Knowledge representation for generating locating gestures in route directions

Kristina Striegnitz[1], Paul Tepper[2], Andrew Lovett[2], Justine Cassell[2]

[1]Computer Science Department, Union College

[2]ArticuLab, Northwestern University

kris@union,edu, {ptepper,andrew-lovett,justine}@northwestern.edu

Abstract

When humans give route directions, they often use gestures to indicate the location of landmarks. The form of these gestures reflect one of several perspectives that speakers take when producing them. They may locate the landmark with respect to the speaker, with respect to the person following the route, or with respect to other landmarks. A corpus study shows that the perspective chosen is partly determined by the function of the discourse segment these gestures occur in. Since locating gestures are so prevalent in direction-giving, in this paper we address the kinds of dialogue information and knowledge representation that is needed to generate them automatically.

1. Introduction

When giving route directions, humans may use gestures for a variety of purposes, such as indicating turns and movement direction, to describe the location of landmarks, and to depict their shape. In previous work (Kopp, Tepper, Ferriman, Striegnitz & Cassell, 2007), we have studied how gestures are used to describe the *shape* of landmarks and how such gestures can be generated in an embodied conversational agent (ECA). In this paper, we look at the way humans use gesture to indicate the *location* of landmarks. Emmorey, Tversky, and Taylor (Emmorey, Tversky & Taylor, 2001, Taylor & Tversky, 1996) have found that people alternate between different perspectives when giving directions. We examine the use of these different perspectives in our data (Section 2). Next, we formulate requirements on knowledge representation for generating such gestures in an ECA (Section 3), and we propose a way of implementing these requirements (Section 4). We then sketch how this information is used in a direction giving ECA (Section 5). Finally, Section 6 relates our results to previous work before we conclude in Section 7.

2. Gestures in direction giving dialogues

2.1 Data

The observations described in this paper are based on videos of people giving directions across Northwestern University's campus to another person who (they believe) is unfamiliar with the campus. In addition to transcribing the speech, we have identified and coded gestures referring to landmarks, annotated them with their referents (a basic name for what they seem to depict) and information about the perspective used (as described below). Utterances have, furthermore, been marked for the dialogue moves that they accomplish, using a coding scheme that was inspired by the DAMSL coding scheme (Allen & Core, 1997) and by the scheme for classifying instructions in route directions introduced by Denis (1997). The scheme is also similar to the one used by Muller and Prévot (2008) to annotate French direction giving dialogues with dialogue moves.

We coded 5 direction giving dialogues which altogether consist of 753 utterance units by the person giving the directions and 234 utterance units by the person receiving the directions. In this paper, we are interested in the direction giver's language and will, therefore, concentrate on his contributions to the dialogue.

Utterance units are annotated along five different dimensions. First, they are classified with respect to their communicative status and information level. 640 of the direction giver's utterance units are interpretable and directly pertain to the task. All others were either abandoned, otherwise uninterpretable, or meta-communication about the task or conversation.

The second dimension marks utterance units that make assertions contributing to the route description as statements. We distinguish six types of statements: instructions to reorient or to reorient with respect to a landmark (labeled as reorient and reorient+lm, respectively), instructions to move or to move with respect to a landmark (move/move+lm), statements that mention a landmark without an instruction to reorient or move (lm), and statements describing cardinal directions (dir), such as "*north is that way*". 597 of the 640 utterance units by the direction giver (that is, 93%) are statements. Table 1 shows the distribution of utterance units over statement types. [INSERT TABLE 1 HERE]

Our third and fourth dimensions look at queries and responses marking clarification questions (Q-clarif), requests for feedback (Q-feedback), and other requests for information (Q-other), and answers to clarification questions (A-clarif), backchannel feedback (A-ack), and other answers (A-other). 18 of the direction giver's utterances (3%) are queries and 185 (29%) are responses. 172 of the responses are answers to clarification questions and 13 are backchannel feedback. Note that the statement, query and response dimensions are not mutually exclusive. For example, many statements (158) are part of a response. Therefore, the totals for statement, query and response type utterance units do not add up to 640 or 100%.

Finally, we mark utterance units that belong to an elaboration on a landmark or action (elab), such as the second utterance in "The Allen Center is to your left. It's really big.", or that are part of a redescription of a route segment that has previously been introduced and described (repeat). In our data, 227 utterance units are annotated as elaborations and 75 as part of a redescription. All of them are statements.

2.2 Perspective of locating gestures in direction giving dialogues

The literature on route descriptions discusses two perspectives that people use for describing space along the route (Taylor & Tversky, 1996). In **route perspective**, landmarks are described in the frame of reference of a person walking the route. The **survey perspective** is like a birds-eye view. Buildings are described relative to each other or to an absolute frame of reference (for example, cardinal directions). These two different perspectives are also reflected in the gestures that accompany speech (Emmorey, Taylor & Tversky, 2001), and we find examples of both perspectives in our data. In our data, we also find gestures that do not fall into these two categories. First, we find gestures that seem to be purely shape depicting, and do not refer to the location of the referent landmark at all. Second, we find gestures which locate the object with respect to the speaker's actual position and orientation.

