

*This talk highlights the CONCERT corpus, a dialogue corpus collected during the summer of 2006. We define the research question, outline the corpus properties, highlight a few interesting aspects of the corpus, and finally talk about future directions in this research.*

## 1 Research Questions

Spatial frames of reference are formally defined as “a unit or organization of units that collectively serve to identify a coordinate system with respect to which certain properties of objects... are gauged” ([2], pg 24) The research we conducted with the CONCERT corpus was geared to identify and answer two overall questions.

**How are these coordinate systems formed?** Context plays a role in the appropriateness and salience of reference frames. If we could identify situations in which certain coordinate systems are preferred, then we could begin to construct models of the world that take spatial context into account.

**How can we pick out the appropriate referent based on these coordinate systems?** Using the speaker’s utterance, we should be able to identify an appropriate frame of reference. Then, using this information, we would be able to identify which item was the intended referent.

Previous research into spatial reference by linguists and cognitive scientists [1] [4] has identified three general strategies for constructing spatial frames of reference.

### 1.1 Intrinsic

In intrinsic reference, coordinate systems are generated with respect to **salient features of the referent object**. An *intrinsic spatial relator*  $R$  is a binary spatial relation, with arguments  $F$  (the found object) and  $G$  (the grounding object) where the relator typically names a part of the grounding object and the origin of the coordinate system is always on the volumetric center of  $G$ . The relation asserts that  $F$  lies in a search domain extending from the grounding object on the basis of an angle projected from the center of  $G$  through an anchor point (usually the named part) outwards for a determined distance.

- “the man in front of the house” – **front**( ‘man’, house )
- “the man on the left side of the house” – **left**( ‘man’, house )
- “the man in front of me” – **front**( ‘man’, speaker )
- “the woman to the left of the bus” – **left**( ‘woman’, house )

### 1.2 Absolute

Absolute coordinate systems are derived with respect to **salient features of the environment**. More formally, an *absolute relator*  $R$  expresses a binary relation between  $F$  and  $G$ . It asserts that the found object can be found in a search domain at the fixed bearing from  $G$ . The origin  $X$  of the coordinate system is nearly always centered on  $G$ , and the system of terms anchored by reference to a conceptual ‘slope’.

- “the man north of the house” – **north**( ‘man’, house )
- “the woman south of the bus” – **south**( ‘woman’, bus )
- “the thing under the chair” – **under**( ‘thing’, chair )

	INTRINSIC	ABSOLUTE	RELATIVE
<b>Relation:</b>	binary R(F,G)	binary R(F,G)	ternary R(V,F,G)
<b>Origin on:</b>	Ground	Ground	Viewpoint V
<b>Anchored by:</b>	A within G	'slope'	A within V
<b>Transitivity:</b>	no	yes	yes ( <i>under constant V</i> )
<i>Constance under rotation of:</i>			
<b>Whole array:</b>	yes	no	no
<b>Viewer:</b>	yes	yes	no
<b>Ground:</b>	no	yes	yes

Table 1: Summary of properties of different frames of reference; adapted from [2], pg 53

### 1.3 Relative

The final coordinate system is derived based upon an **explicit (or presupposed) viewpoint distinct from the objects being described**. A *relative relator*  $R$  expresses a ternary spatial relation. Its arguments include the viewpoint  $V$ , the located item  $F$  and the grounding item  $G$ . There will always be a primary coordinate system with its origin on  $V$ . There may also be a secondary coordinate system with its origin on  $G$ .

- “the man to the left side of the house” – **left**(  $V_s$ , ‘man’, house ); origin on  $V_s$
- “the man in front of the tree” – **front**(  $V_s$ , ‘man’, tree ); origin on  $V_{tree}$
- “the woman to the right of the bus, from your point of view.” – **right**(  $V_h$ , ‘woman’, bus ); origin on  $V_h$
- “the man to the left side of the house” – **left**(  $V_h$ , ‘man’, house ); origin on  $V_h$

## 2 Corpus

This summer, we collected a corpus of two-party dialogues in the Quake II embodied-task environment. The outcome of this work is the OSU CONversational, Collaborative, Embodied, Rearrangement-Task (**CONCERT**) Corpus. The CONCERT Corpus was designed to test the participants’ spatial reference strategies.

