NL descriptions of indoor scenes with *disjoint* relations, there is a need for more specificity than elemental spatial prepositions can provide due to the uncertain nature of the spatial relationship.

Room context conditions appear to have less impact on spatial preposition choice than was expected with a few exceptions. For example, *against* was preferred over *on* in all OS settings in both small and large room sizes. The term *on* was preferred in SS settings in both small and large rooms, and there was no difference in the use of spatial prepositions in *disjoint* OS settings. The term *against* was the preferred spatial preposition in a right or left orientation in comparison to settings with objects in a front orientation. In addition, *against* was the preferred term used in far distance OS *contact* relations in comparison to mid-distance conditions.

Finally, the text analysis of structured prompt responses helped to identify additional room structures, such as *corner* and *middle* in the descriptions of object-structure relations within the indoor scenes, pointing to some utility in designating physical unbounded features within rooms. These implicit room structures may work as additional containment structures for objects when a clear *contact* relation was not discernable because of a *disjoint* relation between the object and structure in question.

5.6 Conclusions

This chapter provided details regarding the three experiments conducted to investigate patterns of spatial preposition use in indoor scenes. The experiments were based on patterns observed in the analysis of indoor scene description data in Chapter 3 and were designed to isolate key variables influencing spatial expressions by creating spatial images in a virtual reality environment. Despite the large variation of terms used in the open response prompts to describe spatial relationships in indoor scenes, there were consistent and statistically significant patterns in the terms people used to describe spatial concepts such as *contact, partof* and *disjoint* relations within the indoor scenes. The next chapter provides an expanded discussion of the implications of the findings and the application for their use in the design of an intelligent indoor scene description agent.

CHAPTER 6

DISCUSSION, IMPLICATIONS, AND APPLICATIONS

This chapter concludes the thesis. It provides an overall summary of the study investigating the alignment of spatial relations with natural language spatial expressions in indoor vista space. This thesis investigates the alignment of conceptual spatial relations and natural language (NL) semantics for *contact, disjoint*, and *partof* relations within indoor scene descriptions This chapter provides a synthesis of the analysis of indoor scene descriptions and the findings of the set of experiments designed to investigate this alignment. It provides a discussion of the research questions (Chapter 1) contextualized in relation to existing knowledge and theories about how spatial prepositions convey spatial information at different spatial scales.

1. *How do people conceptualize and communicate spatial relations when they describe an indoor scene in natural-language?*
2. *What spatial prepositions do people use to describe topological and conceptual relations between objects in a room?*
3. *What are preferred spatial prepositions to express spatial relations between objects in indoor scenes?*
4. *Do descriptions of indoor scenes differ based on sensory constraints of the intended recipient of the description?*
5. *What role does object function serve in the choice of spatial prepositions in the description of indoor scenes?*
6. *Are there differences in the preference of level of specificity in spatial prepositions used in scene descriptions?*

This work applied a Naive Geography approach to the alignment of conceptual spatial relations to NL spatial prepositions within vista scale space. It considered abstractions of spatial concepts and employed human-subject based experiments to test assumptions about how spatial relationships are conveyed in NL spatial expressions. While there is a large body of work using this approach at tabletop and geographic scales, there has been less work using this approach within indoor settings. The associated corpus development provides a valuable contribution to machine learning techniques on which to train the NL algorithms used to generate image captions.

For this dissertation research, it was necessary to return to earlier methodology to better understand some of the most basic questions about spatial relations in indoor space, such as 1) the types of entities and relations included in scene descriptions, 2) the ordering of entities and their importance to the entire description, 3) the spatial prepositions used to communicate spatial relationships, and 4) the similarity, clarity and preference of spatial prepositions within indoor vista scale scenes.

6.1 Discussion of research questions

A large body of research provides evidence that the ways in which humans communicate about space provides clues as to how multimodal sensory input helps to create a conceptual model of space (e.g., Miller & Johnson-Laird, 1976; Montello, 1993; Tversky, 1993, 2001; Tversky, 2009). In alignment with a Naïve Geography perspective, this dissertation research used both a cognitive and a linguistic approach to understanding the spatial prepositions used for spatial relations with a spatial behavior task (e.g., scene descriptions). Based on the indoor scene description analysis (Chapter 3) and the results of Experiments 1-3 (Chapter 5), there are some basic questions we can answer about the types of entities, the spatial relations and spatial prepositions used in indoor scene descriptions at a room size vista space scale.

6.1.1 Research question 1: Conceptualization and Communication of Indoor Scenes

How do people conceptualize and communicate spatial relations when they describe an indoor scene in natural-language?

Understanding how spatial relations are conceptualized and communicated in indoor scenes involves an examination of: (1) what objects are being related to one another; (2) what are the types of relations being conceptualized and communicated within the description of the spatial configuration. There were several sub-parts to this first research question. The hypothesis was that there would be no difference in frequency of use in the types of spatial prepositions used to describe relations in oral versus typed-text based descriptions.

What objects did participants relate to one another in descriptions of an indoor scene?

In the analysis of open scene descriptions, participants most frequently identified smaller, moveable objects (e.g., desk/table, file cabinet and bookcase) in relationships with larger, immoveable structure objects or regions (e.g., wall, side, room) as the primary entities in NL scene descriptions. These spatial triples consisted of an "object (trajector) + spatial preposition + structure (landmark)", although in many cases, there were other spatial triples used within in a single spatial utterance that linked the primary trajector and the landmark pair. This is an illustration of how additional reference objects are often used to create the topological link between the figure and the ground (e.g., the chair in the corner, in front of the larger chair; Herskovits, 1980). Unlike the open descriptions (Chapter 3), the open structured prompts (Experiment 1-Chapter 5) provided the target trajector and landmark objects. However, additional objects used in the open structured prompts do provide some additional insight into what types of objects were more frequently used to topologically link the targeted trajector and landmark objects provided in the prompt. While additional linking objects were found in only about half of all prompt responses, the objects that were used were most frequently room region areas (corner, side) and room structure objects (wall, window, and door). Few smaller room objects were used as additional topological links between the targeted trajector and landmark objects. This is important because it emphasizes the importance of the room structures, both physical objects such as walls and windows, as well as perceived abstract regions such as corners and sides of the room.

There was also a dominant trend in the open descriptions of participants relating objects to vertical structure objects (e.g., walls) rather than horizontal structure objects (e.g., floor or ceiling). Finally, participants most often communicated relations between objects using an *intrinsic* (rather than *absolute* or *relative*) frame of reference in the open scene descriptions.

What spatial relations were conceptualized and communicated in NL indoor scene descriptions?

In the open scene descriptions, participants used primarily *contact* and *qualitative proximity* relations, as well as a few other relations such as *contains*, *covers/covered* by (e.g., window in middle of the wall, chair pushed into desk). Participants also seemed to favor using underspecified spatial prepositions such as *on* and *in* in spatial expressions although the total variation of spatial prepositions and the level of spatial information detail used was broad. Overall, the scene description analysis found that the spatial preposition *on* was the most frequently used spatial indicator and was used primarily in the *contact* sense (e.g., TR [moveable object] *on* LM [structure]). There were few instances of *on* being used as a spatial relation in the *support* sense. This is consistent with previous research on the assignment of figure and ground dependencies where an object whose location is at question, the figure, most often precedes the preposition and the ground is typically larger and less mobile (Talmy, 1978).

Based on results of the scene description analysis, there are indications that conceptualization and communication about objects in indoor vista-scale spaces differ from both tabletop space and geographic space. Geographic space is interpreted as 2D space where horizontal and vertical dimensions are separated and the 3rd dimension is represented as an attribute (position) rather than an equal dimension (Mark, Egenhofer, 1994). Indoor space at the vista-scale, used in both the open scene descriptions and Experiments 1-3, seems to be interpreted as a 3D space, even in a virtual environment, except perhaps the case of structure objects such as windows and doors. The relation between the window and door types of structure objects and other structure objects, such as walls may be conceptualized in a similar way to 2D relations, as two flat surfaces in a *covers* relation. The scene descriptions seem to demonstrate a significant difference from tabletop space in reliance on moveable objects relationship with structure objects, illustrating the importance of the boundedness represented by the walls of the room and the hierarchical nature of the indoor environment (e.g., (room within building (object location within room)). This observation supports a hierarchical model of indoor space that can provide different levels of detail based on the context, user need, and desired spatial behavior task. This approach to representation may help to reduce some level of uncertainty in the topological configuration of objects and structures in indoor scenes using a prescribed set of spatial prepositions associated with semantic annotation data. The additional spatial information can enhance the representation accuracy of NL descriptions of indoor scene image datasets as well as descriptions of indoor spaces used for NL guides in public buildings. An example is guidelines to the length and structure of short descriptive expressions relating objects in an indoors scene. On average, there were approximately five nouns (e.g., objects/structures), two verbs and three prepositions used per utterance in the scene descriptions. This observation and similar results found in the experiments (Chapter 5) suggest a possible optimal length and structure for sentences describing spatial relationships within indoor scenes. Specifically, based on the findings of this dissertation, a concise spatial triple should take the following form and length: trajector (≤ 3 words) + spatial preposition or prepositional phrase (≤ 4 words) + landmark (≤ 3 words) = spatial triple (≤ 10 words).

6.1.2. Research Question 2: Use of Spatial Prepositions in Indoor Scenes

What spatial prepositions do people use to describe topological and conceptual relations between objects in a room?

For *contact* relations between room objects and structures, although the open response descriptions found that on was the preferred term for *contact* relation, there is strong evidence in Experiments 1 -3 for the preference in the use of *against* in almost all room context conditions (room size, orientation, distance) and across all question types (i.e., open response, sorting, and ranking). Responses to *contact* relation images showed less variation in the number of unique spatial relation terms used and a much larger variety of spatial prepositions recorded in the analysis of scene descriptions. It is important to note that *against* and *on* were chosen often at the same frequency levels to describe *contact* relationships between moveable objects and structure objects (OS). Other less frequent spatial prepositions used were *along, in front of,* and *touching*. but these terms usually did not achieve more than 15% of frequency response across *contact* relation items as well as across question format.