Figure 1 shows an example of a gesture where the speaker takes on the perspective of the person following the route (the route perspective). He speaks and gestures as if he has the position and orientation that an imaginary direction-follower would have at this point along the route. Therefore, the location of his gesture (to the left of his body) corresponds to the location of the landmark relative to the location and orientation of the imaginary direction-follower. This perspective is by far the most common in our data (54.2% of all gestures referring to landmarks). [INSERT FIGURES 1 AND 2 HERE]

Another way in which people use their hands and the space around their bodies is to lay out virtual maps using a birds-eye view, as shown in Figure 3. Map gestures are unique in that after one gesture is made, the hand is held in place, while the next location is depicted relative to the first, by placing the other hand relative to the position of the first. As Figure 3 illustrates, the right hand representing University Hall is the anchor, held in exactly the same position throughout the three-gesture sequence, while the locations of Kresge and Harris Hall are shown relative to it. Kresge is shown using an almost identical gesture, a flat hand shape facing downwards, placing the building with respect to University. This probably indicates a survey perspective for these two gestures. Harris is not actually placed in the same way; rather it is pointed to in a kind of deictic gestures that assumes the route perspective, or the perspective of the imaginary direction follower. This mixed-perspective interpretation is supported by her language, which serves to place the first two landmarks, University and Kresge, and indicates that the third, Harris, is not placed on the left or the right of the follower, but “straight ahead” of the follower. Overall, the virtual map is oriented in the same way, such that it matches up with the direction a person walking the route would be facing. We found that 16.3% of the landmark-depicting gestures in our data are survey perspective map gestures. [INSERT FIGURE 3 HERE]

It is important to note that gestures referring to landmarks do not necessarily have a locating function. For example, after having located the Allen Center to the left of the direction-follower, the speaker in Figure 1 continues by saying *and it's really big*. He accompanies this elaboration with the gesture shown in Figure 2, which refers to the landmark's shape by indicating its horizontal extent. This gesture does not locate the landmark to the left, which would be its position with respect to the point of view assumed for the previous utterance. Instead the gesture is carried out it in front of the speaker's body. In our data, 15.8% of the gestures referring to landmarks are of this non-locating kind.

However, often gestures are neither purely locating nor purely shape depicting. For instance, the gesture used in Figure 1 seems to indicate the wall of the building being described, as the shape of the hand is flat and vertically oriented. It thus has a shape depicting component in addition to its locating function. In this paper, we are concerned with the locating function of gesture and will not address the issue of how to determine which shape features to depict and how to depict them (but see Kopp et al., 2007 and Sowa & Wachsmuth, 2008 for more on these questions).

Finally, gestures may be used to locate objects with respect to the speaker. That is, the speaker simply points to a real object. This type of gesture is extremely rare in our data (only 1.9% of all gestures referring to landmarks fall in this class). Table 2 shows the distribution of perspective among gestures referring to landmarks in our set of direction giving dialogues. [INSERT TABLE 2 HERE]

2.3 Perspective and dialogue structure

In order to generate locating gestures with different perspectives, we must address the following question: When are the different perspectives used? As the following results show, the use of these perspectives seems to be determined in part by the dialogue move that the speaker is trying to perform.

In our data, most of the direction giver’s gestures referring to landmarks occur with utterance units marked as statements. In fact, *all* of the survey perspective, route perspective, and non-locating gestures, which are the gestures we are most interested in, co-occur with statements. Table 3 shows which statement types the different gesture perspectives co-occur with. Unsurprisingly, gestures of any perspective that are referring to landmarks co-occur with utterances that mention a landmark in the speech. (Remember that we are not looking at gestures depicting actions in this paper.) [INSERT TABLE 3 HERE]

None of the gestures under consideration co-occur with queries, but some of them co-occur with statements that are also marked as an elaboration, as a re-description of previously explained route segments, or as a response to a clarification question (we do not have cases of co-occurrence with other response types). Tables 4-6 show the frequency with which gestures of the different perspectives co-occur with utterance units with these labels. Table 7 shows how often gestures of the different perspectives co-occur with plain statements, that is, statements which are not marked as a response, a query, an elaboration or re-description. The tables also show the percentage deviation for those frequencies, which measures how much the frequency differs from the frequency we would expect if gestures were equally likely to co-occur with utterance units of any dialogue function. [INSERT TABLES 4, 5, 6, AND 7 HERE]

Survey perspective gestures occur much more often than we would expect in answers to clarification questions and in re-descriptions of route segments. They occur much less often than expected in plain statements. This indicates that speakers switch to survey perspective when they need to re-explain a portion of the route. It also fits findings of a previous study on direction-giving, which differed from our own in that the subjects could use a physical map (Cassell et al., 2002). In that study, subjects only refered to the map if their purely verbally given directions were not sufficient.

In contrast, route perspective gestures occur more often than expected in plain statements and less often in statements marked as A-clarif, elab, or repeat. So, the route perspective seems to be the default when gesturing about landmarks.

