CHOOSING FRAMES OF REFERENECE: PERSPECTIVE-TAKING IN A 2D AND 3D NAVIGATIONAL TASK

Farilee E. Mintz
ITT Industries, AES Division
Alexandria, VA

J. Gregory Trafton, Elaine Marsh, & Dennis Perzanowski
Naval Research Laboratory
Washington, DC

This study investigates how frames of reference are chosen in a dynamic navigational task. Participants issued verbal instructions to an animated robot and were provided with one of three views for navigating the animated robot around a virtual world. The different views included a flat two-dimensional (2D) North-up map, a three-dimensional (3D) robot's eye view of the world, and a 3D view from behind the robot (3D-Camera) in which depth cues were manipulated. Our results show people adopt an egocentric frame of reference when depth cues are salient and an exocentric reference frame when depth cues are absent. The results suggest the absence or presence of depth cues is a critical component in choosing a reference frame. We discuss the extension of Bryant and Tversky's (1999) theoretical framework to a dynamic environment, such as navigation.

INTRODUCTION

Many researchers and consumers want to interact with automobile navigation aids, virtual environments, and Geographic Information Systems (GIS) via natural language. However, many current natural language systems that deal with space were built with a more traditional 2D view of space in mind (e.g. Dale, Geldof, & Prost, 2003; Wauchope, 1996) and will most likely not be able to handle 3D space using the same semantics and pragmatics. For example, if the utterance "Go up" were given to a traditional GIS/Map based system, the intent might be to go north (which might be towards the top of the screen). However, in today's current virtual reality systems, "Go up" might also mean to proceed in the vertical (z) direction.

Determining what frame of reference is being used could resolve this ambiguity. A **frame of reference** is a perspective a speaker chooses to talk about space. Several different reference frames can be used to describe space. The **egocentric** frame of reference exploits a speaker's' own perspective whereas **addressee-centered** reference frames requires a speaker to adopt another speaker's perspective. In an **object-centered** reference frame, the object itself or its features (sides, top, or bottom) orients a speaker. An **exocentric** frame of reference facilitates a top-down or world-based perspective where absolute compass degrees or cardinal directions (North, East, South, and West) orient a speaker in a location. For example, how would speaker two in Figure 1 describe the location of the light gray flowers? Table 1 shows several examples paired with the reference frame speaker two is using.

Several factors can influence a reference frame. For instance, while multiple reference frames are active at once (Carlson-Radvansky & Logan, 1997) people choose which reference frame to use based on a number of relationships. The functional relationship of an object can influence the reference frame when they are referenced (Carlson-Radvansky & Radvansky, 1996) (e.g. above/below, left/right), as can the features of different objects (Fillmore, 1975; Levelt, 1984) (e.g. in front of/in back of), the communicative aspect of tasks, such as identifying the relative location of circles to another speaker, (Schober, 1993) and the perspective adopted in a scene (Bryant & Tversky, 1999).

Where are the light gray flowers?

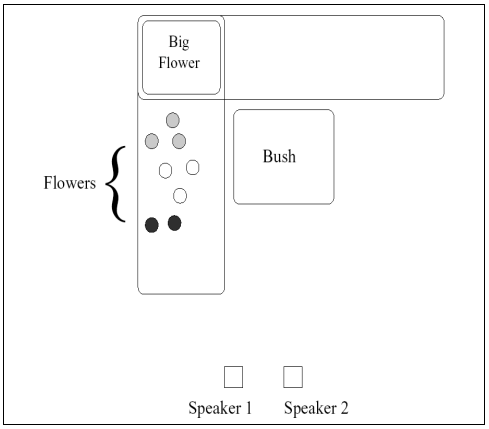

Figure 1. A simple figure for displaying multiple frames of reference.

Frame of Reference	Speaker 2's utterance
Egocentric	To my left
Addressee-Centered	To your left
Object-Centered	To the left of the bush
Exocentric	South of the big flower

Table 1. Different frames of reference for describing where the light gray flowers are in Figure 1.

We are primarily interested in how frames of reference are selected in a navigation task. While Carlson-Radvansky and colleagues (1997) have described how different reference frames are chosen, their work focuses on how a person describes an object, not on navigational tasks. Further, the task presented by Carlson-Radvansky and Logan (1997) examined frames of reference in a 2D static

environment and found spatial terms such as *above, below, left, and right* were used to describe an object's orientation; however, navigation is dynamic and traditionally more 3D. Thus, the most applicable theoretical work for navigation is Bryant and Tversky (1999), which we summarize below.