Patterns in the open response format showed a much wider variation of spatial prepositions used to describe the scene, however *against* and *on* were the most frequently chosen terms. The results of the experiments suggest that the two spatial prepositions can be used interchangeably to represent *contact* relations between objects in indoor vista- scale, although *against* is clearly preferred. The original hypothesis that the underspecified term *on* may serve as a minimum specificity term for the *contact* relation was supported by the frequency with which *on* was chosen in all *contact* relations and the strength of its similarity, clarity and preference rankings that directly aligned to the term *against*. The implication of this finding is that when designing an assistant for NL scene descriptions, the term *on* may be used as the minimum specificity term for *contact* relation between objects and structures, however, *against* should be the preferred term used to describe these spatial relations.

The spatial preposition *on* was significantly preferred in all room context conditions featuring structure-structure (SS) *partof* relations (e.g., window, door, and wall). The frequency for the use of *on* to describe the relationship between the window and the wall ranged from 45-66% in each item and was almost exclusively chosen at a statistically significant level (p<.01). There were very few other terms used for this type of relation the most frequent being *in the middle of*. Although this term did not reach a statistically significant level, it was chosen by over 40% of the participants who did not choose *on* for the same item.

While this is consistent with the patterns observed for the indoor scene description analysis, the results imply a disconnect in how these structures such as windows and doors were classified in this study as being *partof* the wall structure and perceived by the vast majority of participants. The strong preference among participants for the use of *on* to describe these relations suggests that embedded room structures like windows or doors within walls may be understood perhaps as a 2D *support* relation rather than *partof* by the participants. This pattern may be an example of viewing a particular object for a specific purpose, ignoring specific characteristics of the object (Talmy, 1978). Likewise, when considering the alignment of spatial concepts to NL spatial terms, the statistically significant preferred term, *on* can be considered an acceptable term to use in a NL indoor scene descriptions to describe these types of embedded structure-structure relations.

Spatial prepositions such as *next to* and *near* were preferred at statistically significant levels for all types of distinct *disjoint* relationships across all question formats. Descriptions of *disjoint* relationships between objects and structures experienced the greatest variation in spatial preposition use, with *next to* often being chosen to complete open response prompts, but never at a statistically significant level. Other terms used for *disjoint* relations were *near,* indicating a proximity/distance relation, or *to* (*the right/left*) indicating a directional relation. In a few cases, *against* was used. In these cases, the relation may have been perceived as a fuzzy boundary situation, although the trajectory object was clearly not in contact with the landmark object, it was perceived as 'close enough' to use *against*, a *contact* relation.

In settings with object-structure *disjoint* relations, the hypothesis about more frequent use of minimum specificity terms such as *on, at,* and *by* was not supported. Although both *on* and *by* were frequently used to describe OS *disjoint* relationships, this did not occur at statistically significant levels. These results suggest NL indoor scene descriptions for *disjoint* relations should consider the *distance* relation terms *near* and *next to* as preferred terms for *disjoint* relations, and this may indicate that there is a need for more specificity in *disjoint* relations than elemental spatial prepositions can provide due to the uncertain nature of the spatial relationship.

How similar is one spatial preposition in comparison to another for a given indoor scene?

In the similarity ranking task (Experiment 3), the terms *against* and *touching* were both ranked as similar to *on* at a statistically significant level. In the clarity rankings, both *on* and *against* were found to have the same statistically significant level to describe clarity of these term for contact relations. In MDS maps, these terms cluster together both in terms of similarity and clarity. For *partof* relations, the spatial term *on* was ranked at statistically significant levels of similarity only to the term *at* and was ranked as the most clear term for this type of relation. Likewise, for images with *disjoint* relations, *near* and *next to* were the only statistically significant terms for similarity and clarity.

When summarizing the patterns observed in the choice of spatial prepositions for the description of indoor scenes, there was evidence supporting Feist's (2000) attribute values of spatial scenes, in which the choice of spatial preposition conveys key pieces of spatial information such as a *contact/disjoint relation* between the figure and ground and the primacy of objects on the *vertical axis* as spatial references. The use of *on* for *partof* relations between room structure objects, such as windows and walls, also supports Feist's observation that choice of spatial prepositions conveys information about the *inclusion* of the figure by the ground, as well as the *nature of the support*, if any, provided to the figure by the ground.

There is evidence to support a spatial gradient of spatial prepositions based on *contact* and *support* sense for the prepositions *on* (Levinson, 2003). While the multiple semantics of *on* can be distinguished by the *support* sense and/or the *contact* sense, neither represents the use of *on* in a *partof* relation as was observed in window and wall relation. Based on Levinson's classification (Figure 6.1), the spatial preposition *in*, representing the *contains* or *inside* relations, is clearly separate from the preposition *on*, which is classified according to the *contact* and/or visual *support* in the figurative sense.

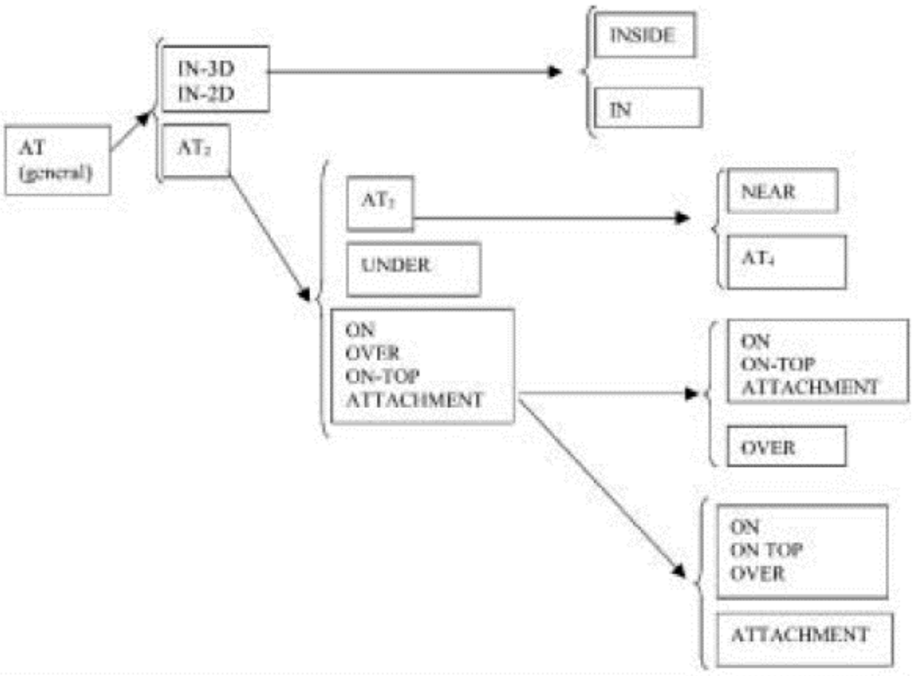


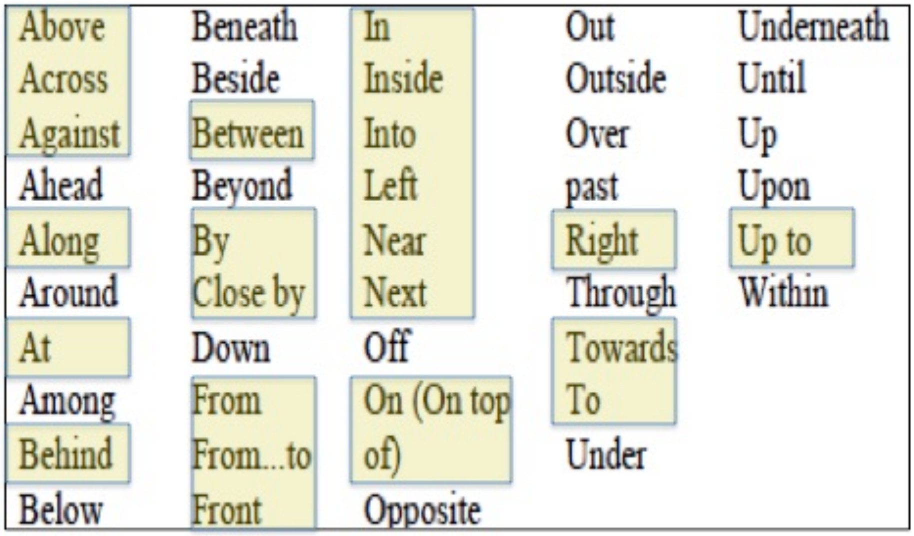
Figure 6.1: *Implicational scale of English spatial prepositions (Levinson, 2003).*

Levinson's (2003) classification does not permit the use of *on* in the *partof* sense and so the frequency level of scene descriptions using *on* to describe the relationship between a window and a wall suggests there is more to explore in these types of indoor relations. It is possible the vista-scale scene was limited or Levinson's gradient scales were using relation semantics extrapolated from tabletop and geographic space as a proxy for indoor space. However, the results may also point to the possible unique alignment of spatial relations and prepositions in indoor space that are not typically present at the other spatial scales.

6.1.3. Research Question 3: Preferences in Spatial Prepositions

What are preferred spatial prepositions to express spatial relations between objects in indoor scenes?

The results of this dissertation research on spatial preposition use in indoor scene descriptions contributes to the development of models for NL spatial expressions for indoor space. The results provide support for the refinement of the list of common prepositions used in object to object spatial reference expressions (in English). It is interesting to note that out of the 44 prepositions included in a recent NL model for indoor space proposed by Sithole and Zlatanova (2016), the results of this dissertation only provide support for the use of approximately 20 terms within scene descriptions and experiment results (highlighted terms-Figure 6.2). Furthermore, in the Sithole and Zlatanova models of indoor space, the spatial preposition *on* is only defined in its *support* sense, not the *contact* sense. The results of this dissertation provide strong evidence that the *contact* relation set of spatial prepositions should include *against* and *on* as primary terms for indoor scenes and *along*, and *touching* as secondary terms.