Non-locating gestures, finally, occur much more often than expected in elaborations and much less often in plain statements. They occur slightly more often than expected in answers to clarification questions. This can be explained as follows. After having introduced a landmark, probably using a gesture that locates the landmark, speakers give further information about the visual properties of the landmark, such as its shape or size. This is reflected in their gestures, in which the locating component may be absent or deemphasized.

3. Requirements on knowledge representation

To generate any kind of route description, a map of the relevant area is needed. Minimally, the map must include the paths that can be taken, so that the system can calculate the route. Unlike direction-giving systems like MapQuest, our system gives directions using landmarks to indicate reorientation points and other important points along the path. Therefore, our map representation has to include the landmarks located along these paths.

As the data presented in Section 2 show, gestures referring to these landmarks may express different perspectives. The perspectives differ in whether or not and how relative location in the map representation is reflected in the placement of gestures in the gesture space. This requires information about the position and orientation of both the imaginary direction follower and the speaker as well as mechanisms for inferring spatial relations between entities in the map and mapping them to the speaker’s gesture space.

For survey and route perspective gestures, we need to keep track of the position and orientation that a person following the route would have at each point of the description, and to generate gestures which locate landmarks relative to the speaker, we need the position and orientation of the person or ECA giving the directions in the map. The system also requires mechanisms for inferring spatial relations between the entities in the representation. For example, the system needs to be able to infer the location of landmarks relative to paths, other landmarks, the speaker, and the direction-follower. This is necessary for the decision which landmarks to mention in the route description; landmarks that are mentioned at a specific point in the description should be visible to the direction-follower when he/she reaches the corresponding point of the route. In addition to these inference mechanisms, the system needs an appropriate mapping from positions in the map representation to positions in the gesture space in order to place route as well as survey perspective gestures correctly in the gesture space. For example, the position of route perspective gestures should reflect the relative location of the landmark with respect to the direction follower, and the positions of the different gestures in a survey perspective sequence should reflect the relative location of the landmarks to each other and to the direction follower. Additionally, the discourse history has to contain information about the current location of the hands and which landmark they stand for, such that multimodal anaphoric expressions can refer back to these landmarks in later utterances.

Finally, landmarks and paths must be associated with semantic information. For instance, a description of a landmark could draw upon information about its name, type (building, lake, monument, etc.), size, color, and shape. For paths, we may specify what type of path it is, a street, parking lot, courtyard, etc. This information is necessary for generating descriptions of landmarks together with gestures depicting their shape and/or size. In the next section, we propose a way of implementing the knowledge requirements formulated above in an ECA.

4. Locating landmarks in space

The basis for generating locating gestures is a map representation consisting of two interlinked components: (i) a graph, where edges represent the paths that can be walked and nodes (path points) represent points on the map where the direction-follower may have to change his direction, and (ii) a set of landmarks. Landmarks are associated with areas and path points are associated with points in a common coordinate system (see Figure 4). In addition, path points can be linked to landmarks by qualitative relations specifying whether a path point is the entrance of a building or whether it is next to a landmark (Figure 5). Finally, landmarks and path points are associated with semantic information as described above (type of landmark, size, color, shape, etc.). Also see Shi and Tenbrink

(2008) for a discussion of representing spatial information for direction giving and following. [INSERT FIGURES 4 AND 5 HERE]

4.1. Locating landmarks with respect to the direction-follower's and the speaker's perspective

When gestures are used to locate landmarks with respect to the *direction-follower's* point of view, they depict the landmark at a location in the gesture space. This location corresponds to the location of the landmark relative to the position and orientation that the direction-follower would have in the world at that moment if he/she were walking the route. This holds whether it is a simple pointing gesture or a gesture that depicts some aspect of the landmark's shape, as in Figure 1. In order to generate such gestures, we need to keep track of the position and orientation of the direction-follower in the map representation. These values change continually over the course of the dialogue, as the description (and the imaginary direction-follower) progresses along the route.

Given a route between two points on the map graph, we can derive the direction-follower's orientation for each point along this route, based on the location of the previous point on that route. This allows us to calculate the angle at which landmarks are located with respect to the direction-follower's orientation, which can then be mapped to different positions in the speaker's gesture space. Since these gestures are normally only used to locate the landmark with respect to the direction follower and do not represent relative location to other landmarks, we use a coarse mapping that maps ranges of angles to five different positions in the gesture space: left, right, font left, front right, and front (see Figure 6). [INSERT FIGURE 6 HERE]

Gestures that locate objects with respect to the *speaker* can be generated using the same mechanisms, given that the location and orientation of the speaker are recorded within the map representation. Note that in our current application the speaker is our ECA, which is part of a stationary information kiosk. The agent is displayed on a fixed screen, so its position and orientation remain the same over the course of an interaction.

4.2. Generating map gestures

In their simplest form, map gestures resemble the act of placing objects in the horizontal, tabletop plane in front of the speaker. While they can get more complicated than this, for example, by also depicting information about the shape of the objects, here we will just consider this basic case of positioning objects. Neither are we currently modeling map gestures where route and survey perspective are mixed, as in the example in Figure 3.