Bryant and Tversky (1999) describe a series of experiments, which predicts when a person will use an egocentric or exocentric frame of reference. They suggest people mentally put themselves in the place of the character or in the scene and use their own head, feet, front, back, and right, left to make judgments when there is a strong sense of depth in a scene. That is, with depth cues people use the egocentric frame of reference. Conversely, people take the outside perspective and adopt an exocentric frame of reference when there is not a strong sense of depth.

Bryant and Tversky's (1999) theory was supported by a series of experiments that used a static environment featuring a doll. The primary measure in the task was memory. Participants viewed the doll for unlimited time in a natural environment, such as a kitchen, which contained objects like a fork, a spoon, a plate, etc. Next, participants identified the object the doll faced in a computer task. Bryant and Tversky's (1999) experiment provided strong support for their theory based on memory measures. Our goal is to apply the same theoretical framework to a different task —dynamic navigation— and use natural language utterance types as our measure of reference frames.

If Bryant and Tversky's (1999) framework is correct, then we would expect exocentric utterances to be used in a 2D North-up display when depth cues are absent and egocentric utterances to be used with 3D displays when depth cues are present.

Experiment

We selected navigation because of its application toward natural language systems for narrative tasks. Hence, we opted for a video game that included both 2D and 3D displays. We included a 3D-Camera view, which integrates the exocentric and egocentric perspectives (see methods section for a complete description of this particular display) because performance differences have been cited between the two 3D views. For example, Olmos, Wickens & Chudy (2000) showed a difference in travel time for a simulated aircraft favoring the regular 3D view, whereas response time to avoid an obstacle was smaller for the 3D-Camera view. We wondered whether the differences between these two displays would reveal a difference in the proportion of reference frames used between the displays. However, we believed that because the display remained 3D, our prediction would still be correct.

The advantage of using these different perspectives is that we can manipulate depth cues and examine the types of utterances people used across displays in an experimental setting. A Wizard-of-Oz format, where a user interacts with a computer interface/system and a human operator controls its response, was used for this experiment. A video camera recorded the interface display along with participants' verbal utterances.

METHOD

Participants

18 employees from the Naval Research Laboratory (NRL) (13 men, five women) in Washington, D.C., volunteered to participate in this study. Their ages ranged from 22 to 49, with an average age of 36. Completed education ranged from high school diploma to Doctorate. 14 participants reported spending less than one hour per week playing video games. Only one participant reported having previously played Mechwarrior: Mercenaries II—the video game used in this task.

Materials and Task

The video game Mechwarrior: Mercenaries II featured a 2D North-up, 3D, and 3D-Camera display. The 2D North-up display did not contain depth cues. It provided a top-down perspective of the world below and was depicted like a map. Note though, a 2D track-up display was not available for this task. Depth cues were immersed in the 3D view and showed the simulated environment through the robot's eye (Wickens, 1999). In the 3D-Camera display, depth cues were tethered (Wickens, 1999) so the viewpoint was from several different horizontal angles but always from behind the robot. Figure 2 illustrates these featured displays.

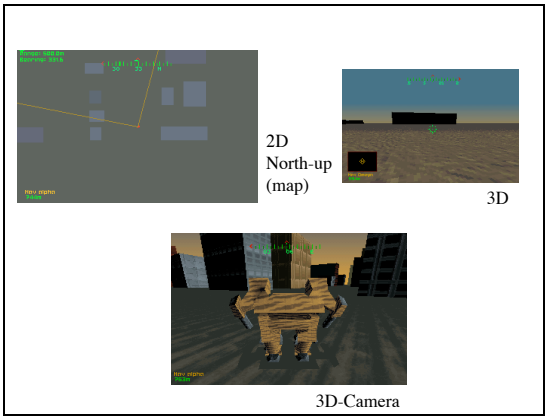

Figure 2. Snapshots from the three conditions. Squares and rectangles represent buildings in the 2D view. The robot is the small object at the center of the display.