Above
Across
Against
Ahead
Along
Around
At
Among
Behind
Below
Beneath
Beside
Between
Beyond
By
Close by
Down
From
From...to
Front
In
Inside
Into
Left
Near
Next
Off
On (On top of)
Opposite
Out
Outside
Over
past
Right
Through
Towards
To
Under
Underneath
Until
Up
Upon
Up to
Within

Figure 6.2: *Spatial prepositions in model for indoor space (Sithole and Zlatanova, 2016).*

For *partof* relations between room structures, the results provide strong support for using *on* as the preferred term to describe spatial relationships such as between a window embedded within a wall. There was also a statistically significant difference in preference for using '*in the middle of*' to describe this same spatial relationship between two structure objects. Although a more spatially intuitive term, '*in the middle of'* is both conceptually and semantically different, however, the two terms are used for the same relation but appear to communicate two different types of topological relations.

The preferred terms for *disjoint* relations between objects and structures support systems using terms such as *near* and *next to* interchangeably over other possible terms. While the terms *by* and *to (right/left)* can be considered similar alternatives for *disjoint* relations between object and structure pairs, the use of the more vague proximity term *by* or direction term *to* was not strongly supported by the results of the analyses.

Table 6.1: Sets of preferred spatial prepositions for target relations

	Preferred term	Similar/Clear term	Alternate terms
Contact OS	Against	On	Touching, Along
Partof SS	On	--------------------	In the middle of
Disjoint OS/OO	Near/Next to	By	To (right/left)

6.1.4 Research Question 4: Sensory Constraints and the Intended Recipient of a Scene Description

Do descriptions of indoor scenes differ based on sensory constraints of the intended recipient of the description?

The results of this dissertation suggest there were no statistically significant differences in frequency of terms used for *contact, disjoint* or *partof* relations for the hypothetical intended users. In addition, there were no statistically significant differences in the mean number of words used to complete the prompt nor the number of words used to complete the oral and typed-text formats. Although participants in the scene descriptions were given explicit directions to create a description for someone who could not directly view the scene, most of the descriptions used underspecified spatial prepositions such as *on* and *in* with a high frequency. The results suggest there seemed to be little awareness that these terms might contribute to uncertainty and produce ambiguous spatial semantics for a person who could not directly view the scene or the scene image. This outcome is particularly important in the potential problems in the practice of the use of general training sets for neural networks that are created from crowd-sourced descriptions of object relations by sighted annotators. Similar to the room being modeled as a list of objects in a container without relations or context, these types of descriptions are not likely to be of practical use for users who are members of the BVI community. Additional research on which descriptive terms are the most effective or preferred in creating accurate scene descriptions for users in the BVI community will be part of plans to extend the work of this dissertation.

6.1.5. Research Question 5: Spatial Prepositions and Object Function in Indoor Scenes

What role does object function serve in the choice of spatial prepositions in the description of indoor scenes?

The analysis of both the indoor scene descriptions and the results of the open response prompt identified additional features, such as *corner* and the *middle* in the descriptions of the VE scenes. These concepts signaled evidence of functional features within bounded rooms that serve as types of containment structures for objects when a clear spatial relation with an explicit structure was not easily identified because of a *disjoint* relation with the object and structure in question.

While analyses of indoor scenes (Chapter 3) did point to the importance of structure objects in describing the spatial configuration of objects of indoor space, the more structured experiments (Chapter 5) provide further evidence that structure objects function to convey the *boundedness* of the space and these features are central to the description of indoor scenes. Structural objects serve a function as defining the edges of the space and the connected nature of the interior boundaries (left wall>far wall>right wall) serve a function as a description order strategy.

There were few explicit instances of spatial prepositions conveying functional spatial roles within descriptions of spatial configurations. In some cases, terms used to name objects pointed to implicit functional properties of objects and relations such as a noun choice of *map* versus *picture* suggesting a possible *activity use* function or *desk* versus table suggesting a *work/write* versus more general *activity use* function. Based on arguments for how functional attributes are conveyed through spatial prepositions and central to discerning context in scene descriptions (Vandeloise, 2006, Langacker, 2010), the lack of these types of contextual cues was surprising. It is possible that the indoor scenes being described did not contain enough variation in objects or the type of indoor setting (i.e., office workspace) was too generic.

The ordered networks (Chapter 3) provided evidence that descriptions moved in either a dominant near/far or far/near access pattern with right and left entities following (e.g., near right and near left). The network analysis that included structure object orders illustrated this distance-related description strategy over the ‘round-about’ description pattern observed in the original analysis of the scene descriptions (Kesavan, Giudice, 2012). Wall nodes in the network were primarily ordered in terms of connectivity from wall (left) to wall (far) to wall (right), suggesting some general rules for structuring scene descriptions and for a method of grouping objects and structures within descriptions of indoor scenes. For example, based on these results it would make sense to develop rules that group all objects in a *contact* relation with each of the walls and then deliver the description based on an order of near-left, near-right, far wall, and other moveable objects in the room that are not in contact with a structure object. This would only be the recommendation if a user has not specified an object of significant salience for the description or the user's spatial task is unknown.

6.1.6. Research Question 6: Impact of Context Factors on Preferred Spatial Prepositions

Are there differences in the preferences of level of specificity in spatial prepositions used in scene descriptions based on room context factors?

The analysis of scene descriptions (Chapter 3) demonstrated a preference for underspecified spatial prepositions (*on, in, by, at*) and while in the experiment responses (Chapter 5) these same terms did reach levels of statistical significance, they were not the preferred terms. Instead, when given a choice between minimally specified terms such as *on, in, at,* and *by* along with a list of spatial terms with an increasingly greater level of specificity such as *connected to and projecting out from*, the most preferred terms were moderately specified terms such as *against, near,* or *in front of.* These results were not impacted to a statistically significant level by any aspects of room context that were identified as potential factors for impacting the use of spatial prepositions. T-tests for dependent groups indicated that room size (small, large), orientation (right/left, front), and distance from observer (near, mid, far) did not impact the preference for the use of an underspecified term (*on-contact, by-disjoint*) over a term with more spatial information (*against-contact, next to-disjoint*). The only exceptions to this were the results for a statistically significant level of preference for the use of the underspecified term *on* to describe a spatial relation of a structure object (window/door) with another structure object (wall). However, there were no other feature pair types that had a statistically significant impact on the specificity preference of spatial prepositions. The unexpected patterns of descriptions of windows as they relate to walls in the indoor scenes across all response collection formats is an important area for future investigation to better explain this finding.

6.2. Limitations

All studies encounter some limitations and this research was no exception. Some problems were due to assumptions made in the design process such as not isolating the room context factors more fully in the image prompts. For example, when evaluating the impact of object orientation on spatial preposition choice within a bounded space, it appears that considering just the horizontal axis changes of the object (Figure 6.3) is not sufficiently constraining. Object height, in combination with directional placement, may have impacted prepositional choice more than anticipated. The comparison of a tall bookcase in a *contact* relation in a sorting task with a long set of desks also in a *contact* relation with a wall in a similar scene may have influenced the patterns of sorting responses and in labeling of the preferred spatial prepositions (Figure 6.4).

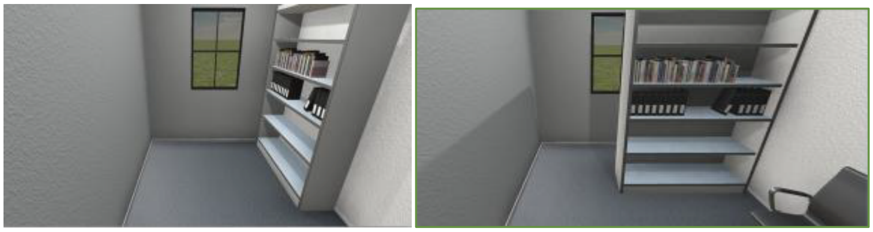


Figure 6.3: *Contact-relation Single Item*

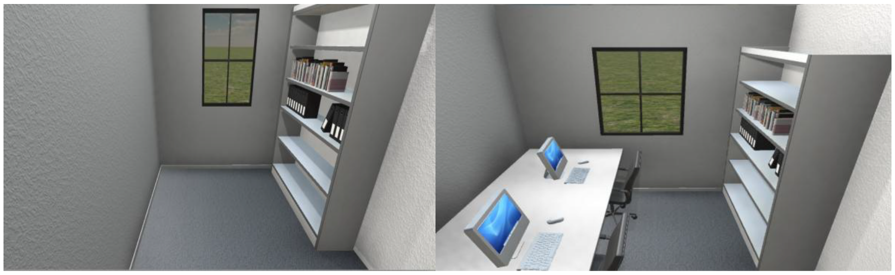


Figure 6.4: *Contact-relation Multiple Items.*

This large set of room context variables associated with the images made it difficult to create enough image prompts to run a factor analysis with an acceptable amount of reliability. Future studies will need to address a smaller number of room context variables for each spatial relation in order to determine if associations are statistically significant.

6.3 Conclusions

The goal of this research was to align NL specifications to effectively describe spatial concepts and relations within a simple indoor scene. *In particular, the thesis focused on identifying a controlled vocabulary of spatial prepositions for a small set of spatial prepositions to convey spatial relations between objects in indoor environments and to be used in automated scene descriptions.* The research questions, experimental design and methodology was grounded in the theoretical framework of Naive Geography (1995) which seeks to model spatial knowledge from a *common-sense perspective*. This set of theories is concerned with understanding space from the human user perspective, and uses human-subjects testing to better understand how people conceptualize and communicate about object relations in indoor scenes.

Based on the findings in this thesis, there is evidence that the perception and communication of indoor vista-scale space follows similar patterns identified in previous work in Naive Geography. For example, there was significant evidence in the open scene descriptions of variation in the frames of reference and level of spatial detail participants used to describe simple indoor scenes. The preference for underspecified spatial prepositions was a particularly significant pattern observed in the open scene descriptions as was the reliance on room structures (i.e., boundaries) as preferred reference objects in these spatial expressions. Based on the findings in the open scene descriptions, structured prompts were created to test observed patterns in the alignment of the spatial prepositions used in natural-language expressions describing simple indoor scenes. Previous work had investigated natural-language use in a variety of other spatial scales but this thesis is the first known research using this framework specifically in indoor vista-scale settings. A next logical step would be the design of similar experiments that allow for the comparison of the targeted spatial relations investigated in this thesis (contact, disjoint, part of) and the dominant prepositions used in indoor vista scale and at least one (or more) spatial scale. An immersive virtual testing environment would allow for similar variables to be tested and context to be highly controlled.

