Running head: TIME TO UNDERSTAND A CLUTTERED SPATIAL SCENE

Time will not help unskilled observers to understand a cluttered spatial scene

Thora Tenbrink

School of Linguistics and English Language, Bangor University, Bangor LL 572DG, UK;

t.tenbrink@bangor.ac.uk

Evelyn Bergmann

University of Bremen, Enrique-Schmidt-Str. 5, 28359 Bremen, Germany;

e.bergmann@uni-bremen.de

Christoph Hertzberg

University of Bremen, Robotics Group, Robert-Hooke-Str. 1, 28359 Bremen, Germany;

chtz@informatik.uni-bremen.de

Carsten Gondorf

Center for Cognitive Science, University of Kaiserslautern, 67663 Kaiserslautern,

Germany; gondorf@rhrk.uni-kl.de

Abstract

When searching for people in collapsed buildings, Urban Search and Rescue workers need to comprehend a complex cluttered scene observed through an endoscope under time pressure. This paper addresses the effects of time pressure and spatial ability on the comprehension of a film showing a mock-up collapsed room that was explored using endoscope-like technology. Participants' task was to find the objects that were hidden in the rubble, describe where they had found them, and draw the scene. Analysis focused on coherence and spatial specificity. Results indicate that spatial skills were most decisive for understanding and conceptually integrating the scene. Time pressure only affected the amount of objects found, not the degree of conceptual integration as reflected in the descriptions.

Time will not help unskilled observers to understand a cluttered spatial scene

Introduction

How do we make sense of an unstructured environment? People involved in Urban Search and Rescue, USAR workers, are confronted with this challenge when searching for trapped people in collapsed buildings. Not only has the building itself lost its structure as a result of the disaster, but access is typically severely restricted, including perception: rescuers may not be able to see any further than the next heap of stones. After possible survivors have been detected (e.g., by canine search) endoscopes are often employed allowing USAR workers to see inside structures that are as yet inaccessible from the outside. Recorded or transmitted images and films then enable the rescuers to do a preliminary search before actually entering the scene.

For a rescue operation to be successful, it is vital for USAR workers to make sense of what they see, i.e., to develop a coherent mental representation of the collapsed building. They need to determine whether it is safe to enter the scene, and the locations need to be identified where persons could be covered by debris. Upon entering a room, USAR workers will act on the basis of the representation they have developed, and communicate about the situation with their co-workers. All of this poses a major cognitive challenge, and depends to a high extent on the quality of the view provided by the endoscope-like device, causing Casper and Murphy (2003) to recommend research on perceptual user interfaces.

Unfortunately, current state-of-the art devices of this kind come with a number of problems (Hamp, Gorgis, Labenda, & Neumann, 2013). They are not easy to operate, nor are the views they provide simple to interpret. When navigating the endoscope through the scene, a main challenge is to keep track of the camera's position and orientation as it

gets out of the operator's sight. Due to the angled and tiltable camera head, orientation constantly changes, creating substantial psychological difficulty similar to angled laparoscopy in medicine (Hegarty, Keehner, Cohen, Montello, & Lippa, 2007). Within camera images, it is hard to estimate distances and object sizes, especially if reference objects are lacking in the environment, as is often the case with collapsed buildings. USAR workers have to find out how to navigate within the scene, and establish a sense of the route they are taking so as to cover and search within the whole room without neglecting less visible parts. Based on intuitions and inferences about the scales and distances seen in the scene, they need to gradually build up their spatial knowledge of the situation by relying on whatever landmarks they can recognise. Thus, an object configuration that may easily be understood when observed at once and in daylight will be hard to conceptualise under these perceptually hard conditions.

How can such a complex task be solved, and what are the factors that influence task performance most decisively? Intuitively, at least two factors should be relevant in this scenario. On the one hand, the skills of the observer may be crucial, as individuals are known to differ extensively with respect to spatial abilities (Hegarty, Montello, Richardson, Ishikawa, & Lovelace, 2006). On the other hand, the USAR scenario involves serious time pressure. Trivially, time will restrict the range of observations that can be made. However, beyond mere quantity, the implications of time pressure might furthermore affect the *quality* of observations made. In a visually cluttered scene with many distractors, it is substantially more demanding to remember where things are than to simply register their existence (Körner & Gilchrist, 2008). Given enough time, will even unskilled workers be able to comprehend a complex cluttered scene sufficiently to remember the relevant spatial relationships? Here we address this question by focussing on the communication of object locations within an unstructured scene. The ability to communicate information about a spatial scene indicates the speakers' comprehension of

it, since only those parts of the scene can be communicated that have been sufficiently understood to be verbalised. Moreover, communication is a crucial feature of a realistic USAR scenario (Casper & Murphy, 2003), which typically involves a team of people working together. In such a situation, understanding a complex spatial scene is only part of the problem. When one person begins to understand the situation, crucial information needs to be communicated to others so that appropriate actions can be taken. Since language production can be instrumental for building up mental representations (Hermer-Vazquez, Moffet, & Munkholm, 2001), the ability to verbalise the spatial situation successfully is crucial in this scenario.

Strategies for communicating spatial structure

In general, verbal descriptions of spatial environments are highly structured, reflecting underlying principles of conceptualising spatial configurations. Locally, the relationship between a target object and a salient reference item (henceforth called *relatum*) that is accessible in the discourse or situational context (Couclelis, Golledge, Gale, & Tobler, 1987; Miller & Johnson-Laird, 1976; Talmy, 2000) is crucial for both simple and complex settings. If no single object is found to be sufficiently salient to serve as an unambiguous relatum, speakers tend to start by identifying a reference region (Beun & Cremers, 2001) so as to reduce the search space. If the search space is as complex as a house (Plumert, Carswell, de Vet, & Ihrig, 1995), or even a city (Tomko & Winter, 2009), reference regions may be hierarchically structured by either zooming in or zooming out.

When describing complex spatial layouts the entire scene is typically conveyed sequentially, using a continuous trajectory by specifying the spatial relationships between multiple objects (Levelt, 1982). Structures such as perceived rows or clusters are readily used to organise a spatial description (Andonova, Tenbrink, & Coventry, 2010). Further aspects affecting the coherence of the representation include the viewpoint during

acquisition (Taylor & Tversky, 1992) and the order in which objects have been perceived (Buhl, 1996). Sequential strategies are highly efficient: they make sure that no item is omitted, and they facilitate picturing the scene for the addressee. The specific description strategy used in each case depends on scale. Scenes that involve navigation in environmental space invite an imaginary tour, like a route description (Linde & Labov, 1975), with changing perspectives. Scenes in figural or vista space that can be apprehended at once (Montello, 1993) invite a consistent perspective as basis for description (Ullmer-Ehrich, 1979). Altogether, descriptions of this kind convey a sense of continuous space in that the relationship between all parts of the whole scene are specified, and the individual objects are localised coherently with respect to each other as well as the overall scene.

Representation of complex spatial scenes

When committing spatial information to memory, structural features of the scene influence representation in recall (McNamara, 2003), and meaningful relationships between objects facilitate memory and linguistic representation (Radvansky & Copeland, 2000; Williams, Henderson, & Zacks, 2005). Although the source perspective and manner of acquisition affect recall (Evans & Pezdek, 1980; Sholl, 1987; Thorndyke & Hayes-Roth, 1982), humans are typically able to adopt perspectives flexibly once a scene is fully integrated (Taylor & Tversky, 1992; Brunyé, Rapp, & Taylor, 2008), and they often adapt to their addressee (Schober, 1995; Herrmann & Grabowski, 1994) as well as to the specific requirements in a spatial task (Taylor, Naylor, & Chechile, 1999). The ability to produce a coherent verbal representation from memory is thus based on mental coherence achieved through a good understanding of the spatial configuration in question. Across scenarios, women are consistently better at object location memory tasks than men (Voyer, Postma, Brake, & Imperato-McGinley, 2007), although males perform better at tasks based on

direct spatial experience (Montello, Lovelace, Golledge, & Self, 1999).