The virtual environments featured an animated robot and landmarks, such as buildings and large barriers. A dynamic green arrow (located along the display's perimeter) lead participants along each route. A rotational compass was pointed out to all participants and included atop each display. We provided no explicit indication to participants how to issue commands to navigate the robot to the waypoints.

The task presented to participants required them to reach three specified waypoints as quickly as possible in each display. Orange squares indicated navigation (NAV) points in the 2D display as did green text labels (e.g. NAV) in both of the 3D displays. All navigation occurred on the surface and NAV points were grounded.

Design

This experiment used a between subjects manipulation. Participants were randomly assigned to one of three conditions, with six in each condition (2D, 3D, and 3D-Camera) respectively. Participants navigated along two different routes in their particular condition. Route order was counterbalanced and condition order was randomly determined.

Procedure

We used a Pentium III personal computer and a 20-inch monitor for this experiment. Participants sat in front of the monitor and issued verbal commands in English (during training/the experiment). Moreover, participants were told to regard themselves as the robot as it navigated along the simulated urban environment. Participants worked individually and served as navigator to the experimenter/robot operator who sat diagonally behind the participant and used a computer keyboard to control the robot.

Each course began with the robot stationary. If the "robot" understood a command, then it responded by performing the instruction. However, there were a couple constraints. First, if the participant issued a command that either contained the phrase, NAV point, required additional information (e.g. turn), or memory (e.g. 'Go back to where you were before'), the "robot" responded with, "I'm sorry, I don't understand." The participant then either rephrased the utterance or issued a new command. Second, the robot only could walk; however, its pace was on a continuous scale and participants could issue verbal commands to decrease/increase the robot's speed unless it was already operating at minimum/maximum speed.

Participants practiced in their specific condition before the experiment began. The same features (e.g. compass, etc.) existed in the demonstration, but the course routes and landmarks were different from the experimental routes. Trees were the only landmarks in the demonstration.

The experiment was divided into two sessions. The task was the same in both sessions, but we used different courses for each session to prevent participants from navigating in a familiar environment. Each session concluded when the final NAV point was reached.

Coding

Utterances were transcribed and segmented by instruction. Instructions were defined as a single utterance that included a noun (e.g. North) or adjective (e.g. left) to identify for the robot the direction to rotate. Commands only containing verbs, such as turn, were marked as incomplete commands because they require additional information from the participant on where to go. Incomplete commands were discarded from further coding. Off-task remarks and comments not influencing the outcome of the task were also eliminated from further coding.

Remaining instruction utterances were categorized according to either egocentric or exocentric reference frame. Stop and/or proceed commands were also categorized, but not by frame of reference since they do not possess specific reference frames. Table 1 shows examples of commands participants verbalized and Table 2 provides command examples coupled with its frame of reference.

Instruction	Description
Instruction	Turn left.
	Turn to 270 degrees.
	Go counter-clockwise.
	Turn due East.
Proceed	Increase speed.
	Maintain bearing.
	Walk forward.
Stop	Stop.
	Halt.
Incomplete utterance	Turn.
	Turn to 39*[1]
Off-Task	Bingo!
	He may be dead.

Table 2. Example instruction utterances issued by participants. [1]The * indicates the participant excluded the last digit, so the utterance is coded as incomplete.

Frame of Reference	Description
Egocentric	Go right.
	Turn left between the buildings.
	Turn to your left.
Exocentric	Go North.
	Go to 85 degrees.
	Turn around.

Table 3. Instruction utterances classified according to the type of frame of reference.

One author coded all the inclusive utterances and a second author coded a 10% subset of the corpus data. Inter-rater reliability (IRR) was 99%, Kappa = .99, $p<.001$

RESULTS

There was no difference in occurrence of reference frame type between the two routes, so we combined the data for each display. Software problems prevented one participant from completing his session, but the data were included since it occurred 115 meters from the final NAV point.

We predicted exocentric commands to be used in a 2D North-up display and egocentric commands in the 3D and 3D-Camera display. We found participants verbalized exocentric commands 97% of the time in the 2D North-up display, whereas when depth cues were present in the 3D and 3D-Camera displays respectively, participants verbalized exocentric commands 29% and 28% of the time. This difference was significant, $\chi^2(2)=60.9$, $p<.01$.

Further, egocentric commands were used 3% of the time in a 2D North-up display and 71% and 72% of the time in the 3D and 3D-Camera displays respectively. This difference was significant, $\chi^2(2)=64.3$, $p<.01$.