6.4. Directions for Future Research

6.4.1. Annotation of Spatial Property Graphs

Based on the original motivational problem scenarios, in order to generate correct and concise automated NL descriptions for indoor scenes, an intelligent system needs have the ability to:

1. collect spatial data from a variety of sources (e.g., computer vision, localization sensor networks and human input);
2. integrate collected heterogeneous spatial data with spatial reasoning structures;

3. use NL processing tools to synthesize and communicate relevant and accurate information about indoor environments
4. convey as much contextual information about the indoor space as possible to reduce spatial uncertainty.

In the problem scenarios, the goal was to better understand conceptual and linguistic patterns that would help to generate correct and concise automated indoor scene descriptions for a user who is unable to directly view the scene. In the second scenario, we assumed data capture and processing through a mobile device camera to produce a spatial graph that can generate an accurate scene description from the user's perspective. Assuming a perspective where the agent shares the same *in-the-container* perspective as the user, one approach would be to integrate spatial data and reason about the entire set of spatial information available to the system using a spatial property graph which could be annotated with spatial roles (e.g., object type/function, location, plausible mobility vs. static structure classification). From there, automated spatial descriptions could be generated based on spatial role labels and a machine learning algorithm employing preferred spatial prepositions to linguistically represent spatial relations between objects. For example, the spatial property graph (Figure 6.5) could be collected through computer vision and along with annotated scene descriptions with topological, geometric, and context cues would be available to provide a rich description of the space and the objects for those who could not see it.

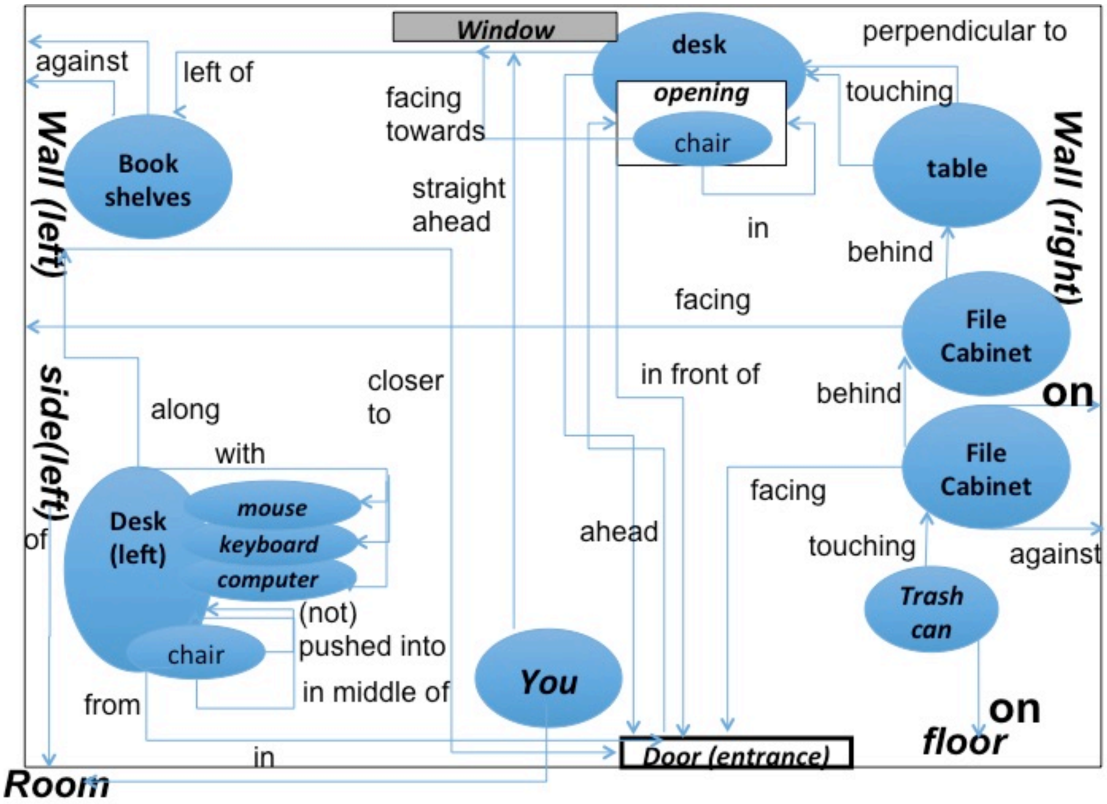


Figure 6.5: *Spatial Property Graph for Room-1*.

6.4.2. Guidelines for Indoor Scene Descriptions

Based on the findings from this study aligning spatial relations and NL spatial prepositions in indoor scenes, the human subject testing results suggest a preliminary set of guidelines, based on the following observations.

Guideline 1: The GUM-Space has *Connection* relation (*Contact*) does not specify what is the preferred NL spatial preposition to convey the scene below (Figure 6.6).

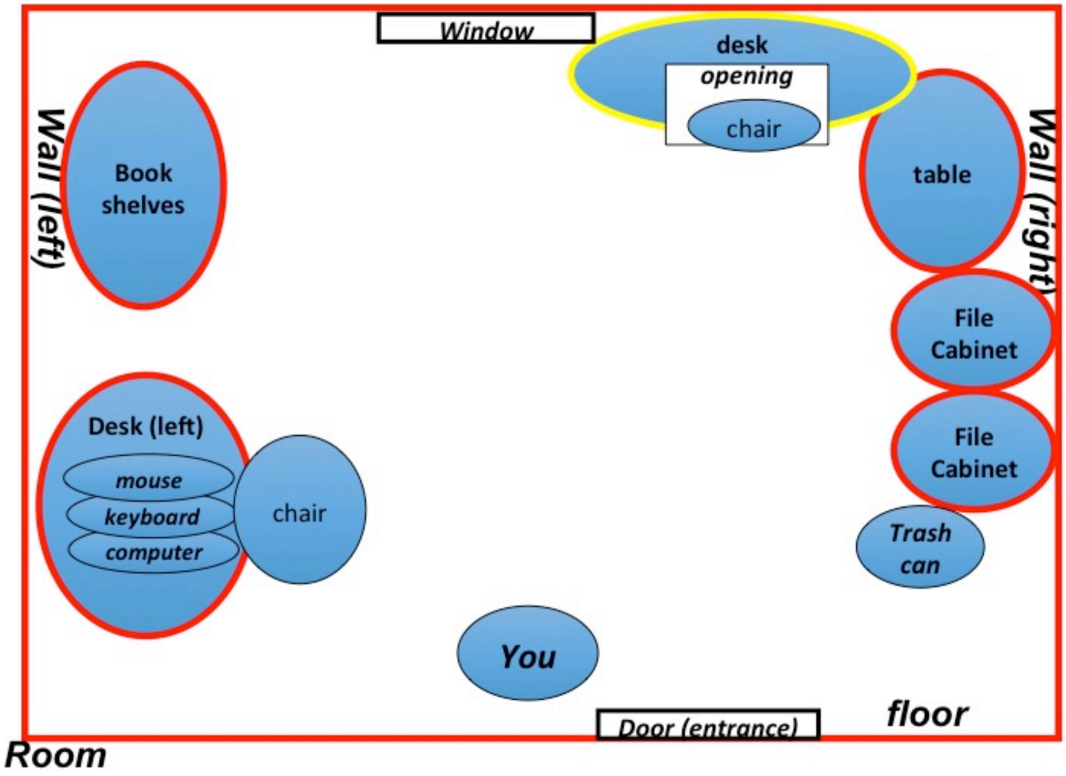


Figure 6.6: *Contact relation: A desk is against far wall.*

Based on the results of this study, guidelines for this example might specify that the spatial triple should consist of no more than three words to describe the trajectory (desk), no more than four words to describe the spatial relation *(connection/contact)* and no more than three words to describe the landmark (far wall). In addition, the preferred terms for this spatial relation would be against with alternative terms being *on, touching,* and *along* in that order.

Guideline 2: GUM-Space lacks rules to order object and/or structure relations with *contact* relation based on potential movement of objects, size of objects or scale of space (Figure 6.7). GUM-Space could use additional context information annotation classes to more precisely describe objects, structures and their interactions using principles, such as the relative size of the trajector to landmark, the potential mobility of each entity, and the scale of the indoor environment.

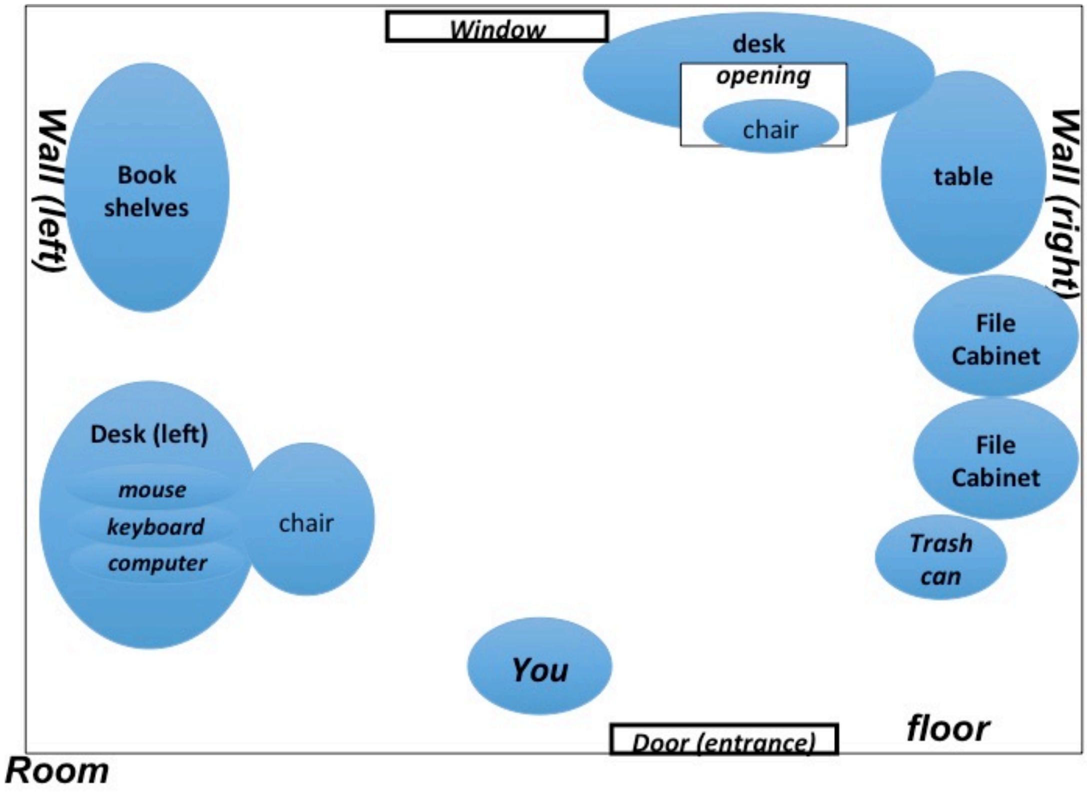


Figure 6.7*: Proximity 'moveable' objects*

Guideline 3: GUM-space does not have a way to classify spatial preposition use of objects/structures based on different scales of hierarchical indoor space (vista scale versus tabletop scale) (Figure 6.8). For example, "*There is a file cabinet in front of another file cabinet on the right wall.*" as opposed to "*You can use the mouse on the desk to operate the computer*". This hierarchical distinction between objects within indoor scenes and its related annotation will be necessary for an intelligent indoor scene description agent to provide salient NL descriptions depending on user needs and intended spatial tasks.