Each map gesture depicts a limited section of the map of the world. This section contains the target landmark and a number of other visible landmarks. We choose landmarks which either could easily be confused with the target or can help in distinguishing it. For example, if the target landmark is a building which is to the left of the direction follower and there is another building which is also to the left or to the left and front, then the target could easily be confused with this second landmark based on their location. Or if, for example, the target is a path turning only slightly left and there is another path continuing straight, these two paths can easily be confused and would both be included in a map gesture.

Once we have identified which landmarks to include in the map gesture, we compute the angles at which those landmarks are located with respect to the current position and orientation of the direction follower in the map or, in the case of paths, the angle at which the path leaves this point. Those angles are then mapped to positions on an imagined circle which is centered slightly in front of the speaker's body in the tabletop plane. Positions on this circle are described in terms of the three-dimensional coordinate system representing the speaker's gesture space. Figures 7 and 8 show examples of this mapping. If we assume the target landmark in Figure 7(a) is building B, there is one building (building A) which could easily be confused with the target. So the relevant section of the map for the map gesture contains buildings A and B. Figure 7(b) shows the positions in the gesture space they are mapped to. Let us now assume that the target is the path labeled C in Figure 8 (a). This path could easily be confused with path E, while building D can help to distinguish them. Figure 8 (b) shows how paths C and E and building D get mapped to the gesture space. [INSERT FIGURES 7 and 8 HERE]

The next step is to decide what gestures to use to indicate these locations and how to order them. We use a static gesture for buildings, which places a hand with a flat hand shape and the palm pointing down at the point in the gesture space determined by the mapping. For paths we use a dynamic gesture which “draws” a line from the center of the imagined circle to the (end) point determined by the mapping. A pointing hand shape (where the index finger is extended and all other fingers are curled up) is used.

The order of the gestures making up the map gesture is determined as follows. Generally, the target is mentioned first and then all other landmarks going either clockwise or counterclockwise from the target. If the target is a path and some three-dimensional landmarks are involved in the map gesture, the three-dimensional landmarks are mentioned first, then the target and then all other landmarks.

Finally, we propose to store information linking the agent’s hands to their locations and to the entities they represent in the dialogue context. This information needs to be updated appropriately as the relations between hands, locations, and landmarks change. This allows later utterances to make use of the information, for example, in order to generate appropriate multimodal anaphoric references to landmarks, where the ECA continues using the same hand and location to refer to the same landmark as long as the direction-follower's position and orientation remains stable.

5. Architecture of a direction giving ECA

Now, we move on to describing the architecture of our ECA called NUMACK, illustrated in Figure 9. First, we discuss the dialogue management module and its central data structure, the Information State. Next, we describe the content planning stage, which includes a route planner that employs a map representation specialized for gesture and natural language generation (see Section 4). The content planner also determines the perspective used in each gesture. Lastly, we give a brief description of the multimodal microplanner and surface realization components. [INSERT FIGURE 9 HERE]

At the center of the system is the Information State (Traum & Larsson, 2003). This is a data structure that keeps track of the dialogue history, the private knowledge of the system, the shared knowledge of user and system, and the current state of the system. In addition to this kind of information, which is commonly found in any Information State, we also use the Information State to store the output of the content planner, and to keep track of the point in the route the description has reached. We are still working on integrating the information necessary for producing anaphoric gestures as described in the previous section into the Information State.

The Dialogue Move Engine determines how to integrate user dialogue moves into the Information State and chooses the moves of the system. We use Midiki, the MITRE Dialogue Kit (Burke et al., 2003) in our implementation, which provides a Dialogue Move Engine, lets us specify the rule system for the Dialogue Move Engine, and maintains the Information State for us.

Once the system has determined where the user wants to go and where he wants to leave from, the route planner calculates the shortest path between these two points. The map representation that the route planner currently works with has been coded by hand. Ultimately, we would like to automatically derive the necessary information from existing sources of geographic information. The output of the route planner is a sequence of path points and the task of the next step, which is content planning, is to map this to a sequence of preverbal messages, which can then be turned into multimodal utterances by the multimodal microplanner. More specifically, the content planner (i) chooses which path points to mention, (ii) decides which instruction types (that is, reorient, reorient+lm, move, move+lm, or lm) to use for describing each step in the route, (iii) selects landmarks that can be used to identify path points to the user, and then (iv) determines the semantic content of the expressions referring to those landmarks. In step (iv), the content planner chooses the properties of the landmark that need to be expressed either in the language or in gesture to distinguish the landmark from its surroundings. It also determines the perspective that should be used with respect to gesture.