In large-scale space, the process of accumulating such integrated spatial representations is traditionally believed to proceed in successive stages of spatial knowledge: from landmark via route to survey knowledge, following Siegel and White (1975). Here, knowledge about the identity of individual landmarks incrementally extends to knowledge about routes, which eventually integrate into a survey-like conceptual representation of the environment. In this process, *time* spent in the spatial environment in question is assumed to be one of the most decisive factors. Arguably, time allows for deeper cognitive processing and repetitive memorisation cycles, supporting a better cognitive integration. Challenging this view, (Montello, 1998) argued for a more continuous developmental progression, where quantitative (metric) information is gradually accumulated without any major qualitative shifts. In line with this, Ishikawa and Montello (2006) found that survey knowledge may be acquired already from first exposure. This effect was mediated decisively by individual abilities. Spatial abilities are indeed known to affect performance in many contexts, ranging from mental rotation (Geiser, Lehmann, & Eid, 2006) to spatial layout and environmental learning (Hegarty et al., 2006; Allen, Kirasic, Dobson, Long, & Beck, 1996) and navigation (Wolbers & Hegarty, 2010). Clearly, such effects are task-dependent and affected by scale; like verbal descriptions, spatial skills vary considerably depending on the features and challenges at hand.

With respect to scale, USAR workers are confronted with an unusual situation. Although the space they explore is fairly limited, they nevertheless have considerably less visual access than is normally available in uncluttered small-scale (i.e., non-navigational) space. This transcends the scale differentiations into figural, vista and environmental (navigational) space suggested by Montello (1993). The endoscope provides only a narrow view, which results in a procedural exploration of the scene, comparable to sequential

exploration during navigation. The USAR scenario therefore evokes a range of interesting research questions. Motivated by the insights summarised so far, we address the following:

- Previous research on spatial descriptions suggests that speakers rely heavily on the structure perceived in a spatial setting. However, the known difficulties in keeping track of an endoscope's position and orientation may preclude developing a coherent concept that could be verbalised. *What kinds of description strategies emerge in such a situation?*

- Previous research suggests that both time and abilities can be decisive in developing coherent mental representations (corresponding to survey knowledge). Since time is substantially constrained in an USAR scenario, these effects are crucial. *To what extent do time and spatial abilities affect the quality of mental representation as represented by the coherence and specificity of a verbal description?*

Empirical Study

Since we were interested in spatial skill rather than expertise, our participants were as untrained as those in related earlier studies (as cited above). We designed a scenario that matched the USAR challenge in crucial respects but did not require any technological knowledge or USAR expertise. The effects of time pressure were addressed by using a between-participants design, *with* versus *without time constraints*. This addresses time pressure as such, rather than the considerable anxiety involved in a real USAR situation where lives are at stake, which we consider a separate issue.

Individual spatial skills were assessed by the Santa Barbara Solids Test (Cohen & Hegarty, 2012), a 30-item multiple choice test that addresses the ability to identify the two-dimensional cross section of a three-dimensional geometric solid. We chose this test because understanding a two-dimensional representation of a complex three-dimensional structure is highly relevant to the USAR scenario targeted in our study. Specifically, this test deals with recognising and comprehending structures as if visually penetrating the

depicted space (Cohen & Hegarty, 2012). In contrast, the classic domain-general spatial ability tests such as the Vandenberg Mental Rotation Test (Vandenberg & Kuse, 1978) and the Paper Folding Test (Ekstrom, French, Harman, & Dermen, 1976) focus on visualising and mentally manipulating objects, which is not required in the present scenario. The skill tested by the Solids Test has been shown to be distinct from the spatial skills addressed in the classic tests.

Participants

37 students took part in the study for course credit (32 female, 5 male; mean age 22.8, SD 3.4). One data set (from a female participant) had to be excluded due to technical problems. Seventeen participants completed the task without time pressure, and nineteen with time constraints.

Materials

We used a mock-up of a collapsed room that consisted of a $1 \times 2\,\mathrm{m}^2$ rectangle made of styrofoam boards. Alongside the walls, styrofoam clutter mimicked rubble. Eight objects were placed at random positions at the walls, beside or on top of rubble and protrusions (see Figure 1). They differed systematically with respect to visibility.[1] At the beginning, only one object was clearly visible (the tennis ball), and two further objects were non-salient but discernible (the table and the bottle opener). All other objects were occluded and only came into view when the camera approached the walls. They appeared at the margins of the screen rather than in the center. To confirm the differences in visibility in the film we collected participants' ratings of perceived visibility for each object.

The scenario was filmed using a camera head that was originally built for Hertzberg, Wagner, Birbach, Hammer, and Frese (2011) and improved later. It is equipped with two cameras and an inertial measurement unit which measures the current orientation of the

camera and was used to calculate an artificial horizon. For this study we mounted the camera head via a servo motor to a pole (see Figure 2). This allowed the cameras to look in all directions by using a combination of tilting the pole and turning the servo. In this study only the images of the right camera were used. The resulting view corresponds to state-of-the-art endoscope technology used in USAR situations with respect to view restrictions and turning and re-orientation behaviour. In particular, due to technical reasons, the camera head could not rotate for the back view without an image rotation of 90° to the side, which is equivalent to tilting one's head to the side to look back.

The film proceeds as follows. The camera enters a rectangular room at the narrow side and proceeds towards the far side. After having traversed one third of the room, the view gradually turns $\approx 140°$ back towards the entrance, and then back again, allowing to look into the indents. The camera first turns to the right and then to the left. After that, it proceeds further towards the far end. It rotates back again after two thirds of the way, and then again when arriving directly in front of the wall on the far side. When played without interruption at normal speed, the film takes 1:51 minutes.

The camera rotation affected how objects came into view. Since most of the objects were hidden in the indents, they could not be perceived while the camera was in normal upright position, but first came into view with a rotation of 90°. An artificial horizon was displayed on the screen to help maintain orientation (see Figure 3). Nevertheless, the conceptual challenge was much higher than it would be if the observer simply tilted their head. With head movement, the brain normally has no problems adjusting to a tilt. In contrast, an externally induced change of orientation needs to be processed with considerable cognitive effort (Hegarty et al., 2007). As a result, it was difficult to determine whether an object was standing or lying on the surface.

Participants never saw the actual mock-up arrangement. They could control the speed and were able to pause and play the film forwards and backwards, but were

otherwise restricted by the way the film was recorded. In addition to the camera head rotations, further perceptual challenges resembling real emergency scenarios included darkness and low contrast; also, the scale of objects and distances in the environment were hard to determine.

Procedure

Film Observation Task. Participants were led to the computer and told that they would be presented with a film which was shot in an environment that resembles the interior of a collapsed building. They heard the following main instruction (translated from the German original):

"Your task is to find all 8 objects. Please also memorise WHERE you found them and how they were standing or lying there, so that you can describe this later."

They were shown how to use the joystick to play the film forwards and backwards and to control speed, and were asked to click on any item that they considered to be an object. They practised the procedure until they felt safe to begin the main film. The practice film resembled the camera movement and perspective rotation as shown in the main film, but was shot outside the mock-up.

Participants in the *no-pressure condition* were not restricted in time, and the film's default speed corresponded to the speed of production (10 frames per second). Participants in the *time-pressure condition* were instructed to find the objects within three minutes. A timer was visible at the top right corner which counted down the minutes, and the film was initially displayed at an increased frame rate of 30 frames per second. However, participants in both conditions could adjust the speed (frame rate) according to their preferences.[2]

Description Task. Next, participants were asked two questions by the experimenter, in the form of a natural dialogue:

1. *"What do you remember? Where were the objects and how did they stand or lie there?"*

2. *"Can you tell me something about the shape of the room?"*

Participants were reminded to provide the location of the objects in case they did not spontaneously do so: *"Could you also say **where** the object was?"*

Drawing Task. Following verbal description, participants were asked to draw their perception of the scene, using the following prompt: *"Could you please roughly sketch the shape of the room, and also sketch the objects in?"*

Questionnaires. Directly after the tasks, participants were asked to rate the objects' visibility on a scale between 1 (very good) to 6 (very bad). Then they performed the Santa Barbara Solids test (Cohen & Hegarty, 2012), shown in color on a computer screen. The test was terminated after 5 minutes. Participants proceeded through the test self-paced and chose answers on a paper sheet in front of them. The number of correct answers was operationalised as a measure for spatial ability.

Analysis

The description task was audio recorded and transcribed. The language data were segmented into *utterances*[3] and systematically annotated in an iterative process by two independent coders, using CODA (Cognitive Discourse Analysis) (Tenbrink, 2015). We focused on two main aspects as follows.

Coherence To assess coherence as a measure of participants' ability to comprehend the spatial scene shown in the film, we visualised the order of mention of individual objects by drawing *trajectories* into photographies of the scene, following Tenbrink, Coventry, and Andonova (2011). This visual inspection provides a first indication of any sequential

ordering principle that the participants might have conceptualised. However, only linear patterns could be clearly identified in this way; a cluster configuration with a salient relatum at its center will generate a chaotic visual representation, but can be perfectly coherent in terms of its linguistic description. In a second step we therefore assessed linguistic coherence by identifying description strategy categories based on object localisation patterns (see Appendix). Overall scene descriptions were classified as belonging to a category if the participant used the same localisation type for description of at least two-thirds of the objects described. As such, the identification of description strategies is a qualitative result of the present study.