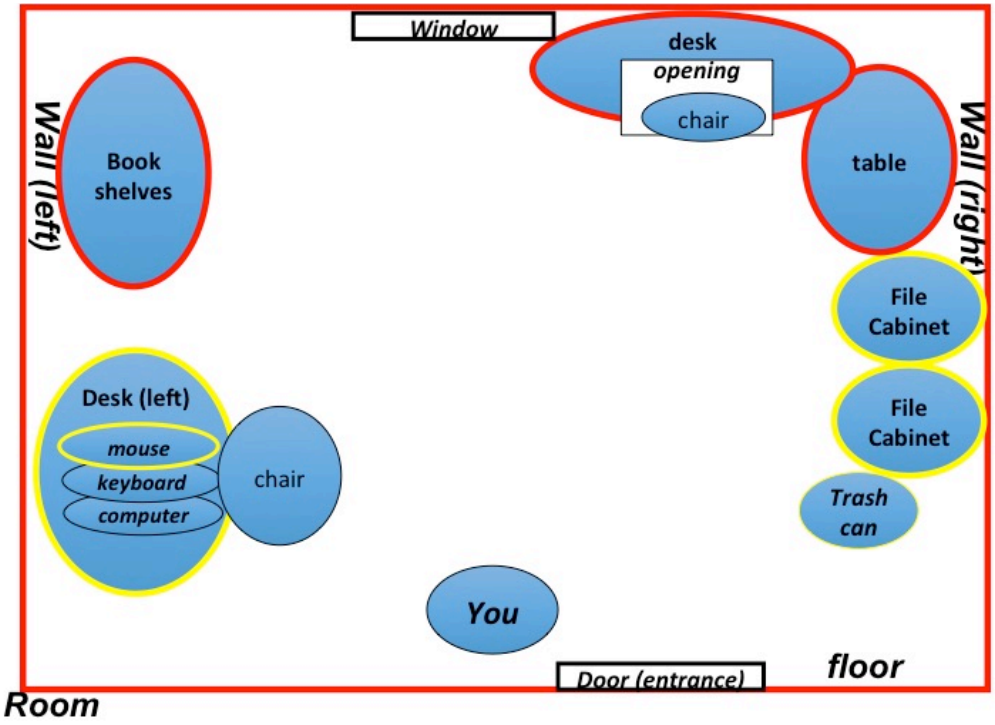


Figure 6.8: *Moveable objects in contact relation using against/on.*

Guideline 4: GUM-Space classifies *support* as a functional relation but does not currently have the capacity to classify objects by topological and functional relationships.

Based on the results of this dissertation, the scene below (Figure 6.9) could be represented as *the desk is on the left wall* or *the desk is on the wall to your left* and *the desk has a computer, keyboard and mouse on it*. Expanded annotation about preferred spatial prepositions that can be used in conjunction with object functions would allow for a richer representation of a collection of objects. This type of description would take into account both the topological configuration and relations based on each object's typical functions, creating a more precise NL description of the scene: *the computer is on top of the desk against the left wall.* (Figure 6.9)

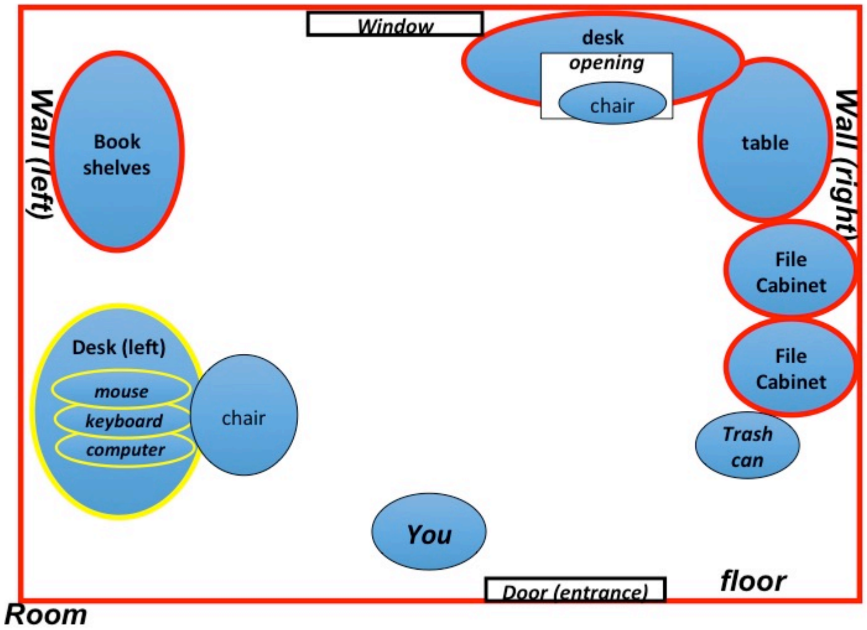
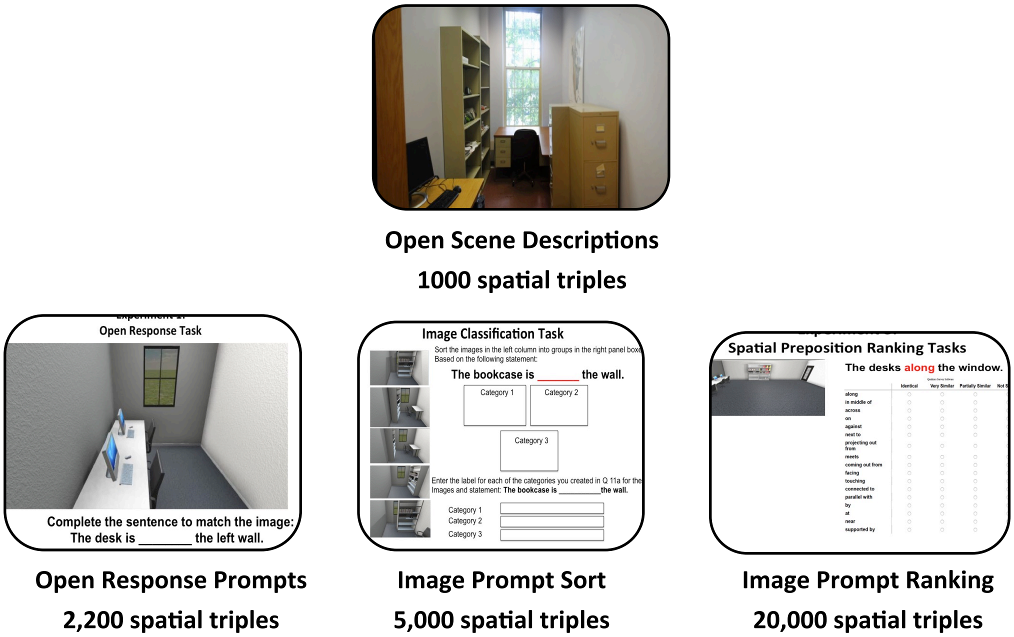


Figure 6.9: Multiple Uses of *on* in Functional and Topological Relations

6.4.3. Development of an Indoor Scene Corpus

Given the known challenges of existing spatial annotation schemes in NL research, this study hopes to contribute to the body of spatially annotated corpora with both the corpus of indoor scene descriptions and the annotated results of Experiments 1-3. There were over 28,000 spatial triples generated by this research (Figure 6.10). The spatial triples are mapped to images and are annotated with GUM-Space classification labels, Spatial Role Labels, and Room Context Labels. These types of resources with a fine detail level of spatial linguistic annotations are necessary to help researchers better understand the concepts at different spatial scales, spatial cues for anticipated motion detection, and frame of reference identification. This set of structured spatial data alone provides a substantial contribution to research on indoor scenes by providing additional resources to train machine learning models to recognize and automatically generate linguistic spatial concepts, reason about different spatial scales, and develop more intuitive descriptions for 3D objects in a variety of real-world spatial situations/scales. It also should help to develop better models for reducing uncertainty through probabilistic rankings of utterance semantics based on identified and validated indoor setting contextual cues.


Open Scene Descriptions
1000 spatial triples

Open Response Task

Complete the sentence to match the image:
The desk is ________ the left wall.

Image Classification Task

Sort the images in the left column into groups in the right panel boxe
Based on the following statement:

The bookcase is _______ the wall.

Enter the label for each of the categories you created in Q 11a for the Images and statement: **The bookcase is __________the wall.**

Category 1 ____
Category 2 ____
Category 3 ____

Spatial Preposition Ranking Tasks

The desks along the window.