So, it is in these last two steps that the data structures described in the previous sections come to bear. By default, the system assumes the route perspective. Figure 10(a) shows an example of a route perspective gesture, which accompanies the words "*Pass the Allen Center on your left.*" Non-locating gestures are only used in elaborations on landmarks that do not mention the location of that landmark (for example, Figure 10(b): "*Dearborn Observatory is the building with the dome*"). As our system's capabilities to accept and react to clarification questions are still very limited, we only use map gestures for re-descriptions of route segments. Such re-descriptions are triggered if a reorientation occurs at a point where one or more turns are possible that can easily be confused with the target turn (cf. the situation in Figure 8(a)), or if the destination landmark can easily be confused with neighboring landmarks (cf. the situation in Figure 7(a)). Figure 10(c) shows an example of such a map gesture. The accompanying speech is "*Annenberg Hall is here and the Seminary is here*" where the first occurrence of *here* refers to the position of the right hand and the second one to the left hand. [INSERT FIGURE 10 HERE]

The output of the content planner specifies the structure of the route description and the semantic content that needs to be expressed by each utterance. It is stored in the information state. Based on user feedback, the dialogue manager chooses when to send the next utterance specification to the microplanning and realization modules. The multimodal microplanner determines the form of the utterance, including the actual words as well as the form of the gestures and the coordination between language and gesture (Kopp et al., 2007). Finally, the surface realization component turns the utterance specification produced by the microplanner into speech and movements of the animated character on the screen (Kopp & Wachsmuth, 2004).

6. Related work

Most literature on deictic gestures in multimodal interfaces concerns the *interpretation* of such gestures (see, for example, Bolt, 1980, Johnston & Bangalore, 2000). There are systems which *generate* deictic gestures, such as the COMIC system (Foster, 2004), DFKI's PPP Persona (André, Rist & Müller, 1998), but these systems only handle pointing gestures that point to objects presented on the screen. They are, hence, what we have called gestures that locate objects with respect to the speaker.

Another body of research that is relevant to our application is the existing work on generating natural language route descriptions. For example, Dale, Geldof and Prost (2005) generate driving directions from GIS data. Look, Kottahachchi,, Laddaga, and Shrobe (2005) produce walking directions, but concentrate on the representation of the information necessary for planning the route rather than the planning and realization of the natural language output. Habel (2003) concentrates on the architecture of a generation system for route directions, arguing for an incremental processing model. None of these systems model face-to-face dialogue and, hence, none of them look at generating the gestures that humans use when giving route directions.

More recently, Theune, Hofs, and van Kessel (2007) describe an ECA that generates route directions in a virtual environment. However, they do not generate words and gesture in an integrated way – the words are generated first, then gestures are added – and while their system has a mechanism for choosing between different kinds of gestures they do not consider survey gesture and seem to mostly rely on pointing gestures from the direction follower's point of view. As part of this research, Evers, Theune, and Karreman (2007) also investigate the effect that the orientation of the direction giver with respect to the person receiving the directions has and found no evidence that whether the ECA is facing that person or is positioned to look into the same direction as that person influences the effectiveness of the directions. The directions are perceived as more natural, however, when the ECA is facing the user. As Figure 10 shows, NUMACK is facing the user.

7. Conclusions and future work

Previous work on human face-to-face dialogue has shown that speakers assume different perspectives when giving route directions (Taylor & Tversky, 1996). In particular, they use the route perspective, which refers to landmarks with respect to an imaginary direction-follower’s point of view, and the survey perspective which locates landmarks using a birds-eye view. Our data supports this finding and also shows that, in addition to route perspective and survey perspective gestures, people use non-locating gestures and gestures that locate landmarks with respect to the speaker’s point of view. The distribution of these gestures is partly determined by the dialogue move of the utterance they occur in. Our goal is to model the different uses of locating gestures in a direction giving ECA in order to produce route descriptions which are more natural and easier to understand. To the best of our knowledge the issue of perspective in locating gestures has never been addressed with the aim of generating such gestures in a virtual agent.

In this paper, we have discussed the knowledge necessary for generating such gestures and we have proposed a way of representing this knowledge in an implemented system. More specifically, we have argued that we need a suitable map representation (representing not only the paths that can be walked on but also landmarks in relation to these paths as well as additional semantic information about properties of paths and landmarks) and that we have to be able to keep track of the position and orientation of entities in this map (that is, landmarks as well as the direction-follower and the speaker). This information is necessary for generating route perspective and survey perspective gestures as well as gestures that locate a landmark with respect to the speaker’s point of view. In the case of map gestures, the position of the speaker's hands needs to be recorded, linked to landmarks, and this information needs to be appropriately updated as the discourse proceeds.

The proposal made in this paper is implemented in a direction giving ECA. We are currently preparing a study to evaluate the way this ECA uses gestures. Furthermore, we are working on making the system more interactive. The main goal is to make it more effective by taking user feedback into account, but this will also allow us to further integrate our findings on how dialogue moves influence gesture perspective.

Acknowledgements**:** We would like to thank Magan Sethi and Laura Cristina Stoia for their help with implementation. This research was supported by a fellowship from the Postdoc-Program of the German Academic Exchange Service (DAAD).

References

Allen, J. & Core, M. (1997). Draft of DAMSL: Dialogue Markup in Several Layers. Retrieved on February 10, 2008 from http://www.cs.rochester.edu/research/speech/damsl/RevisedManual/.

André, E., Rist, T. & Müller, J. (1998). WebPersona: A Life-Like Presentation Agent for the World-Wide Web. *Knowledge-Based Systems*, 11(1), 25-36.