Specificity Next, we addressed the specificity of object location descriptions as another prime indicator of the participants' ability to integrate the spatial relationships into a coherent mental representation. We reasoned that if participants were able to describe the specific localisation of most objects this would indicate an integrated conceptual representation of the room. On the other hand, if the scene was not well understood and spatial integration was hard for the participants, they might resort to unspecific (local) descriptions. Since descriptions of similar kinds have not been analysed linguistically before, coding definitions were operationalised iteratively based on inspection of the data (see Appendix).

As an additional measure related to specificity, we analysed mention of spatial *axes*. In contrast to most scenarios in the literature, in our setting all three axes (lateral, frontal, and vertical) were relevant for successful localisation, since the objects were scattered at various positions in the cluttered scene. This may be especially important for communication in the absence of a clearly ordered sequential description. If individually described objects were specified with respect to all three axes of the room, this would serve as an optimal substitute for the inferences typically allowed by a coherent

description. Mention of axes could be based on projective terms (e.g., left/right for the lateral axes, front/back for the frontal, and above/below for the vertical), or on meanings implied by the verb or prepositional phrases: verbs such as *stand, lie, hang* and phrases such as *on the ground, from the ceiling* imply a vertical position relative to a relatum (ground/ceiling).

Sketch Performance We obtained scores for the quality of sketches as follows. As an operationalisation of representation completeness, two independent coders rated the rooms' shape as well as object placements. To ensure reliability of the codings, Krippendorff's Alpha (Krippendorf, 2004) was computed yielding excellent results ($\alpha = 0.884$).

The following scheme was designed after inspection of the drawings:

Room Shape

- 5 points were given if the drawings consisted of a rectangular shape with correct orientation
- 4 points: rectangular shape, wrong orientation
- 3 points: shape is four-sided
- 2 points: shape is distorted
- 1 point: no discernible shape

Object Placement 1 point was given for each individual object that was placed correctly inside the room by way of adding up:

- 0.25 for the *presence* of an object in the sketch
- 0.25 if the object was placed on the correct *side*
- 0.25 for placement in the correct *quadrant*
- 0.25 if correct *vertical* placement was indicated

An additional 0.25 points were given for correct object *orientation*; this was only relevant for three objects. Thus, a maximum of 13.75 points could be reached.

Finally, to assess the relationship between the factors used and identified in our study, we computed a generalised linear model (Bates, Maechler, Bolker, & Walker, 2014) with condition, sketch accuracy and spatial ability as fixed effects and total number of specific markers produced as response variable using the statistic software R (R Development Core Team, 2008).

Results

We first report on general task performance, before we turn to the analysis of language use and sketch accuracy in relation to spatial ability and time constraints. Participant numbers did not allow us to test systematically for gender differences. The data from our five male participants were consistently within the overall scope of results.

Task performance: Object identification

Participants did not have much difficulty discriminating objects from rubble. Only 12.6% of clicks made under time pressure and 29% of clicks made without time pressure hit non-objects. The tennis ball was found and remembered by all participants (see Table 1). As expected, participants without time pressure found significantly more objects ($M = 6.9$) than participants under time pressure ($M = 4.9$; $t(28.89) = 5.15$, $p < .0001$), and also remembered more objects ($M = 5.4$) than participants under time pressure ($M = 3.9$; $t(25.57) = -3.7$, $p < .01$).

Moreover, participants under time pressure consistently conceived of object visibility as worse ($M = 4.1$, $t(3.5)$, $p < .001$) than participants without pressure ($M = 3.2$) for most objects (see Table 2 for mean ratings per condition). Specifically, participants without time pressure rated the table significantly better ($M = 4.1$) than participants with time pressure ($M = 5.2$), $t(1.8)$, $p < .05$; likewise for the bauble ($M = 2.3$ without;

$M = 3.8$ with time pressure, $t(2.3)$, $p < .05$), the cube ($M = 1.8$ vs. $M = 2.2$, $t(2.5)$, $p < .05$), and (marginally) the lamp ($M = 2.2$ vs. 4.1, $t(2.9)$, $p < .01$). Consistently across groups, best marks were given for the tennis ball and the lowest marks were given for the piano and the glue stick. These results indicate that time pressure affects perceived visibility, and that the objects in this study were consistently perceived as differing widely and systematically with respect to visibility (as planned).

Order of mention

Across both conditions, the salient tennis ball was mentioned first in 83% of cases. Visualisation of the trajectories of object localisation in the verbal descriptions revealed that the order of experience was followed only in six cases. No other orderly patterns could be identified on this basis, such as circular or row-based trajectories as previously reported in studies of structured environments. Although all objects were positioned alongside the walls, there were only two references to object clusters at one wall.

Coherence and specificity

Linguistic analysis revealed considerable diversity in object localisation strategies. The description types specified and exemplified in the Appendix highlight the participants' creativity in establishing coherence. While some descriptions were consistently organised around the tennis ball as an anchor or the room area as a general frame, or following the film's trajectory, others employed combinations of these, or appeared generally incoherent. Crucially, there were no differences according to condition, as shown in Table 3.

While the diversity in establishing coherence is qualitatively illuminating, analysis of specificity lends itself more directly for quantification. We distinguished **specific** descriptions that locate the object's position within the configuration (either relative to an object, as in *to the left of the tennis ball*, or relative to the scene, as in *on the right side of the room*) from **unspecific** ones that pertain to features that could apply anywhere in the

overall scene (such as *at the wall*). Table 4 shows the number of objects per condition that were described using specific localisations relative to the overall scene or to another object, or (only) unspecific localisations. Since none of the objects were described relative to the overall scene as well as to another object, these categories are mutually exclusive. Again, the distributions do not differ significantly across conditions ($\chi^2 = 0.09$, $p > .1$). In both conditions specific localisations amount to roughly two thirds of the localisations.

Table 5 shows the number of objects for which an axis or a combination of axes was encoded linguistically. Most object locations were described by reference to only one axis (most often the vertical, followed by the lateral). For one quarter of the objects in the time-pressure condition and one third in the no-pressure condition, reference was based on two axes (mostly a combination of vertical and lateral). The distribution of axes did not differ significantly across conditions ($\chi^2 = 0.5$, $p > .4$).

Drawings

Participants either sketched the scene as viewed from above, or used other perspectives that allowed them to incorporate some of the vertical information that was relevant in the verbal descriptions as just shown. Accuracy varied from near-complete schematic and vivid 3D visualizations to scarce impressions of single objects. All participants represented the location of at least the snapshot view onto the back wall, which they had been facing repeatedly without rotation. Many sketches give little indication of object relations to the global environment (see Figure 4), while showing some traces of a local environment. Based on our rating scheme that allowed for a total of 13.75 possible points, drawings scored a mean of 5.5 points (SD 1.6) and ranged between 1.5 and 10.75 points. Participants without time pressure generated significantly better drawings ($M = 6.6$) than those with time pressure ($M = 4.5$) ($t(26.4) = -3.3$, $p < .01$), as expected due to the higher number of objects found without time pressure.

Relations between factors

The generalised linear model revealed that spatial ability affected the use of specific markers; participants who scored higher on the Santa Barbara Solids test also produced more specific markers ($p < .05$). Moreover, participants who drew better sketches also produced more specific markers ($p < .05$), and this explained the variance significantly better than spatial ability ($\chi^2 = 1.59$, $p < .001$, AIC = 446). This simple model was the best fit. Including condition did not lead to a better model fit, and since spatial ability and sketch accuracy correlated with each other, including both variables did not lead to better model fit (AIC values > 446).

Discussion

Searching for people trapped in the debris of a collapsed building is a matter of life and death, and therefore puts a high amount of pressure on USAR workers. Apart from the extreme time pressure and many other challenges in such a scenario, workers may have to deal with a highly unstructured and potentially distorted image conveyed by a camera whose movement is not easily accessible to intuitive understanding. In our study we focussed on the ability and strategies used by untrained humans to make sense of a cluttered scene containing objects, conveyed by a film created with state-of-the-art USAR endoscope technology. We were particularly interested in the effects of time pressure and spatial skill on participants' understanding, as reflected in the coherence and specificity of verbal descriptions, as well as the quality of drawings.