	Identical	Very Similar	Partially Similar	Not S
along	○	○	○	
in middle of	○	○	○	
across	○	○	○	
on	○	○	○	
against	○	○	○	
next to	○	○	○	
projecting out from	○	○	○	
meets	○	○	○	
coming out from	○	○	○	
facing	○	○	○	
touching	○	○	○	
connected to	○	○	○	
parallel with	○	○	○	
by	○	○	○	
at	○	○	○	
near	○	○	○	
supported by	○	○	○	

Open Response Prompts
2,200 spatial triples

Image Prompt Sort
5,000 spatial triples

Image Prompt Ranking
20,000 spatial triples

Figure 6.10: *Indoor Scene Description Corpus Components*

6.4.4. Future Experiments

The next logical step in this line of research is to move the venue from a static 2D image and non-immersive VR environment to a fully immersive VR environment that would allow participants to perform a variety of spatial tasks and allow researchers to observe a variety of spatial behaviors and more precisely measure the outcomes. The VEMI lab recently completed an indoor navigation environment that would make an ideal experimental setting in which to isolate room context variables (Figures 6.11 and 6.12). This environment will allow participants to move through indoor space based on spatial scene descriptions. It would also allow for testing of spatial updating and spatial preposition use within an endless variety of indoor scenes.

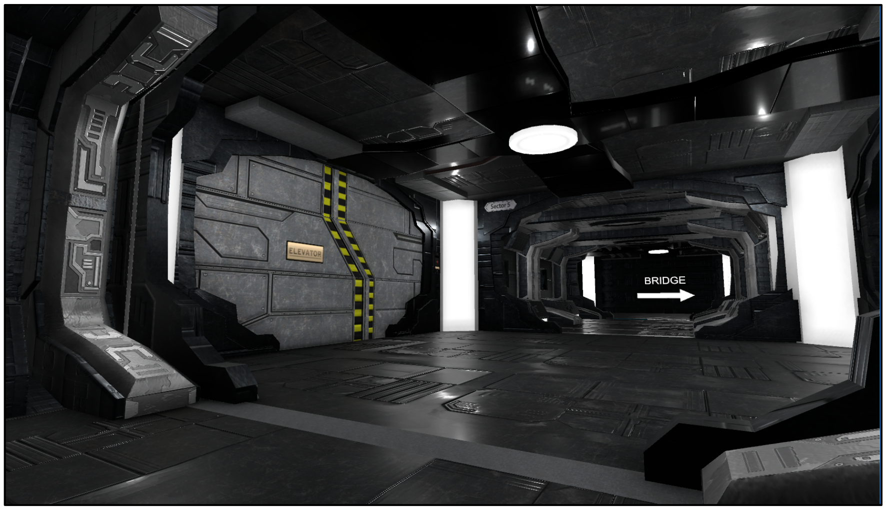


Figure 6.11: *VEMI Indoor navigation transition scene* (credit-John San Diego).

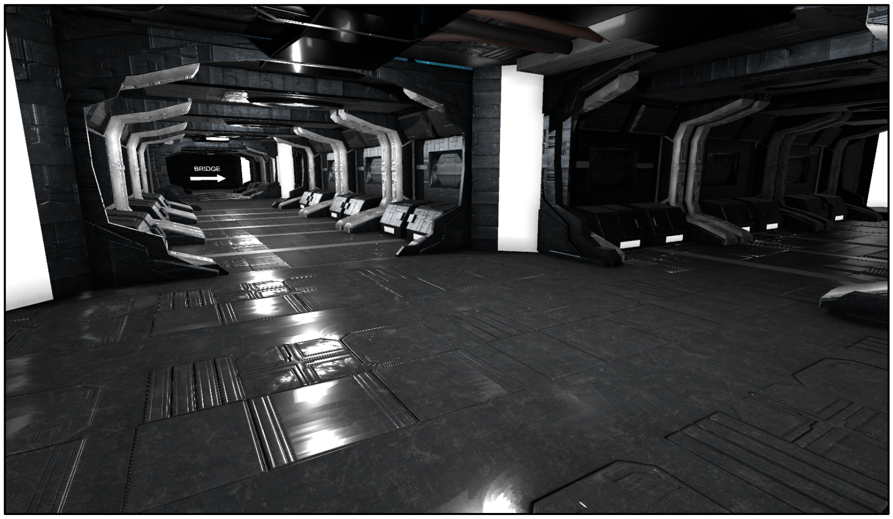


Figure 6.12: *VEMI Indoor navigation corridor scene* (credit-John San Diego).

This ability to create immersive environments that can be precisely controlled and manipulated provides additional benefits for experiments that specifically compare spatial language use in different scale spaces. The use of immersive virtual environments will help to provide more evidence to the assertion that ‘space is not space' when it comes to human psychology." (Montello, 1993) through the more precise testing of differences in spatial behaviors and tasks in environments that represent the size and perspective of the human body at different spatial scales.

6.4.5 Scene Descriptions and Virtual Assistants

To date, most of the voice-activated assistants, such as Alexa, Siri, Google, and Cortana, are limited to connecting into pre-existing knowledgebases or other 'Internet of Things' (IoT) enabled devices (e.g., lights, thermostats, security systems) to control different parts of an indoor environment. In the future, these devices will help people who are unable to easily locate items (e.g., blind, low vision or memory impaired) to be able to have the assistant survey the indoor environment and have the assistant provide spatial information about the target object within an indoor setting in real time. Many of the existing skills of these devices are already creating spatial networks of connected devices. Adding relevant topological and geometric data through the use of a combination of wireless beacons and RFID tags with available devices (e.g. smartphones and home assistants) would be the next step to building a model of indoor environments that could be queried in ways not possible by current systems. These devices can also learn about the indoor environment from their owner's scene descriptions, providing more information for the system to use at a later date.

REFERENCES

Abarbanell, L., & Li, P. (2015). Left-right language and perspective taking in Tseltal Mayan children. In *Proceedings of the 39th Annual Boston University Conference on Language Development* (Vol. 1, pp. 1-13).

Aditya, S., Yang, Y., Baral, C., Fermuller, C., & Aloimonos, Y. (2015). From images to sentences through scene description graphs using commonsense reasoning and knowledge. *arXiv preprint arXiv:1511.03292*.

Afyouni, I., Cyril, R., & Christophe, C. (2012). Spatial models for context-aware indoor navigation systems: A survey. *Journal of Spatial Information Science*, *1*(4), 85-123.

Allen, G. L. (2000). Principles and practices for communicating route knowledge. *Applied Cognitive Psychology*, *14*(4), 333-359.

American Physical Society. (2008). *Energy future: Think efficiency*. Washington, DC: American Physical Society.

Anderson, A.H., Bader, M., Bard, E., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J, McAllister, J., Miller, J., Sotillo, C., Thompson, H., & Weinert, R., (1991). The HCRC map task corpus. *Language and Speech* 34 (4) 351–366.

Barclay, M., & Galton, A. (2008, April). A Scene Corpus for Training and Testing Spatial Communication Systems. In *AISB 2008 Convention Communication, Interaction and Social Intelligence* (Vol. 1, p. 26).

Bateman, J. A., Henschel, R., & Rinaldi, F. (1995). The generalized upper model 2.0. In *Proceedings of the ECAI94 Workshop: Comparison of Implemented Ontologies*.

Bateman, J., Tenbrink, T., & Farrar, S. (2007). The role of conceptual and linguistic ontologies in interpreting spatial discourse. *Discourse Processes*, *44*(3), 175-212.

Bateman, J. A., Hois, J., Ross, R., & Tenbrink, T. (2010). A linguistic ontology of space for natural language processing. *Artificial Intelligence*, *174*(14), 1027-1071.

Bennett, D. C. (1975). *Spatial and temporal uses of English prepositions: An essay in stratificational semantics* (Vol. 17). Longman Publishing Group.

Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A., & Plank, B. (2016). Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures. *Journal of Artificial Intelligence Research (JAIR)*, *55*, 409-442.

Bowerman, M. (1996). Learning how to structure space for language: A crosslinguistic perspective. *Language and space*, 385-436.

Brodaric., B, Hahmann, T., & Gruninger, M.(under review) Water Features and Their Parts. *Applied Ontology.* (January, 2017).

Brugman, C., & Lakoff, G. (2006). Radial network. *Cognitive linguistics: basic readings. Berlin: Mouton de Gruyter*, 109-140.

Burigo, M., & Coventry, K. R. (2010). Context affects scale selection for proximity terms. *Spatial Cognition and Computation, 10,* 292-312.

Casati, R., & Varzi, A. C. (1999). *Parts and places: The structures of spatial representation*. MIT Press.

Chen, X., Fang, H., Lin, T. Y., Vedantam, R., Gupta, S., Dollár, P., & Zitnick, C. L. (2015). Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Choi, W., Chao, Y. W., Pantofaru, C., & Savarese, S. (2013). Understanding indoor scenes using 3d geometric phrases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 33-40).

Clark, H. H. (1973). Space, time, semantics, and the child. *Cognitive development and the acquisition of language*, *27*, 63.

Cohn, A. G., Bennett, B., Gooday, J., & Gotts, N. M. (1997). Representing and reasoning with qualitative spatial relations about regions. In *Spatial and Temporal Reasoning* (pp. 97-134). Springer Netherlands.

Coventry, K. R., Carmichael, R., & Garrod, S. C. (1994). Spatial prepositions, object-specific function, and task requirements. *Journal of Semantics*, *11*(4), 289-309.

Coventry, K., & Garrod, S. C. (2005). *Spatial prepositions and the functional geometric framework. Towards a classification of extra-geometric influences.*

Crowston, K. (2012). Amazon Mechanical Turk: A research tool for organizations and information systems scholars. In *Shaping the Future of ICT Research. Methods and Approaches* (pp. 210-221). Springer Berlin Heidelberg.

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS one*, *8*(3), e57410.

DiManzo, M., Adorni, G., & Giunchiglia, F. (1986). Reasoning about scene descriptions. *Proceedings of the IEEE*, *74*(7), 1013-1025.

Downs, R. & Stea, D. (1977) *Maps in Minds: Reflections in Cognitive Mapping*. New York Harper and Row.

Eberhard, K. M., Nicholson, H., Kübler, S., Gundersen, S., & Scheutz, M. (2010). The Indiana "Cooperative Remote Search Task"(CReST) Corpus. In *LREC*.