Bolt, R. (1980). “Put-that-there”: Voice and gesture at the graphics interface. In *Proceedings of the 7th annual conference on Computer graphics and interactive techniques,* 262-270.

Burke, C., Doran, C., Gertner, A., Gregorowicz, A., Harper, L., Korb, J. & Loehr, D. (2003). Dialogue complexity with portability? Research directions for the Information State approach. In *the Research Directions in Dialogue Processing Workshop at the 2003 HLT-NAACL/NSF Human Language Technology Conference.*

Cassell, J., Stocky, T., Bickmore, T., Gao, Y., Nakano, Y., Ryokai, K., Tversky, D., Vaucelle, C., Vilhjálmsson, H. (2002). MACK: Media lab Autonomous Conversational Kiosk. *Proceedings of Imagina02*. February 12-15, Monte Carlo.

Dale, R., Geldof, S., & Prost, J.-P. (2005). Using Natural Language Generation in Automatic Route Description. *Journal of Research and Practice in Information Technology*, 37(1), 89-105.

Denis, M. (1997). The description of routes: A cognitive approach to the production of spatial discourse. *Current Psychology of Cognition* **16**, 409-458.

Emmorey, K., Tversky, B., & Taylor, H. A. (2001). Using space to describe space: Perspective in speech, sign, and gesture. *Spatial Cognition and Computation, 2(3)*, 1-24.

Evers, M., Theune, M., & Karreman, J. (2007). Which way to turn? Guide orientation in virtual way finding. In *Proceedings of the ACL 2007 Workshop on Embodied Language Processing*, 25-32.

Foster, M. E. (2004). Corpus-based Planning of Deictic Gestures in COMIC. In Belz, A.; Evans, R. & Piwek, P. (Eds.), *Proceedings of the Third International Conference on Natural Language Generation*, (pp. 198-204), Springer, Lecture Notes in Computer Science, Vol. 3123, 198-204.

Habel, C. (2003). Incremental Generation of Multimodal Route Instructions. In Proceedings of the AAAI Spring Symposium on Natural Language Generation in Spoken and Written Dialogue, 44-51.

Johnston, M. & Bangalore, S. (2000). Finite state multimodal parsing and understanding. In *Proceedings of the International Conference on Computational Linguistics (Coling),* 369-375.

Kopp, S., Tepper, P., Ferriman, K., Striegnitz, K. & Cassell, J. (2007). Trading Spaces: How Humans and Humanoids use Speech and Gesture to Give

Directions. In Nishida, T. (Ed.) Conversational Informatics: An Engineering Approach (133-160). John Wiley and Sons.

Kopp, S. & Wachsmuth, I. (2004). Synthesizing Multimodal Utterances for Conversational Agents. *The Journal Computer Animation and Virtual Worlds.* 15(1), 39-52.

Larsson, S. (2002). *Issue-Based Dialogue Management.* PhD thesis, Goteborg University.

Look, G., Kottahachchi, B., Laddaga, R. & Shrobe, H. (2005). A Location Representation for Generating Descriptive Walking Directions. In *IUI '05: Proceedings of the 10th International Conference on Intelligent User Interfaces*, 122-129.

Muller, P. & Prévot, L. (2008). Grounding information in route explanation dialogues. In Coventry, K., Tenbring, T. & Bateman, J. (Eds.) *Spatial Language and Dialogue*, Oxford University Press.

Shi, H. & Tenbrink, T. (2008). Telling Rolland where to go: HRI dialogues on route navigation. In Coventry, K., Tenbring, T. & Bateman, J. (Eds.) *Spatial Language and Dialogue*, Oxford University Press.

Sowa, T. & Wachsmuth, I. (2008). A Model for the Representation and Processing of Shape in Coverbal Iconic Gestures. In Coventry, K., Tenbring, T. & Bateman, J. (Eds.) *Spatial Language and Dialogue,* Oxford University Press.

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

Theune, M., Hofs, D. & van Kessel, M. (2007). The Virtual Guide: A direction giving embodied conversational agent. In *Proceedings of Interspeech 2007*, 2197-2200.

Traum, D. and Larsson S. (2003). The Information State Approach to Dialogue Management. In Smith & Kuppevelt (Eds.), *Current and New Directions in Discourse and Dialogue* (325-353). Kluwer.

statement type	**# of utterance units**
reorient	32
reorient+lm	24
move	51
move+lm	119
lm	367
dir	3
	597

Table 1: Distribution of statement utterance units over statement type.