Our findings show that time pressure affected the number of objects found and consequently the quality of drawings, since number of objects found enhanced the score. However, crucially, time pressure did not affect the quality of descriptions, as measured (qualitatively) by coherence and (quantitatively) by specificity. Instead, the main decisive factor affecting participants' performance, both for specificity and sketch quality, was

spatial skill as assessed by the Santa Barbara Solids Test (Cohen & Hegarty, 2012). Our prediction that this test would suitably measure individual ability to conceptualise complex three-dimensional relationships from a two-dimensional representation was thus borne out.[4] In the following, we will take a closer look at the participants' strategies and performance in creating descriptions before returning to the influence of time and skill.

Coherence Usually, humans easily find structure in the environment to use for organization of descriptions. Verbal descriptions of spatial layouts (of any kind, judging from the literature so far) are highly coherent and represent the speaker's underlying conceptualisation of the configuration at hand. This includes the underlying perspective, the understanding of functional relationships between objects, and an overall trajectory reflecting the 'gestalt' of the scene or the way in which it is experienced (Ehrich & Koster, 1983; Grenoble, 1995; Levelt, 1982; Shanon, 1984; Tenbrink et al., 2011). The multiple perceptual challenges involved in our scenario made the structure of the scene much harder to grasp; furthermore, participants lacked experience in this particular type of description task. This to some extent explains the high variety in description types that we found in our data.

Nevertheless, participants could have relied on the order of acquisition, resembling earlier findings both for spatial object configuration descriptions (Buhl, 1996; Taylor & Tversky, 1992) and more generally in spatial experience (Gander, 2004). The fact that speakers tend to report events in the order they happened has become one of the classic Gricean Maxims (Be Orderly) (Grice, 1975), and appears so self-evident that temporal markers like *before* and *after* are not needed to understand the order of events (Anderson, 1980; Tenbrink & Schilder, 2003). Strikingly however, judging from visual inspection of the trajectory in participants' descriptions, only 6 out of 36 participants relied on experience by reporting the objects in the order they had observed them in the film. The fact that

most of our speakers refrained from making use of this simple practice to represent the spatial scene is a clear indicator of the difficulty in establishing an integrated conceptual representation from the confusing visual input conveyed by the camera movement.

Instead, all participants except one mentioned the tennis ball first, in spite of the fact that it was not the object closest to the observer in the film. This object was the one uniformly rated as most visually accessible, and therefore salient. Thus, it served as an excellent landmark in the cluttered scenario. The strategy of orienting at salient landmarks is well known in the literature, particularly when navigating routes in complex environments (Caduff & Timpf, 2008). When describing individual objects in small scale space, salient objects are frequently taken as a relatum to describe less salient objects (Talmy, 1983).

Apart from the unanimous choice of a salient landmark relatum as a starting point, the description types we identified in the data resemble previously identified strategies of describing spatial scenes only to a limited extent. The barely used strategy of following the camera's trajectory explicitly is similar to a route description (a continuous description with changing viewpoints following a path through the scene), as investigated widely for outdoor environments and also identified for apartment descriptions (Linde & Labov, 1975). Apart from that, some participants adopted a static view on the scene and described objects in relation to other objects; this is comparable to the 'gaze tour' (a continuous description from a fixed viewpoint) previously identified, for instance, by Ullmer-Ehrich (1979) and Ehrich and Koster (1983). However, temporally and sequentially consistent strategies were rather exceptional in our data. Clearly, continuous gaze tour and route description strategies presuppose a completely integrated conceptual representation of the scene—and this was hard to obtain in the present scenario.

The lack of continuity in the participants' conceptual representation has several implications, both for memory and for description. Since only eight objects were

contained in the scene, it should have been possible to recall all objects found during navigation. However, this was not the case; descriptions consistently contained a lower number of objects than participants found during navigation. Generally, when memorizing lists of items, strategies include categorization and clustering of items (Gobet et al., 2001; Baddeley, 2003); it is well known that retrieval of objects is facilitated by a structuring strategy (Ericsson, Chase, & Faloon, 1980). The lack of coherence in the participants' internal representation may have prevented such a strategy. Also, the lack of a continuous trajectory in description poses major problems for communication. If there is no spatially ordered strategy of description that allows to infer the location of an object by reference to the previous one, object positions need to be specified individually. Using a trajectory along the walls, as in more structured environments (Ullmer-Ehrich, 1979; Shanon, 1984; Ehrich & Koster, 1983; Grenoble, 1995), would have reduced the effort of communicating. Moreover, it would have conveyed the form of the room itself, facilitating representation and comprehension on the part of the listener.

Despite the overwhelming lack of continuity, 15 out of 36 descriptions exhibited a degree of coherence by consistent reference to a single relatum, namely either the salient tennis ball or the overall scene. While consistency, in general, is a well-known feature of spatial descriptions (Vorwerg, 2009), constant reference to the same relatum has not to our knowledge been reported for scenarios involving gradual exploration rather than a full view on the scene. Furthermore, speakers generally prefer a good spatial relationship over saliency; when the spatial relationship to a salient item cannot be captured by a simple spatial term, they choose a less salient object as relatum (Carlson & Hill, 2009). Clearly, not all of the items in our scenario had a simple spatial relationship to the tennis ball. Nevertheless, for some people, the spatial relation to this conceptual anchor must have been the most accessible one under present circumstances. For others, it appeared to be easier to refer to the sides of the overall scene, relative to the observer. Both of these

strategies may have been supported by the camera's movement from the prominent middle position to the sides, and back.

Rather than enhancing a continuous description strategy, the presentation format therefore supported coherence in the spatial representation by focussing on one single conceptual frame of reference. However, this only accounts for less than half of the participants. Many failed to establish coherence of any kind, and described object locations on a primarily ad-hoc basis. Several people commented that it was hard to maintain orientation within the film; this challenge accordingly led to the ensuing lack of coherence in the descriptions.

Specificity. Complementing the coherence analysis, detailed analysis of object description specificity showed that about a third of the objects, on average, were described only locally without ever specifying their position relative to another object or within the room. Although providing a spatial description without actually specifying the location of objects is not a typical phenomenon known from the literature, it relates to the widely used notion of *landmark knowledge* as suggested by Siegel and White (1975). This type of knowledge means that the existence of objects or landmarks is known, but there is no clearly established concept of the relative location of these entities to each other. In a large-scale environment, this entails that humans will not know how to get from one of these landmark locations to another, and they will not be able to point to them. Again, this suggests an incomplete mental representation of the spatial configuration. Crucially, although the descriptions in our study differed with respect to specificity, this was unrelated to the amount of time available. Accordingly, there was no discernible qualitative shift that allowed to build up a more survey-like representation of the scene based on sufficient exposure. Those participants who managed to build up a spatial representation that included object position apparently did so from the start, for the

objects they could identify in the time available. This corresponds to previous findings showing early acquisition of integrated spatial knowledge (Ishikawa & Montello, 2006).

Furthermore, typically in room descriptions, the room's axes serve as reference for object localisation, as in "in the front right corner of the room." In our scenario, the camera moved towards the back wall, with the long sides of the rectangle on the left and right (relative to the observer). The objects were distributed symmetrically in the front, middle, and back portions of the room, directly at the outer walls. Accordingly, it could be expected that the lateral and frontal axes would be equally relevant for conveying object positions within the scene. Particularly in the absence of a continuous trajectory of description, this information would need to be made explicit for each object individually. Our results showed, however, that the axis mentioned most often was the *vertical* dimension. This specifies the objects' placement on the ground or elsewhere in the clutter, rather than relative to the planar shape of the room. While the lateral placement was also prominent, the frontal axis was rarely encoded. Also, combinations of lateral and frontal axes were rare.

Along with the lack of location specificity, this result highlights the participants' conceptualisation of the configuration. Although most objects were placed on the ground, this was not true for all of them, which may have made the vertical positioning more relevant than in less cluttered everyday scenarios. The potential contrast to the expected (functional) position may have led participants to emphasise this aspect; speakers are known to provide descriptions of just those aspects that they perceive to be *relevant* for the addressee (Sperber & Wilson, 1986). Furthermore, this dimension may have been clearer or more accessible to the participants than the other ones, as it did not require positioning the objects *specifically* within the overall scene. While using the lateral or frontal axes presupposes memory of the object's position, it appears that the vertical axis only indicated, in the present scenario, *how* an object was placed (rather than where).

Furthermore, the lateral axis appeared to be more accessible, or conceptually prominent, than the frontal one. It might have been harder to keep track of the frontal positioning than to encode the lateral sides by following the camera's movements. As a result, even those descriptions that included a localisation of an object relative to the overall scene by using a lateral term still remained vague, since they lacked the (clearly relevant) information about the frontal axis. The analysis of axes therefore serves as a further indicator of the lack of conceptual integration, here with respect to the relative distance to the observer as shown in the film.