Finally, we inspected whether depth cues activated different reference frames across the displays. Bonferroni adjusted comparisons when $p<.001$ showed a significant difference between the proportion of egocentric and exocentric utterances between the 2D North-up and the 3D and 3D-Camera display, but no differentiation between the 3D and 3D-Camera displays themselves. Figure 3 shows these results.

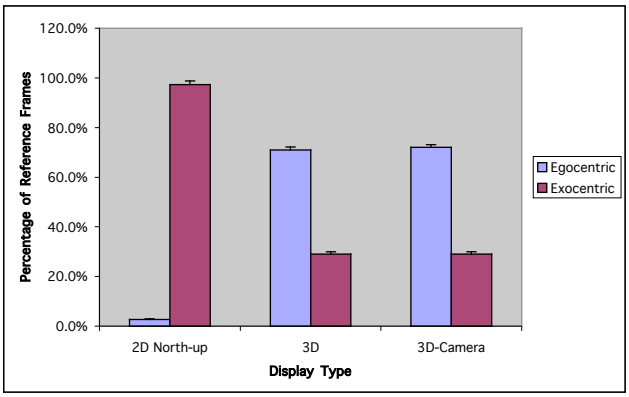

Figure 3. The percentage of reference frames utilized in each condition. Error bars show the standard error of the mean.

We also examined whether reiterating a command influenced frames of reference chosen in the displays. We collapsed commands that were identical and issued where the timing between them was similar. Collapsed commands showed the same patterns as above and thus do not significantly influence or change which reference frame are chosen in each display.

DISCUSSION

A substantial body of perspective taking empirical work has examined frames of reference through a number of different methods including mental tours (e.g. Linde and Labove, 1975), memory tasks and models (e.g. Taylor and Tversky, 1996; Bryant and Tversky, 1999). However, very little work focuses on perspective taking in a dynamic environment, such as navigation. To extend current perspective taking theories toward navigation, we developed a simple navigation task, based on Bryant and Tversky's (1999) theoretical framework, to probe how people talk about space in a dynamic environment. We found 2D North-up displays without depth cues elicited exocentric commands, whereas depth cues in 3D interfaces prompted egocentric commands. These findings replicate and extend Bryant and Tversky's (1999) model toward navigation and suggest the absence or presence of depth cues aid people in choosing reference frames in a navigational task.

We were able to replicate their findings by manipulating depth cues. Thus, there seems to be a close relationship between how people issue navigation commands and how they describe space from memory. Research on spatial descriptions has pointed to people using a route (e.g. egocentric), a survey (e.g. exocentric), or mixed descriptions which integrate instances of the route and survey perspectives (Taylor and Tversky, 1996). Further, Taylor and Tversky (1996) as well as Emorrey, Tversky, and Taylor (2000) showed a survey perspective was adopted when multiple pathways were available and a route perspective occurred when a single pathway existed. In these two studies, entire descriptions, not individual utterances (as in our study) were categorized. Despite this difference, the trends between multiple pathways and survey descriptions map toward the absence of depth cues and exocentric reference frames, whereas single pathways map toward the presence of depth cues and egocentric reference frames.

We can apply this framework toward the 2D and 3D-Camera displays we used in our experiment. In Figure 2 multiple pathways were available for the robot to navigate in the 2D display. In the 3D-Camera display, a single path existed. If we further inspected reference frames by counting the number of pathways available when a command was issued, we would expect to see a pattern between survey and route commands that is closely related to what our results show: exocentric reference frames are analogous to the absence of depth cues and egocentric reference frames are analogous to the presence of depth cues.

Since depth cues seem to influence the reference frame command, designers should consider the selected display type so the user can elicit maximum benefit. For example, use exocentric commands for 2D North-up GIS and use egocentric commands for 3D GIS. We are suggesting to match the display type to the reference frame. However, current in-car navigation systems employ a miss-match between the display and the command, where the display is 2D North-up and the command elicits an egocentric reference frame, which corresponds to a 3D display. Although this current technique contradicts what our study as well as what related literature shows would work, (e.g. Bryant & Tversky, 1999) it has been successful because commands are verbalized at the person-level.