Egenhofer, M. J., & Herring, J. (1990). A mathematical framework for the definition of topological relationships. In *Proceedings of Fourth International Symposium on Spatial Data Handling* (pp. 803-813). Zurich, Switzerland.

Egenhofer, M. J., & Franzosa, R. D. (1991). Point-set topological spatial relations. *International Journal of Geographical Information System*, *5*(2), 161-174.

Egenhofer, M. J., & Franzosa, R. D. (1995). On the equivalence of topological relations. *International Journal of Geographical Information Systems*, *9*(2), 133-152.

Egenhofer, M. J., & Mark, D. M. (1995). Naive geography. In *Proceedings of International Conference on Spatial Information Theory* (pp. 1-15). Springer Berlin Heidelberg.

Egenhofer, M. J., & Vasardani, M. (2007). Spatial reasoning with a hole. In *Proceedings of International Conference on Spatial Information Theory* (pp. 303-320). Springer Berlin Heidelberg.

Elahi, M. F., Shi, H., Bateman, J. A., Eberhard, K. M., & Scheutz, M. (2012). Classification of Localization Utterances using a Spatial Ontology. *P–KAR 2012*, 13.

Evans, V. (2006). Lexical concepts, cognitive models and meaning construction. *Cognitive Linguistics*, *17*(4), 491-534.

Evans, V. (2009). Semantic representation in LCCM Theory. *New Directions in Cognitive Linguistics*, 27-55.

Evans, V. (2015). A unified account of polysemy within LCCM Theory. *Lingua*, *157*, 100-123.

Falomir, Z. (2012). Qualitative distances and qualitative description of images for indoor scene description and recognition in robotics. *AI Communications*, *25*(4), 387-389.

Farhadi, A., Hejrati, M., Sadeghi, M., Young, P., Rashtchian, C., Hockenmaier, J., & Forsyth, D. (2010). Every picture tells a story: Generating sentences from images. In *Proceedings of European Conference on Computer Vision (ECCV 2010)*, 15-29.

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

Feist, M. I., & Gentner, D. (2003). Factors involved in the use of in and on. In *Proceedings of the Twenty-fifth Annual Meeting of the Cognitive Science Society* (pp. 390-395).

Frank, A.U. (1996). The prevalence of objects with sharp boundaries in GIS. In Burrough P. and Frank A. (eds.) *Geographic Objects with Indeterminate Boundaries*. London, Taylor & Francis: 29-40.

Frank, A. U., & Mark, D. M. (1992). Language issues for geographical information systems.

Franklin, N., & Tversky, B. (1990). Searching imagined environments. *Journal of Experimental Psychology: General*, *119*(1), 63.

Freksa, C. (1992). Temporal reasoning based on semi-intervals. *Artificial Intelligence 54* (1) 199–227.

Freundschuh, S. M. (1992). Is there a relationship between spatial cognition and environmental patterns?. In *Theories and methods of spatio-temporal reasoning in geographic space* (pp. 288-304). Springer, Berlin, Heidelberg.

Freundschuh, S., & Blades, M. (2013). The cognitive development of the spatial concepts NEXT, NEAR, AWAY and FAR. In *Cognitive and Linguistic Aspects of Geographic Space* (pp. 43-62). Springer Berlin Heidelberg.

Freundschuh, S. M., & Egenhofer, M. J. (1997). Human conceptions of spaces: Implications for GIS. *Transactions in GIS*, *2*(4), 361-375.

Freundschuh, S. M., & Sharma, M. (1995). Spatial Image Schemata, Locative Terms, and Geographic Spaces in Children's Narrative: Fostering Spatial Skills in Children. *Cartographica: The International Journal for Geographic Information and Geovisualization*, *32*(2), 38-49.

Galton, A. (2012, July). States, Processes and Events, and the Ontology of Causal Relations. In *FOIS* (pp. 279-292).

Gapp, K. P. (1994). *Basic meanings of spatial relations: Computation and evaluation in 3d space* (pp. 1393-1398). Universität des Saarlandes.

Gibson, J. J. (1976). The theory of affordances and the design of the environment. In *Proceedings of the Annual Meeting of the American Society for Aesthetics, Toronto*.

Giudice, N. A., Betty, M. R., & Loomis, J. M. (2011). Functional equivalence of spatial images from touch and vision: evidence from spatial updating in blind and sighted individuals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(3), 621.

Giudice, N., Walton, L. & Worboys, M. (2010). The informatics of indoor and outdoor space: A research agenda. *In Proceedings of the 2nd ACM SIGSpatial International Workshop on Indoor Spatial Awareness*, 47–53. ACM Press.

Greene, M. R., Baldassano, C., Esteva, A., Beck, D. M., & Fei-Fei, L. (2016). Visual scenes are categorized by function. *Journal of Experimental Psychology: General*, *145*(1), 82.

Greeno, J. G. (1994). Gibson's Affordances. *Psychological Review*, *101*(2), 336-342.

Grenon, P., & Smith, B. (2004). SNAP and SPAN: Towards dynamic spatial ontology. *Spatial Cognition and Computation*, *4*(1), 69-104.

Gruber, T. R. (1992). *Ontolingua: A mechanism to support portable ontologies*. Stanford.

Guarino, N. (1998). Formal ontology and information systems, in: N. Guarino (Ed.), *Formal Ontology in Information Systems (FOIS*), IOS Press, Amsterdam, 3–18.

Hahmann, T., & Brodaric, B. (2013). Kinds of full physical containment. In *International Conference on Spatial Information Theory* (pp. 397-417). Springer International Publishing.

Hall, M. M., Smart, P. D., & Jones, C. B. (2011). Interpreting spatial language in image captions. *Cognitive processing*, *12*(1), 67-94.

Hayes, P. (1978) The Naive physics manifesto. in: D. Michie (Ed.), *Expert Systems in the Microelectronic Age*. Edinburgh, Scotland: Edinburgh University Press, (pp. 242-270).

Heeman, P.A., & Allen, J., (1995). The Trains 93 Dialogues, Trains Technical Note 94-2, Computer Science Dept., University of Rochester URL ftp://ftp.cs.rochester.edu/pub/papers/ai/95.tn2.Trains_93_dialogues.ps.gz.

Henderson, J. M., Larson, C. L., & Zhu, D. C. (2007). Cortical activation to indoor versus outdoor scenes: an fMRI study. *Experimental Brain Research*, *179*(1), 75-84.

Henderson, J. M., Zhu, D. C., & Larson, C. L. (2011). Functions of parahippocampal place area and retrosplenial cortex in real-world scene analysis: an fMRI study. *Visual cognition*, *19*(7), 910-927.

Herskovits, A. (1980). On the spatial uses of prepositions. In *Proceedings of the 18th annual meeting on Association for Computational Linguistics* (pp. 1-5). Association for Computational Linguistics.

Herskovits, A. (1986). *Language and spatial cognition: An interdisciplinary study of the prepositions in English, studies in natural language processing*. Cambridge University Press, London.

Hois, J., Kutz, O., & Bateman, J. A. (2008). Similarity-connections between natural language and spatial situations. In *Workshop on Spatial Language in Context: Computational and Theoretical Approaches to Situation Specific Meaning (in Association with Spatial Cognition, 2008)*.

Hois, J., & Kutz, O. (2008). Counterparts in language and space. In *Formal Ontology in Information Systems* (p. 266).

Hois, J. (2010). Inter-annotator agreement on a linguistic ontology for spatial language—A case study for GUM-Space, In *Proceedings of the 7th International Conference on Language Resources and Evaluation* (LREC 2010), LREC, 2010.

Ittelson, W. H. (1973). Environment perception and contemporary perceptual theory. In Ittelson, W. H. (ed.) *Environment and Cognition*. New York, Seminar: 1-19

Kesavan, S., & Giudice, N. A. (2012). Indoor scene knowledge acquisition using a natural language interface. In *Proceedings of Spatial Knowledge Acquisition with Limited Information Displays*, (*SKALID 2012)* 1-6.

Kesavan, S. (2013). Indoor scene knowledge acquisition using natural language descriptions. Master's thesis. University of Maine, Orono.

Klippel, A. (2009). Topologically characterized movement patterns: A cognitive assessment. *Spatial Cognition & Computation*, *9*(4), 233-261.

Klippel, A. (2012). Spatial information theory meets spatial thinking: is topology the Rosetta stone of spatio-temporal cognition?. *Annals of the Association of American Geographers*, *102*(6), 1310-1328.

Kray, C., Fritze, H., Fechner, T., Schwering, A., Li, R., & Anacta, V. J. (2013). Transitional spaces: Between indoor and outdoor spaces. In Proceedings of *International Conference on Spatial Information Theory* (pp. 14-32). Springer International Publishing.

Kuhn, W. (2007). An image-schematic account of spatial categories. *Spatial Information Theory*, 152-168.

Kuipers, B. (1978). Modeling spatial knowledge. *Cognitive Science*, *2*(2), 129-153.

Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A. C., & Berg, T. L. (2011). Baby talk: Understanding and generating image descriptions. In *Proceedings of the 24th Conference on Computer Vision and Pattern Recognition (CVPR 2011).* 1601-1608.

Lakoff, G. (1987). *Women, fire, and dangerous things*. University of Chicago press.

Landau, B., & Jackendoff, R. (1993). Whence and whither in spatial language and spatial cognition? *Behavioral and Brain Sciences*, *16*(02), 255-265.

Langacker, R. W. (1987). *Foundations of cognitive grammar: Theoretical prerequisites* (Vol. 1). Stanford University Press.

Langacker, R. W. (1993). Reference-point constructions. *Cognitive Linguistics*, *4*(1), 1-38.

Langacker, R. W. (2010). Reflections on the functional characterization of spatial prepositions. *CORELA - Numéros Thématiques - Espace, Préposition, Cognition*, 1–19. Retrieved from http://corela.edel.univ-poitiers.fr/index.php?id = 999

Levinson, S.C. (2003). *Space in Language and Cognition: Explorations in Cognitive Diversity,* Cambridge University Press, Cambridge.

Li, H., & Giudice, N. A. (2012). Using mobile 3D visualization techniques to facilitate multi-level cognitive map development of complex indoor spaces. *Spatial Knowledge Acquisition with Limited Information Displays*, *21*, 31-36.