The Influence of Time After having explored the strategies and limitations of spatial location descriptions in our study, it is now time to come back to the main factor that could be expected to hamper the participants' understanding in USAR scenarios—lack of time. We addressed this factor in our study by letting one group of participants explore the scene for as long as they wanted, while the other group of participants was given just enough time to watch the whole film, allowing for very little exploration. Surprisingly, beyond identifying and memorising a higher *quantity* of objects, these two groups did not differ at all with respect to the *quality* of their representations, as reflected in the findings reported so far. Neither coherence nor specificity (including reference to spatial axes) were affected by the conditions of time. Clearly, time is *not* the most decisive factor leading to a better comprehension of a cluttered spatial scene. Given more time, participants were able to find, report, and sketch a higher number of objects—but their descriptions did not become more coherent or specific.

What, then, is the most decisive influencing factor that accounts for the fundamental performance differences in our scenario? Our findings suggest that the participants themselves bring in (or lack) the skills needed to deal with a complex spatial challenge. Performance in the Santa Barbara Solids Test (Cohen & Hegarty, 2012)

correlated with production of specific spatial markers to indicate object position. Again, time pressure did not matter. Those participants who drew more accurate sketches also produced more specific language, regardless of time constraints; this correlation is clearly due to the participants' mental representation of the scene. Importantly, these were also the participants who scored better at the Solids Test. Thus, spatial skill—rather than availability of time—leads to better cognitive integration of spatial information, allowing the participants to draw good sketches and to produce more specific (and coherent) spatial language.

Conclusion and Outlook

Urban Search and Rescue involves substantial challenges for the workers, who (inter alia) need to understand complex and distorted spatial configurations under considerable time pressure. We addressed untrained observers' strategies to deal with a scenario of this kind, and addressed the relative effects of time pressure and spatial skill on the quality of linguistic representations and sketches, of the kind that could be used to communicate observations relevant to an emergency situation.

Our study reveals that time will enable observers to identify a higher number of objects, but not necessarily to develop a better comprehension of the spatial configuration. This result supports the idea that spatial knowledge is acquired by continuously (or quantitatively) adding information (Montello, 1998), rather than as a discrete process that involves a major conceptual (or qualitative) shift allowing for conceptual integration in the sense of Siegel and White (1975). In our context, this means that objects will be identified eventually, but their actual location will remain hard to comprehend and convey to others, no matter how much time is available. The main decisive factor supporting a better understanding is the skill of the observer. In our study, none of the participants were trained for USAR scenarios; however, they differed with respect to their spatial

skills—and this turned out to be decisive for their performance. Clearly, it matters what kinds of skills workers bring into an USAR emergency. Fortunately, spatial skills can be trained to a considerable extent, as shown by a large body of research on spatial learning (Uttal et al., 2013). Since USAR workers receive a fair amount of specific training on their work, this should diminish interindividual differences due to relevant practice (Hegarty et al., 2007). Our research underscores the vital importance of such training, along with the insight that conceptual integration (and, on this basis, efficient communication) can be achieved by skilled workers even under time pressure.

Necessitated by the originality of its research target, our study was explorative and in part qualitative in nature. We specifically devised a coding scheme to assess sketch drawing performance, and identified and operationalised description types that provided insights into coherence. These achievements now enable future studies based on specific hypotheses and pre-defined measures. To allow for more direct and finer-grained sketch quality judgements, participants could be asked to use a particular perspective for drawing rather than using their own preferences. Additional insights can be gained by using eye movement data to reveal cognitive focus during scene comprehension, which typically relates systematically to memory performance (Williams et al., 2005). Moreover, gaze behaviour has been shown to differ between experts and non-experts, in particular under time constraints, in various contexts (Vickers, 2011). Triangulation with eye movement data would therefore illuminate effective cognitive strategies of coping with the challenging USAR scenario.

Future research should also target various design related aspects. The USAR scenario is conceptually demanding in many different ways, due to visual clutter, orientation changes, camera turns, and more. A refined design could disentangle the contribution of each of these factors, pointing to the most urgent improvements that need to be made to USAR visualisation technology. In line with the scope of previous object

location research, our participants were students untrained in the USAR scenario that motivated our study. In future studies, trained and experienced USAR workers will need to be compared with untrained participants, matched for age, general spatial ability, and (ideally) verbal skills, so as to address the benefits of previous USAR experience and training directly.

For recruitment reasons beyond our control, most of our participants were females. While the male participants were no outliers, the patterns in our results may still primarily represent the female gender. Given the abundant literature on gender biases particularly in the field of object location memory (Voyer et al., 2007), this opens up further questions for future research.

Finally, studies will be needed to address more precisely the effects of time constraints, speed (frame rate), and anxiety. In our study, the identified patterns might be due to differences in amount of time, or to default speed, or to differences due to knowledge of limited or unlimited time, or anxiety about time. Actual USAR situations furthermore impose considerable anxiety since lives are at stake. Our study showed that, given more time, more objects could be identified and represented. Surprisingly however, qualitatively the mental representations did not improve. Added psychological pressure might hamper comprehension of the spatial configuration more substantially, which would further add to the challenges of the rescue situation.

Acknowledgements

This research was supported by the SFB/TR8 Spatial Cognition (Deutsche Forschungsgemeinschaft, DFG), projects I6-[NavTalk] and A7-[FreePerspective]. We thank Tobias Hammer for constructing the camera servo mechanism.

References

Allen, G. L., Kirasic, K. C., Dobson, S. H., Long, R. G., & Beck, S. (1996). Predicting environmental learning from spatial abilities: An indirect route. *Intelligence*, *22*, 327–355.

Anderson, J. (1980). *Cognitive psychology and its implications*. New York: W. H. Freeman.

Andonova, E., Tenbrink, T., & Coventry, K. (2010). Function and context affect spatial information packaging at multiple levels. *Psychonomic Bulletin & Review*, *17*(4), 575–580. doi: 10.3758/PBR.17.4.575

Baddeley, A. D. (2003). Working memory: looking back and looking forward. *Nature Reviews: Neuroscience*, *4*, 829–839.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using eigen and s4.* Retrieved from `http://cran.r-project.org/package=lme4` (R package version 1.1-7)

Beun, R. J., & Cremers, A. (2001). Multimodal reference to objects. In H. C. Bunt & R. J. Beun (Eds.), *Cooperative multimodal communication* (pp. 64–86). Berlin: Springer.

Brunyé, T. T., Rapp, D. N., & Taylor, H. A. (2008, August). Representational flexibility and specificity following spatial descriptions of real-world environments. *Cognition*, *108*(2), 418–43.

Buhl, H. M. (1996). Erwerbssituation, mentale Repräsentation und sprachliche Lokalisationen - Blickpunktinformation als Bestandteil der Raumrepräsentation (Acquisition situation, mental representation, and verbal localisations - viewpoint information as a part of the spatial representation). *Sprache und Kognition*, *15*(4), 203–216.

Caduff, D., & Timpf, S. (2008). On the assessment of landmark salience for human navigation. *Cognitive Processing*, *9*(4), 249–267.

Carlson, L. A., & Hill, P. L. (2009). Formulating spatial descriptions across various dialogue contexts. In K. Coventry, T. Tenbrink, & J. Bateman (Eds.), *Spatial language and dialogue* (p. 89-103). Oxford: Oxford University Press.

Casper, J., & Murphy, R. R. (2003). Human-robot interactions during the robot-assisted Urban Search and Rescue response at the World Trade Center. *IEEE Transactions on Systems, Man, and Cybernetics: a publication of the IEEE Systems, Man, and Cybernetics Society*, *33*, 367–385.

Cohen, C. A., & Hegarty, M. (2012). Inferring cross sections of 3D objects: A new spatial thinking test. *Learning and Individual Differences*, *22*(6), 868–874.

Couclelis, H., Golledge, R. G., Gale, N., & Tobler, W. (1987). Exploring the anchor-point-hypothesis of spatial cognition. *Journal of Environmental Psychology*, *7*, 99–122.

Ehrich, V., & Koster, C. (1983). Discourse organization and sentence form: The structure of room descriptions in Dutch. *Discourse Processes*, *6*(2), 169–195.

Ekstrom, R., French, J., Harman, H., & Dermen, D. (1976). *Manual for kit of factor-referenced cognitive tests.* Princeton, NJ: Educational Testing Service.

Ericsson, K. A., Chase, W. G., & Faloon, S. (1980). Acquisition of a memory skill. *Science*, *208*(4448), 1181–1182.