We could imagine a situation where a 3D display is viewed with an exocentric command and a person responds by proceeding in the wrong direction. One reason why the miss-match method may not be as effective as the above example, is that the exocentric reference by definition assumes an environmental perspective and requires a user to have knowledge of an area to accurately determine which direction is indeed North.

We could also imagine that matching the display type to the reference frame would work in current virtual reality (VR) systems. The VR domain would use depth cues to simulate a realistic viewpoint of an environment. If navigation occurred in this domain, such as training sailors to know where things are on a ship prior to their arrival, we would expect people to adopt an egocentric reference frame as our results and others show (e.g. Bryant & Tversky, 1999).

One constraint of 3D systems is that regardless of whether a person is issuing or receiving a command, people should be congruently oriented in a 3D display to make a command as easiest as possible to understand. Facing different directions would increase difficulty and cause environment

descriptions to vary. For example, if Speaker Two stood behind the big flower, he/she would possess a different reference frame from Speaker One. The only ways for these two speakers' perspectives to be congruent to describe where the light gray flowers are located are either to use exocentric commands, or for one speaker to adopt an addressee-centered perspective by mentally rotating the environment or by physically rotating oneself. In domains that match display type to the appropriate reference frame, like air-traffic control and pilot communications, 2D North-up displays and exocentric commands are utilized because controllers and pilots frequently look at displays of the environment from different perspectives.

In summary, our study shows 2D North-up displays elicit exocentric commands and 3D displays elicit egocentric commands. The fact that different reference frames map towards certain displays suggests the absence or presence of depth cues helps determine which reference frame to use in a dynamic navigational task. Thus, designers should focus their efforts on adapting current in-car navigation systems and creating VR systems that match the display type to its correct reference frame.

ACKNOWLEDGEMENTS

This research was supported by grant number N0001402WX20374 to Greg Trafton. We thank Magda Bugajska and three anonymous reviewers for their comments on a previous draft. We also thank Nick Cassimatis for thoughtful discussion on this topic. The views and conclusions in this document are those of the authors and should not be interpreted as necessarily representing the official policies, expressed or implied, of the U.S. Navy.

REFERENCES

Bryant, D. J., & Tversky, B. (1999). Mental representations of perspective and spatial relations from diagram and models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(1), 137-156.

Carlsen-Radvansky, L. A., & Logan, G. D. (1997). The influence of reference frame selection on spatial template construction. *Journal of Memory and Language, 37*, 411-437.

Carlsen-Radvansky, L. A., & Jiang, Y. (1998). Inhibition accompanies reference-frame selection. *Psychological Science, 9*(5), 386-391.

Dale, R., Geldof, S., & Prost, J. P. (2003). Coral: Using natural language generation for NAVigational assistance. In M. Oudshoorn (Ed.), *Proceedings of the 26th Australasian Computer Science Conference* (ACSC 2003) (Vol. 16, pp. 35-44). Adelaide, Australia.

Emmorey, K., Tversky, B., & Taylor, H.A. (2000). Using space to describe space: Perspective in speech, sign, and gesture, *Spatial Cognition and Computation, 2,* 157-180.

Fillmore, C. J. (1975). *Santa Cruz lectures on deixis*. Bloomington, IN: Indiana University Linguistics Club.

Levelt, W. J. M. (1984). Some perceptual limitations on talking about space. In A. J. v. Doorn, W. A. v. d. Gring & J. J. Koenderink (Eds.), *Limits in Perception* (pp. 323-358). Utrecht: NVU Science Press.

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

Olmos, O., Wickens, C.D., & Chudy, A. (2000). Tactical displays for combat awareness: An examination of dimensionality and frame of reference concepts and the application of cognitive engineering, *The International Journal of Aviation Psychology*, *10*(3), 247-271.

Schober, M. F. (1993). Spatial perspective-taking conversation. *Cognition, 47*, 1-24.

Taylor, H.A., & Tversky, B. (1996). Perspective in spatial descriptions, *Journal of Memory and Language, 35,* 371-391.

Wauchope, K. (1996). *Multimodal interaction with a map-based simulation system* (No. AIC-96-027). Washington, D.C.: Naval Research Laboratory.

Wickens, C. D. (1999). Frames of reference for navigation. In D. Gopher & A. Koriat (Eds.), *Attention and Performance* (pp. 113-144). Orlando, FL: Academic Press.