Li, K. J., & Lee, J. Y. (2013). Basic concepts of indoor spatial information candidate standard IndoorGML and its applications. *Journal of Korea Spatial Information Society*, *21*(3), 1-10.

Lin, D. Fidler, S. Kong C. and Urtasun, R. (2015). Generating Multi-sentence Natural Language Descriptions of Indoor Scenes. In Xianghua Xie, Mark W. Jones, and Gary K. L. Tam, editors, In *Proceedings of the British Machine Vision Conference* (BMVC), pages 93.1-93.13. BMVA Press, September 2015.

MacMahon, M., Stankiewicz, B., & Kuipers, B. (2006). Walk the talk: connecting language, knowledge, and action in route instructions. In *Proceedings of the 21st National Conference on Artificial Intelligence-Volume 2* (pp. 1475-1482). AAAI Press.

Mark, D. M., & Egenhofer, M.J. (1994). Modeling spatial relations between lines and regions: Combining formal mathematical models and human subject testing. *Cartography and Geographic Information Systems.* 21 (3):195–212.

Mark, D. M., & Frank, A. U. (1992). *NCGIA Initiative 2: Languages of Spatial Relations, Closing Report*. NCGIA (National Center for Geographic Information and Analysis).

Masolo, C., Borgo, S., Gangemi, A., Guarino, N., & Oltramari, A. (2003). Wonderweb deliverable d18, ontology library (final). *ICT project*, *33052*.

Miller, G.A., & Johnson-Laird, P.N. (1976). *Language and Perception*. Cambridge University Press, Cambridge.

Montello, D. R. (1993). Scale and multiple psychologies of space. In *European Conference on Spatial Information Theory* (pp. 312-321). Springer Berlin Heidelberg.

Montello, D. R. (2009). Cognitive research in GIScience: Recent achievements and future prospects. *Geography Compass*, *3*(5), 1824-1840.

Montello D. R. & Raubal M. (2012). Functions and applications of spatial cognition. In Waller D and Nadel L (eds) Handbook of Spatial Cognition. Washington, DC, American Psychological Association: 249–64

Moratz, R. (2006). Representing relative direction as a binary relation of oriented points. In *ECAI* (Vol. 6, pp. 407-411).

Mozos, O. M., Jensfelt, P., Zender, H., Kruijff, G. J. M., & Burgard, W. (2007). From labels to semantics: An integrated system for conceptual spatial representations of indoor environments for mobile robots. In *ICRA Workshop: Semantic Information in Robotics*.

Museros, L., & Escrig, M. T. (2004). A qualitative theory for shape representation and matching for design. In *Proceedings of the 16th European Conference on Artificial Intelligence* (pp. 858-862). IOS Press.

Nothegger, C., Winter, S., & Raubal, M. (2004). Selection of salient features for route directions. *Spatial Cognition and Computation*, *4*(2), 113-136.

Norman, D. (2002). *The design of everyday things*. New York, NY: Basic Books.

Oliva, A., & Schyns, P. G. (2000). Diagnostic colors mediate scene recognition. *Cognitive Psychology*, *41*(2), 176-210.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research*, *155*, 23-36.

Pacheco, J., Escrig, M. T., & Toledo, F. (2002). Qualitative spatial reasoning on three-dimensional orientation point objects. In *Proceedings of the QR2002. 16th International Workshop on Qualitative Reasoning.*

Pingel, T. J., & Schinazi, V. R. (2014). The relationship between scale and strategy in search-based wayfinding. *Cartographic Perspectives*, (77), 33-45.

Renz, J. (2002). *Qualitative spatial reasoning with topological information*. Berlin: Springer.

Richter, K. F., Winter, S., & Santosa, S. (2011). Hierarchical representations of indoor spaces. *Environment and Planning B: Planning and Design*, *38*(6), 1052-1070.

Riehle, T. H., Lichter, P., & Giudice, N. A. (2008). An indoor navigation system to support the visually impaired. In *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE* (pp. 4435-4438). IEEE.

Rüetschi, U. J. (2007). *Wayfinding in scene space modelling transfers in public transport* (Doctoral dissertation). DOI: 10.5167/uzh-20208

Schwering, A.: Evaluation of a semantic similarity measure for natural language spatial relations. In S. Winter, B. Kuipers, M. Duckham, & L. Kulik (Eds.), *Spatial Information Theory*. 9th International Conference, COSIT 2007, (2007) Melbourne, Australia, Berlin: Springer.

Shariff, A. R. B., Egenhofer, M. J., & Mark, D. M. (1998). Natural-language spatial relations between linear and areal objects: The topology and metric of English-language terms. *International Journal of Geographical Information Science*, *12*(3), 215-245.

Sithole, G., & Zlatanova, S. (2016). Position location, place and area: An indoor perspective. *ISPRS Annual Photogramm Remote Sensing Spatial Information Sci*ence *4*, 89-96.

Talmy, L. (1978). Relations between subordination and coordination. *Universals of Human Language: Syntax*, *4*, 487-513.Talmy, L. (1983). How language structures space. In H. Pick and L. Acredolo (Eds.), *Spatial orientation: Theory, research, and application*. New York: Plenum Press.

Talmy, L. (1983). How language structures space. In H. Pick and L. Acredolo (Eds.), *Spatial orientation: Theory, research, and application*. New York: Plenum Press.

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1* (pp. 173-180). Association for Computational Linguistics.

Tran, K., He, X., Zhang, L., Sun, J., Carapcea, C., Thrasher, C., & Sienkiewicz, C. (2016). Rich image captioning in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 49-56).

Tversky, B. (1981). Distortions in memory for maps. *Cognitive psychology*, *13*(3), 407-433.

Tversky, B. (1993). Cognitive maps, cognitive collages, and spatial mental models. In *European Conference on Spatial Information Theory* (pp. 14-24). Springer Berlin Heidelberg.

Tversky, B. (2001). Spatial schemas in depictions. In M. Gattis (Ed.), *Spatial schemas and abstract thought*. 79–111. Cambridge: MIT Press. Tversky,

Tversky, B. (2009). Spatial cognition: Embodied and situated. In Murat Aydede & P. Robbins (eds.), *The Cambridge Handbook of Situated Cognition*. Cambridge: Cambridge University Press. pp. 201--217

Tyler, A. & Evans, V. (2003). *The semantics of English Prepositions: Spatial scenes, and the polysemy of English prepositions.* Cambridge: Cambridge University Press

Ursini, F. A., & Akagi, N. (2013). Another Look at Modification in Spatial Prepositions. *Iberia*, *5*(2), 38.

Vailaya, A., Figueiredo, M. A., Jain, A. K., & Zhang, H. (1998, December). Bayesian framework for semantic classification of outdoor vacation images. In *Electronic Imaging'99* (pp. 415-426). International Society for Optics and Photonics.

Vandeloise, C. (2006). Are there spatial prepositions? In *Hickmann, M., & Robert, S. (Eds.). Space in languages: Linguistic systems and cognitive categories (Vol. 66). John Benjamins Publishing.* 139-154.

Vasardani, M., Timpf, S., Winter, S., & Tomko, M. (2013). From descriptions to depictions: A conceptual framework. In *International Conference on Spatial Information Theory* (pp. 299-319). Springer International Publishing. Vieu, L. (1997). Spatial representation and reasoning in artificial intelligence. In *Spatial and temporal reasoning* (pp. 5-41). Springer Netherlands.

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3156-3164.

Walton, L., & Worboys, M. (2009). An algebraic approach to image schemas for geographic space. In *International Conference on Spatial Information Theory* (pp. 357-370). Springer Berlin Heidelberg.

Walton, L. A., & Worboys, M. (2012). A qualitative bigraph model for indoor space. In *International Conference on Geographic Information Science* (pp. 226-240). Springer Berlin Heidelberg.

Wang, R. F., & Spelke, E. S. (2002). Human spatial representation: Insights from animals. *Trends in cognitive sciences*, *6*(9), 376-382.

Winter, S. (2012). Indoor spatial information. *International Journal of 3-d Information Modeling. 1*(1) 25-42. Worboys, M. (2011). Modeling indoor space. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Indoor Spatial Awareness*, 1–6. doi:10.1145/2077357.2077358

Wolbers, T., & Wiener, J. M. (2014). Challenges for identifying the neural mechanisms that support spatial navigation: the impact of spatial scale. *Frontiers in human neuroscience*, *8*.

Wu, S., Wieland, J., Farivar, O., & Schiller, J. (2017). Automatic alt-text: Computer-generated image descriptions for blind users on a social network service. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (pp. 1180-1192). ACM.

Yang, L., & Worboys, M. (2011). A navigation ontology for outdoor-indoor space:(work-in-progress). In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Indoor Spatial Awareness* (pp. 31-34). ACM.

Zubin, D. (1989). Natural language understanding and reference frames. In Mark D., Frank A., Egenhofer M., Freundschuh S., McGranaghan M. & White, R. M. (eds.) *Languages of spatial relations: Initiative 2 specialist meeting report technical paper* 89-2 Santa Barbara, CA, National Center for Geographic Information and Analysis: 13-16.

BIOGRAPHY OF THE AUTHOR

Stacy Doore completed a M.S. degree in Spatial Information Science and Engineering in 2011, a B.A. in Cultural Anthropology and a B.S. in Education in 1999 at the University of Maine. Currently, her research interests focus on improving descriptive specifications for spatial environment human-computer interactions. She served as an NSF ADVANCE-IT Internal Evaluator from 2013 to 2017 and was responsible for collecting and analyzing longitudinal data on institutional transformation in the areas of gender equity and diversity of STEM faculty. Stacy started her own evaluation and research consulting business in 2015 and currently serves as an external evaluator for a projects funded by NSF, IES, USDA, and NIH. She is the founder and first president of the student chapter of the University of Maine's Association of Computing Machinery-Women (UMaine ACM-W) funded by a grant from the National Center for Women in Technology (NCWIT) and Google.org. She enjoys developing outreach activities to promote K-12 student interest in STEM and Computer Science education and career paths. Stacy is a candidate for the Doctor of Philosophy degree in Spatial Information Science and Engineering from the University of Maine in August 2017.

.