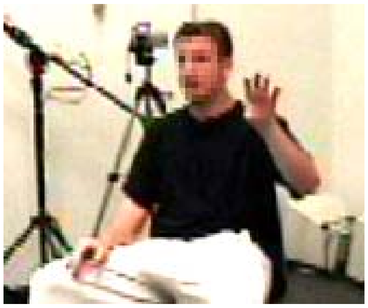


Fig. 1: ‘on your left once you hit this parking lot [is the Allen Center]’

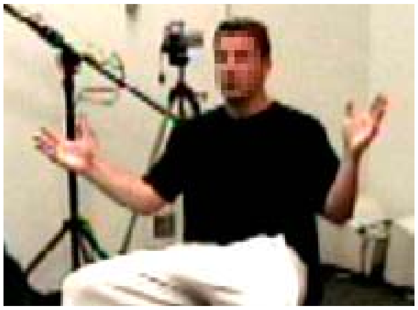


Fig. 2: ‘and [it’s really big]’

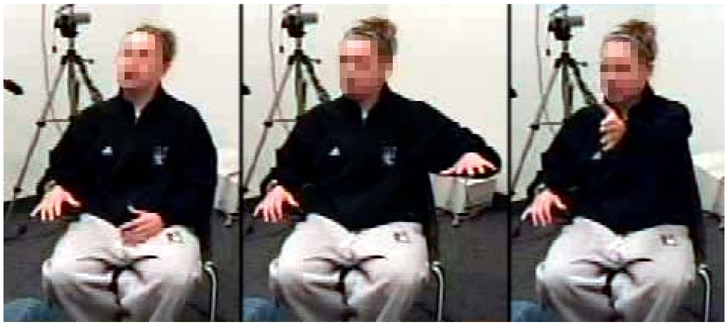


Fig. 3: ‘[University Hall] is on your right, [on the left is Kresge], and [then straight ahead is Harris]’

perspective	**# of gestures**	**%**
route perspective	185	53%
survey perspective	57	16%
non-locating	58	17%
locating wrt. speaker	7	2%
unclear/ambiguous	40	12%
	347	100%

Table 2: Distribution of perspective among gestures referring to landmarks.

type of statement	**# of survey perspective gestures**	**# of route perspective gestures**	**# of non-locating gestures**	**# of speaker perspective gestures**	**# of unclear / unambiguous gestures**
reorient	0	1	0	0	1
reorient+lm	1	4	1	0	0
move	0	0	0	0	0
move+lm	2	23	2	1	6
lm	54	157	55	5	33
dir	0	0	0	0	0
	57	185	58	6	40

Table 3: Distribution of gesture perspective over statement type.

	# of survey perspective gestures		**# of route perspective gestures**		**# of non-locating gestures**		**# of speaker perspective / unclear / unambiguous gestures**		
statement is A-clarif	32	+110%	31	-38%	18	+16%	12	-5%	93
statement is not A-clarif	25	-40%	154	+14%	40	-6%	35	+2%	254
	57		185		58		47		347

Table 4: Frequency of gesture perspective in answers to clarification questions.

	# of survey perspective gestures		**# of route perspective gestures**		**# of non-locating gestures**		**# of speaker perspective / unclear / unambiguous gestures**		
statement is elab	22	-13%	60	-27%	51	+98%	21	+1%	154
statement is not elab	35	+10%	125	+22%	7	-78%	26	-1%	193
	57		185		58		47		347

Table 5: Frequency of gesture perspective in elaborations.

	# of survey perspective gestures		# of route perspective gestures		# of non-locating gestures		# of speaker perspective / unclear / unambiguous gestures		
statement is repeat	16	+144%	13	-39%	7	+5%	4	-26%	40
statement is not repeat	41	-19%	172	+5%	51	-1%	43	+3%	307
	57		185		58		47		347

Table 6: Frequency of gesture perspective in re-descriptions.

	# of survey perspective gestures		**# of route perspective gestures**		**# of non-locating gestures**		**# of speaker perspective / unclear / unambiguous gestures**		
statement is plain	2	-90%	96	+48%	7	-66%	17	+3%	122
statement is not plain	55	+49%	89	-26%	51	+36%	30	-2%	225
	57		185		58		47		347

Table 7: Frequency of gesture perspective in plain statements.

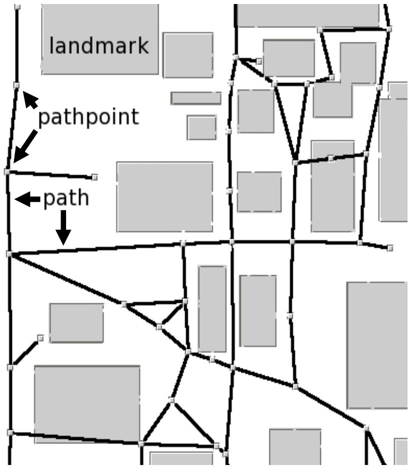


Fig. 4: Map representation showing pathpoints, paths, and landmarks.