Evans, G. W., & Pezdek, K. (1980). Cognitive mapping: Knowledge of real-world distance and location information. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 13–24.

Gander, P. (2004). Spatial mental representations in interactive fiction - what is particular about the interactive text? In S. Porhiel & D. Klingler (Eds.), *L'unite texte* (p. 96-124). Pleyben, France: Perspectives.

Geiser, C., Lehmann, W., & Eid, M. (2006). Separating rotators from non-rotators in the mental rotations test: A multigroup latent class analysis. *Multivariate Behavioral Research*, *41*(3), 261–293.

Gobet, F., Lane, P. C. R., Croker, S., Cheng, P. C.-H., Jones, G., Oliver, I., & Pine, J. M. (2001). Chunking mechanisms in human learning. *Trends in Cognitive Sciences*, *5*(6), 236–243.

Grenoble, L. (1995). Spatial configurations, deixis and apartment descriptions in russian. *Pragmatics*, *5*(3), 365–385.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics 3: Speech acts* (pp. 64–75). New York: Academic Press.

Hamp, Q., Gorgis, O., Labenda, P., & Neumann, M. (2013). Study of efficiency of USAR operations with assistive technologies. *Advanced Robotics*, *27*(5), 337–350.

Hegarty, M., Keehner, M., Cohen, C. A., Montello, D. R., & Lippa, Y. (2007). The role of spatial cognition in medicine: Applications for selecting and training professionals. In G. L. Allen (Ed.), *Applied spatial cognition: From research to cognitive technology* (p. 285-315). Mahwah, NJ: Lawrence Erlbaum.

Hegarty, M., Montello, D. R., Richardson, A. E., Ishikawa, T., & Lovelace, K. (2006). Spatial abilities at different scales: Individual differences in aptitude-test performance and spatial-layout learning. *Intelligence*, *34*, 151–176.

Hermer-Vazquez, L., Moffet, A., & Munkholm, P. (2001). Language, space, and the development of cognitive flexibility in humans: the case of two spatial memory tasks. *Cognition*, *79*(3), 263-299.

Herrmann, T., & Grabowski, J. (1994). *Sprechen. Psychologie der Sprachproduktion.* Heidelberg: Spektrum der Psychologie.

Hertzberg, C., Wagner, R., Birbach, O., Hammer, T., & Frese, U. (2011). Experiences in building a visual SLAM system from open source components. In *Proceedings of the*

IEEE International Conference on Robotics and Automation (ICRA), Shanghai, China.

Ishikawa, T., & Montello, D. R. (2006). Spatial knowledge acquisition from direct experience in the environment: individual differences in the development of metric knowledge and the integration of separately learned places. *Cognitive Psychology*, *52*(2), 93–129.

Klabunde, R. (1999). Logic-based choice of projective terms. In *KI-99: advances in artificial intelligence, 23rd annual german conference on artificial intelligence, bonn, germany, september 13-15, 1999, proceedings* (pp. 149–158).

Körner, C., & Gilchrist, I. D. (2008). Memory processes in multiple-target visual search. *Psychological Research*, *72*(1), 99-105.

Krippendorf, K. (2004). *Content analysis: an introduction to its methodology.* London and Thousand Oaks, CA: Sage.

Levelt, W. J. M. (1982). Linearization in describing spatial networks. In S. Peters & E. Saarinen (Eds.), *Processes, beliefs, and questions* (pp. 199–220). D. Reidel.

Linde, C., & Labov, W. (1975). Spatial networks as a site for the study of language and thought. *Language*, *51*(4), 924–939.

McNamara, T. P. (2003). How are the locations of objects in the environment represented in memory? In *Spatial cognition iii* (pp. 174–191). Berlin, Heidelberg: Springer.

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

Montello, D. R. (1993). Scale and multiple psychologies of space. In A. U. Frank & I. Campari (Eds.), *Spatial information theory: A theoretical basis for GIS* (pp. 312–321). Berlin: Springer.

Montello, D. R. (1998). A new framework for understanding the acquisition of spatial knowledge in large-scale environments. In M. J. Egenhofer & R. G. Golledge (Eds.),

Spatial and temporal reasoning in geographic information systems (p. 143-154). New York: Oxford University Press.

Montello, D. R., Lovelace, K. L., Golledge, R. G., & Self, C. M. (1999). Sex-related differences and similarities in geographic and environmental spatial abilities. *Annals of the Association of American Geographers*, 515–534.

Plumert, J., Carswell, C., de Vet, K., & Ihrig, D. (1995). The content and organization of communication about object locations. *Journal of Memory and Language*, *34*, 477–498.

R Development Core Team. (2008). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org` (ISBN 3-900051-07-0)

Radvansky, G., & Copeland, D. (2000). Functionality and spatial relations in memory and language. *Memory & Cognition*, *28*(6), 987-992.

Schober, M. F. (1995). Speakers, addressees and frames of reference: Whose effort is minimized in conversations of location? *Discourse Processes*, *20*, 219–247.

Shanon, B. (1984). Room descriptions. *Discourse Processes*, *7*(3), 225–255.

Sholl, M. J. (1987). Cognitive maps as orienting schemata. *Journal of Experimental Psychology: Learning, memory, and cognition*, *13*(4), 615–28.

Siegel, A., & White, S. (1975). The development of spatial representations of large-scale environments. In H. Reese (Ed.), *Advances in child development and behavior* (pp. 10–55). New Yord: Academic Press.

Sperber, D., & Wilson, D. (1986). *Relevance: communication and cognition*. Oxford: Blackwell.

Talmy, L. (1983). How language structures space. In J. H. L. Pick & L. P. Acredolo (Eds.), *Spatial orientation: Theory, research and application* (pp. 225–282). New York: Plenum Press.

Talmy, L. (2000). *Towards a cognitive semantics*. Cambridge, MA: A Bradford Book, MIT Press.

Taylor, H. A., Naylor, S. J., & Chechile, N. A. (1999, March). Goal-specific influences on the representation of spatial perspective. *Memory & Cognition*, *27*(2), 309–19.

Taylor, H. A., & Tversky, B. (1992). Spatial mental models derived from survey and route descriptions. *Journal of Memory and Language*, *31*, 261–292.

Tenbrink, T. (2011, February). Reference frames of space and time in language. *Journal of Pragmatics*, *43*(3), 704–722. doi: 10.1016/j.pragma.2010.06.020

Tenbrink, T. (2015). Cognitive Discourse Analysis: Accessing cognitive representations and processes through language data. *Language and Cognition*, *7*(1), 98–137.

Tenbrink, T., Coventry, K. R., & Andonova, E. (2011). Spatial strategies in the description of complex configurations. *Discourse Processes*, *48*, 237–266.

Tenbrink, T., & Schilder, F. (2003). (Non)temporal concepts conveyed by before, after, and then in dialogue. In P. Kuehnlein, H. Rieser, & H. Zeevat (Eds.), *Perspectives on dialogue in the new millennium* (p. 353-380). Amsterdam: John Benjamins.

Thorndyke, P. W., & Hayes-Roth, B. (1982). Differences in spatial information acquired from maps and navigation. *Cognitive Psychology*, *14*, 560–581.

Tomko, M., & Winter, S. (2009, February). Pragmatic construction of destination descriptions for urban environments. *Spatial Cognition & Computation*, *9*(1), 1–29.

Ullmer-Ehrich, V. (1979). Wohnraumbeschreibungen (Living space descriptions). *Zeitschrift für Literaturwissenschaft und Linguistik*, *33*(9), 58–84.

Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C., & Newcombe, N. S. (2013). The malleability of spatial skills: A meta-analysis of training studies. *Psychological Bulletin*, *139*(2), 352-402.

Vandenberg, S. G., & Kuse, A. R. (1978). Mental rotations, a group test of three-dimensional spatial visualization. *Perceptual and Motor Skills*, *47*(2), 599-604.

Vickers, J. N. (2011). Mind over muscle: the role of gaze control, spatial cognition, and the quiet eye in motor expertise. *Cognitive Processing*, *12*(3), 219-222.

Vorwerg, C. (2009). Consistency in successive spatial utterances. In K. Coventry, T. Tenbrink, & J. Bateman (Eds.), *Spatial Language and Dialogue* (pp. 40–55). Oxford: Oxford University Press.

Voyer, D., Postma, A., Brake, B., & Imperato-McGinley, J. (2007). Gender differences in object location memory: A meta-analysis. *Psychonomic Bulletin & Review*, *14*(1), 23-38.

Williams, C. C., Henderson, J. M., & Zacks, R. T. (2005). Incidental visual memory for targets and distractors in visual search. *Perception & Psychophysics*, *67*(5), 816-827.