### 2.1 Task

In each trial, two participants are placed in a “room” in the virtual world with eight identical buttons and four identical boxes that are located on the table in the room. The participants are then asked to work together in order to perform two related tasks:

1. Document what each button in the room does.
2. Rearrange the objects into a designated goal configuration.

One of the participants, the *leader*, is given a diagram containing a final configuration of the four objects arranged on the table. The *follower’s* instructions include blank space in which he is supposed to record what happens when each button is pushed. When all the buttons have been pushed, and the objects are arranged in their appropriate configuration, the subjects indicate that they are finished, and move onto the next trial.

## 2.2 Manipulation

In our experiment, we attempted to manipulate the participant's frame-of-reference strategy used to accomplish reference to items in the world.

Certain conditions were held constant across all trials. Although they are not identical from trial to trial, all four boxes in each individual trial are identical. Tables are either square or 'circular', and do not have any noticeable 'front' or 'back' side. Additionally, all the walls are identical, and each button is spaced evenly from the corner of the room. This was done deliberately in order to force the participants to talk about the boxes in terms of spatial orientation and not in terms of other properties (such as "the red box" or "the large box").

Then, we applied two variables to the trials. First, the table was either located in the center of the room, or it was placed against one of the walls. Second, the participants were either together in the room, or the leader was placed in a separate space above the follower, giving them more of a bird's-eye-view of the task. The leader could see what the follower was doing, but could not push any of the buttons themselves.

## 2.3 Files

The corpus contains 256 trials spread out over seventeen sessions. Trials begin when either participant enters the room and end when the last participant leaves the room. The beginning of a trial overlaps with the end of the previous trial.

For each trial, we will have:

- **WAV** - Raw audio .wav file (16-bit @ 32kHz) [*in progress*]
- **MOV** - H.264 encoded Quicktime movie (640x480); aac audio compression [*in progress*]
- **TEXTGRID** - Praat time-aligned transcript file [*in progress*]
- **POS** - Both participant's position; recorded at 10Hz [*to be started*]
- **EVENT** - Time-aligned event description; recorded at 10Hz [*to be started*]
- **POV** - Description of items in each participant's field of view; recorded at 10Hz [*to be started*]

## 3 Examples

- "the southeast block"
- "one that's diagonal from it"
- "the northmost thing"
- "left front box"
- "the right front"
- "the block that was on my right but just north of the other block"
- "the lower block"
- "the one that is now at the top of the table from my perspective. the one farthest from me."
- "the right center"
- "block one"

## 4 Future Work

The CONCERT corpus will provide additional insight into computational models for spatial reference. Fields like robotics have just begun to ask related questions about spatial frames of reference. There have been recent papers on the importance of perspective taking for robots [5] and robotic reasoning using projective relations [3].

Immediate work on our corpus involved generating and organizing our data. We are currently in the process of converting the video into a sharable format and also cutting the sessions into their component trials. Additionally, we have been working on transcribing the audio recordings using Praat. Soon, we will begin work on extracting positional, event and field-of-view information from the server logs.

Our first research question to answer will be: “did our manipulations have any effect?” We will annotate corpus references with the reference strategy employed. Using the three frame of reference models, will we be able to see any change across the four conditions? If so, what does that tell us about the requirements for the reference models?

After that, we have several interesting directions to go. One possibility is to try to construct a computational model of spatial reference in our corpus domain. Given an utterance, can we pick out the intended referent? Grounding and task strategies would also be an interesting course of investigation. Another potential direction is to investigate what the surface forms of an utterance tell us about the spatial frame employed. Is “on the left” any different functionally than “to the left”, or can they be employed interchangeably?

## References

- [1] K. R. Coventry and S. C. Garrod. *Saying, Seeing and Acting: The Psychological Semantics of Spatial Prepositions*. Psychology Press, 2004.
- [2] S. C. Levinson. *Space in Language and Cognition - Explorations in Cognitive Diversity*. Cambridge University Press, 2003.
- [3] R. Moratz and T. Tenbrink. Spatial reference in linguistic human-robot interaction: Iterative, empirically supported development of a model of projective relations. *Spatial Cognition and Computation*, 6(1):63–106, 2006.
- [4] H. A. Taylor, S. J. Naylor, R. R. Faust, and P. J. Holcomb. ”could you hand me those keys on the right?” disentangling spatial reference frames using different methodologies. *Spatial Cognition and Computation*, 1:381–397, 1999.
- [5] J. G. Trafton, N. L. Cassimatis, M. D. Bugajska, D. P. Brock, F. E. Mintz, and A. C. Schultz. Enabling effective human-robot interaction using perspective-taking in robots. *IEEE Transactions on Systems, Man and Cybernetics*, 35(4):460 – 470, July 2005.