Wolbers, T., & Hegarty, M. (2010). What determines our navigational abilities? *Trends in Cognitive Sciences*, *14*(3), 138-46.

Appendix

Coherence

The following description types were found in our data. We include transcripts exemplifying the main types.

Anchor. Criterion: The location of most objects is described relative to *a single salient relatum* (which was always the tennis ball). Linguistically, typical spatial markers for this description type are relational terms with a specific object relatum, such as 'it was to the right of the tennis ball.'

Example transcript. Also wenn man vom Ausgangs(bild) geht, dann war direkt in der Mitte ein Tennisball eingeklemmt. Und daneben ganz versteckt war noch irgendwas silbernes, mattes, keine Ahnung. Und ähm ein bisschen weiter rechts lag ähm die Öffnung

von so einer Pfanddose. Dann ähm lag auf der linken Seite von dem Tennisball auf jeden Fall dieses ähm Wüf- dieser Würfel, den man drehen kann. ähm was hatte ich denn noch? Genau, vor dem Tennisball lag auch noch meiner Meinung nach sowas Ähnliches wie ein I-Pod oder irgendwas Flaches, dann. Genau, die Lampe, die war ähm äh auch ein Stück rechts von dem Tennisball.

English translation: *If you look from the start view, there was the tennis ball squeezed in in the middle. And next to it completely hidden was something silvery, matt, I don't know. And uhm, a little further to the right was the bottle opener. Then, uhm, to the left side of the tennis ball was this cube, that you can turn. Uhm what else did I have? Yes, in front of the tennis ball there was in my opinion something like an I-Pod or something flat. And yes, the lamp, it was uhm, uh, also a little to the right of the tennis ball.*

Room. Criterion: Most objects are described relative to *areas of the scene*, i.e., the 'room'. Spatial markers are internal projective terms, such as the German adverbs *vorne* (in the front), *hinten* (in the back), *oben* (in the top), and *unten* (below), as well as *rechts* when used without an object as relatum. As internal terms, adverbs such as these presuppose an encompassing relatum such as a room or overall scene (Tenbrink, 2011; Klabunde, 1999). Distance terms such as 'farther away' and 'further towards me' work similarly in that they specify the position inside the room relative to the speaker.

Example transcript. Also am klarsten und deutlichsten war eigentlich der Tennisball, der genau geradezu war. Den man direkt gesehen hat, wie er da eingeklemmt war. Und dann als ich endlich verstanden habe wie dieser Film funktioniert, weiß ich, war auf der vorne also sehr weit vorne direkt auf der rechten Seite muss irgendwo so eine kleine Stehlampe so eine Tischlampe gestanden haben. Und auf der linken Seite vorne war dieser kleine Würfel als Schlüsselanhänger, dieser Zauberwürfel oder wie sagt man dazu, ich weiß es nicht. Ähm dann gehts ja eigentlich immer weiter auf diesen Tennisball zu und

immer mal wieder nach links und rechts und ich glaube ganz hinten links war noch sowas wie ein so ein, ich weiß nicht, ob es ein Parfumflakon war. Und ich sag mal auf dieser Geraden ohne Abzubiegen lag weiter vorne noch sowas wie ein so eine Lasche von einem Dosenöffner. Ähm und dann war noch weiter also wieder ganz hinten rechts sowas ich weiß nicht obs eine Taschenlampe war. Aber der Tennisball der war sehr offensichtlich ähm ja wenn es erstmal klick gemacht hat, wars eigentlich recht verständlich.

English translation: *Well, the clearest was the tennis ball that was straight ahead. Which you could directly see, the way it was squeezed in there. And then, when I had finally understood how the film functioned, I know that at the front, well, at the very front directly on the right side there must have been some kind of little lamp, kind of a table lamp. And on the left side at the front there was this little cube which was a key ring, this magic cube or how do you say that, I don't know. Uhm, then it went straight on to the tennis ball and left and right in between, and I believe at the very back and left there was something like, I don't know, whether it was a perfume flask. And I'll say on this straight line without turns there was further to the front something else like a bottle opener. Uhm and then there was even further on, well, again at the furthest back and to the right, something I don't know if it was a flashlight. But the tennis ball, that was very obvious. Uhm, well, if you finally got it, then it was quite comprehensible.*

View to wall. Criterion: The view into the scene towards the back wall is consistently used as basis for description, either explicitly or implicitly, without using either the ball or the room as a consistent relatum.

Progression. Criterion: Most objects are described following the *trajectory of the camera* through the room. Typical markers are ordinals such as *als erstes, als zweites* (firstly, secondly), motion verbs such as *gehen, abbiegen* (walk, turn), temporal references to the film such as *am Anfang, am Ende* (beginning, end), direction terms such as *nach*

rechts (turn right), and sequential markers such as *dann, danach* (then, after that) (with corresponding contextual meaning; e.g., unordered recall such as 'and then I also had the lamp' was not coded as sequential). Descriptions of the camera's movement—or (metaphorically) the speaker's movement, as in 'the first room I went into'—also reveal temporal structure. By using motion and temporal markers, participants conveyed a coherent linear discourse organization that places objects on a time line. The combination of temporal markers with the *direction* of motion yields a relatively clear spatial localisation of objects.

Example transcript. Als erstes habe ich die Lampe gesehen, die stand so halb hinter einer Styroporplatte, ziemlich am Anfang. Dann glaube ich habe ich als nächstes den Ball gesehen, der war mitten drin in dem Gerüst. Äh, dann der Flaschenöffner war so angelehnt. Die Figur, so eine kleine schwarze, wenn das eine war, war so zwischen Platten so unten drunter. Ich weiß nicht wie man das sagt. Die Christbaumkugel hing glaube ich an einer Styroporplatte. Dieses weiße undefinierbare Ding, was da war ganz zum Schluss, stand auf einer Styroporplatte. Dann war da noch so 'n so wie so 'ne so'ne so'ne Röhre mit an den Enden so weiß und mit glaube ich noch eine Spirale drin. Die war zwischen Platten geklemmt. Dann was gab es noch? Einen Stift glaube ich der war ziemlich im Hintergrund. So schräg hat die Kamera ab und zu auch mal gewechselt. Also ich denke war schräg. Was gab es dann noch? Ah dieses diesen Schlüsselanhänger. Der war so man kann ihn ja so verstellen. Und der war halt so verstellt. Und lag da halt so in den Trümmern.

English translation: *At first I saw the lamp, it stood half behind a styrofoam board, right at the beginning. Then I think I saw the ball next, it was in the middle of the rubble. Uh, then the bottle opener was sort of reclined. The figurine, a kind of small black one, if it was one, was between boards, kind of down below. I don't know how to say that. The bauble, I think, was hanging on a styrofoam board. This white undefinable thing, which was there at the very end, stood on a styrofoam board. Then there was also a kind of like*

kind of tube with kind of white at its ends and with I think a spiral in it. That was squeezed between the boards. What else was there? A pen I believe was quite in the background. Like aslant, the camera was changing sometimes. Yeah, I think it was aslant. What else was there? Ah, this this key ring. It was like you can kind of distort it. And it was distorted. And it just lay there in between the rubble.

Collage. This category captures descriptions that did not fall into one of the preceding types, based on our definitions, but contained more than three objects. Unspecific localisations are frequent in this category (to be detailed below).

Example transcript. An den Ball, weil der ganz oft kam. Der war so zwischen den Steinen eingequetscht. Und dann war davor auf dem Boden war so eine Radkappe und ich überleg grad wie sich das gedreht hat. Das war im Uhrzeigersinn gedreht, und dann oben da lag noch so ein Bieröffner. Und was hab ich denn noch gesehen? Ach so genau diesen Drehwürfel. Ich weiß gar nicht wie die heißen. Aber wo der lag, weiß ich nicht mehr.

English translation: *(I remember) The ball, because it was often displayed. It was kind of squeezed in between the stones. And then in front of it on the ground there was a kind of hubcap and I am just trying to remember how it turned. It turned clockwise, and then on the top there was a bottle opener. And what else did I see? Right, this cube. I don't know what they are called. But where it was, I don't know.*

Sparse. This category captures descriptions that contained only up to three objects.

Specificity

Specific descriptions typically used a projective term such as *left, right, in front of, behind* together with a relatum that was specific enough to determine the object's location, such as the overall scene (or room), or another object. The following linguistic categories were identified as relevant markers and annotated as **specific**:

Room-related terms like *hinten, rechts, oben* (back, right, at the top),

Relational projective terms like *hinter, rechts von, ueber* (behind, to the right of, above) and **projective direction terms** like **nach rechts, geradeaus** (to the right, straight on) could be specific depending on the relatum they used,

Distance terms relating an object to the speaker, and

References to the 'center' of the room or scene.

In contrast, the following linguistic categories were identified as relevant markers and annotated as **unspecific**:

Topological terms specify containment and support relations with the immediate environment, e.g., *in, auf, an* (in, on, at),

Other spatial terms that are independent of a reference frame and do not fall into any of the other categories, such as *neben, in der Naehe* (next to, near),

Relational projective terms with unspecific relata such as 'a stone', 'rubble', 'the ground', 'the wall'.

Footnotes

[1]As indicated above, visual accessibility of objects may affect conceptual and linguistic object localisation (Talmy, 2000)

[2]An analysis of participants' behaviour in this regard was unfortunately outside the scope of this study, although it would be interesting to look at their preferred speed in each condition, and their choices in backing up to revisit parts of the scene.

[3]Utterances were defined based on their syntactic form. Each main clause was considered one utterance, including elliptical or more complex ones that contained one or more subclauses.

[4]Alternatively, as pointed out by a reviewer, a more general skill could lead to better performance in both kinds of task.

Table 1

Number of participants who found a particular object during navigation or mentioned it in the description

	Navigation				**Description**			
Time pressure:	Yes		No		Yes		No	
Object name	#	%	#	%	#	%	#	%
Tennis Ball	19	100.0	17	100.0	19	100.0	17	100.0
Bottle Opener	17	89.5	17	100.0	13	68.4	14	82.4
Cube	15	78.9	17	100.0	12	63.2	17	100.0
Glue Stick	13	68.4	13	76.5	9	47.4	14	82.4
Bauble	10	52.6	14	82.4	9	47.4	12	70.6
Table	6	31.5	6	35.3	4	21.1	6	35.3
Lamp	5	26.3	14	82.4	3	15.8	10	58.8
Piano	5	26.3	6	35.3	1	5.3	3	17.6

Table 2

Object visibility ratings per condition

Mean ratings: 1= excellent, 6=very bad

	bottle opener		piano		glue stick		tennis ball	
	lamp		cube		bauble		table	
Time pressure	4.1**	3.2	2.2*	5.6	3.9*	5.1	5.2*	1.2
No pressure	1.2**	2.4	1.2*	5.7	2.3*	4.6	4.1*	1.0

$^{**}p < 0.01$, $^{*}p < 0.05$

Table 3

Description types

	Anchor	Room	View to wall	Progression	Collage	Sparse
Time pressure	4	4	2	1	6	2
No pressure	3	4	2	2	5	1

Table 4

Specificity of Object Localisations

	scene-based	object-based	unspecific
Time pressure	32	19	34
No pressure	40	24	39

Table 5

Distribution of Axes: l = lateral, f = frontal, v = vertical

	l	f	v	l+v	l+f	v+f	l+v+f
Time pressure	19	6	33	16	2	1	0
No pressure	16	7	39	15	8	4	1

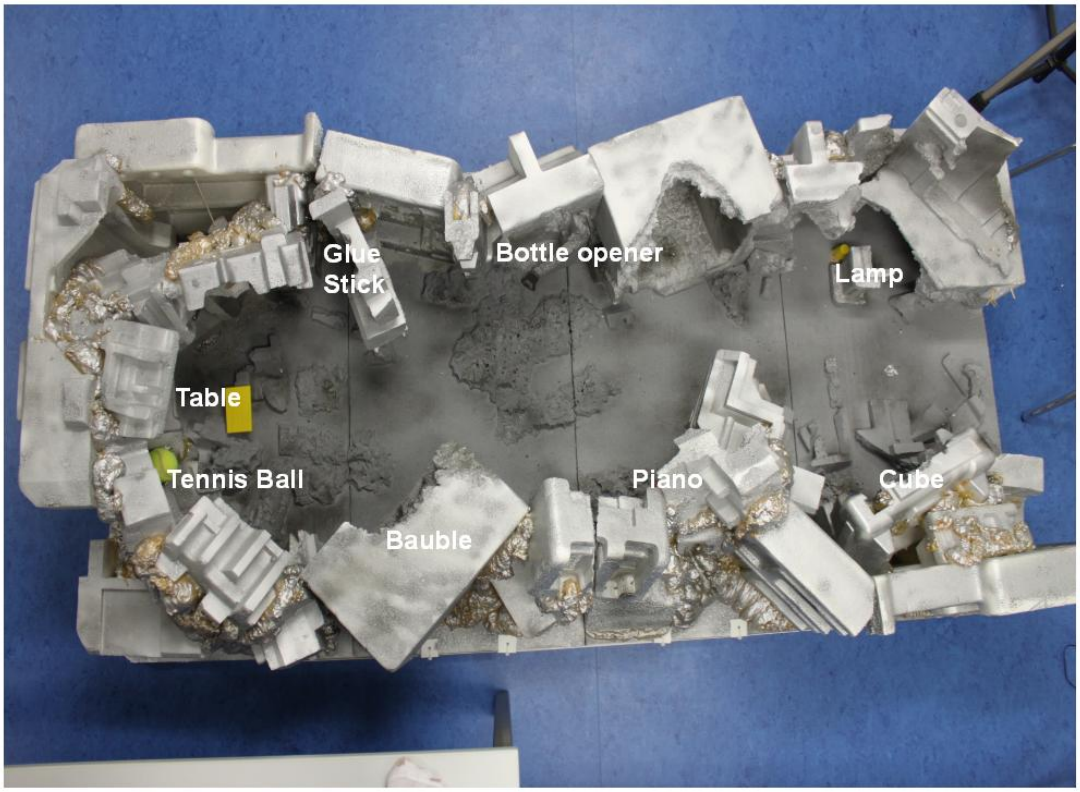
Figure Captions

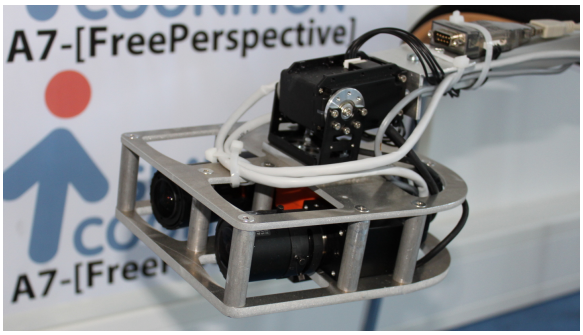
Figure 1. Location of objects inside the mock-up.

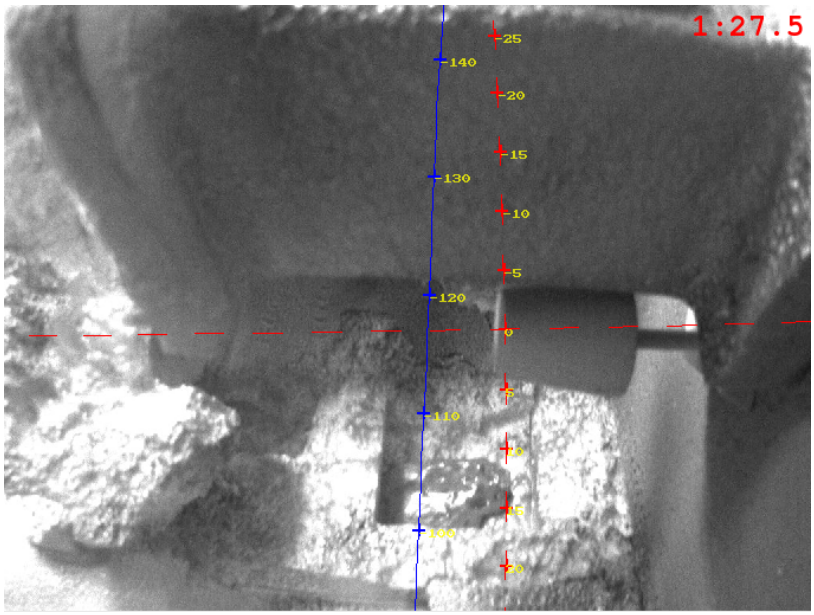
Figure 2. Stereo camera head with inertial measurement unit (orange, partially hidden) mounted to a pole via a servo motor.

Figure 3. Screenshot showing tilted view with artificial horizon, with one non-salient object (the lamp) on the right hand side.

Figure 4. Example room sketch: no indication of object's relation to the global environment.


Glue Stick
Bottle opener
Lamp
Table
Tennis Ball
Piano
Cube
Bauble


A7-[FreePerspective]


1:27.5
25
140
20
15
130
10
5
120
0
110
100