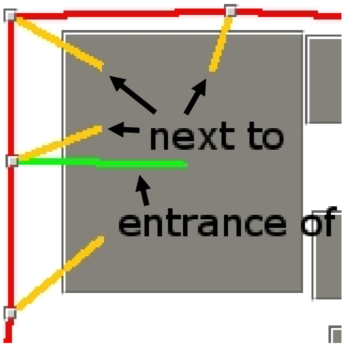
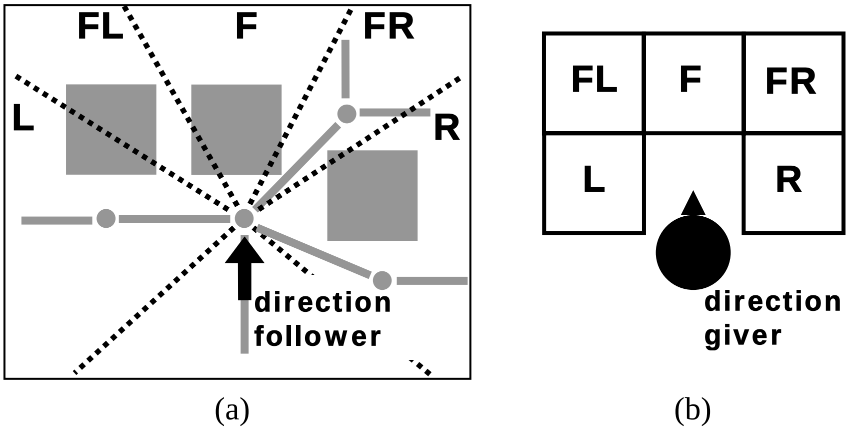
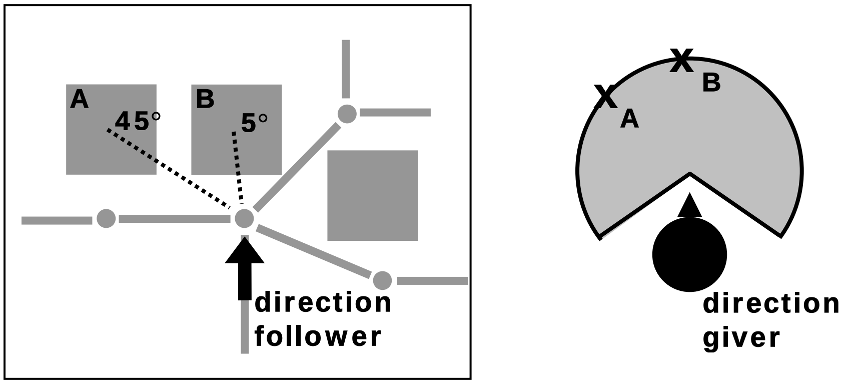


Fig 5: A landmark with qualitative relations to path points.



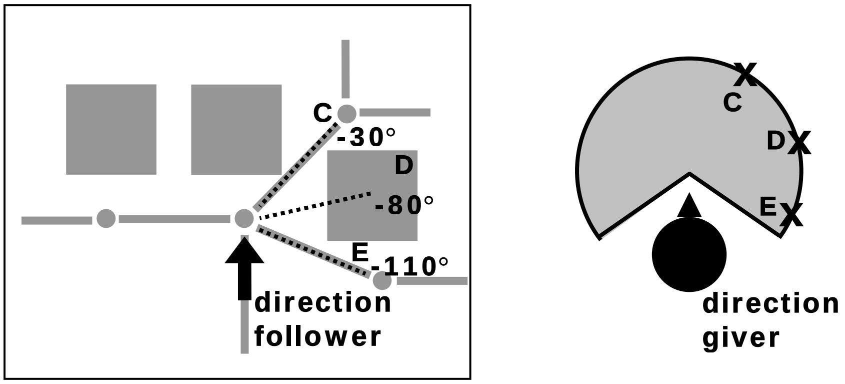
(a) (b)

Fig. 6: Route perspective gestures - mapping landmark location to positions in the gesture space.



(a) (b)

Fig. 7: Map gestures - mapping the location of buildings A and B to positions in the gesture space.



(a) (b)

Fig. 8: Map gestures - mapping the locations of building D and paths C and E to positions in the gesture space.

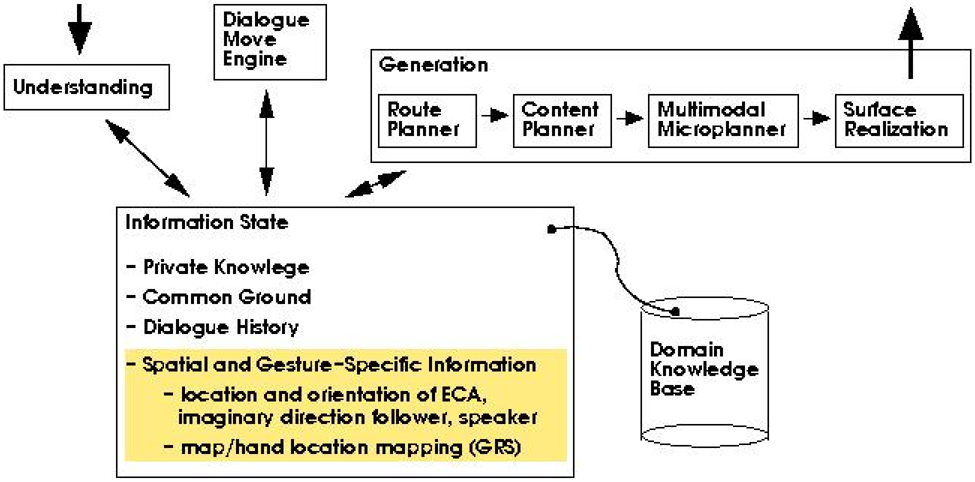


Fig. 9: Architecture of a direction giving ECA.

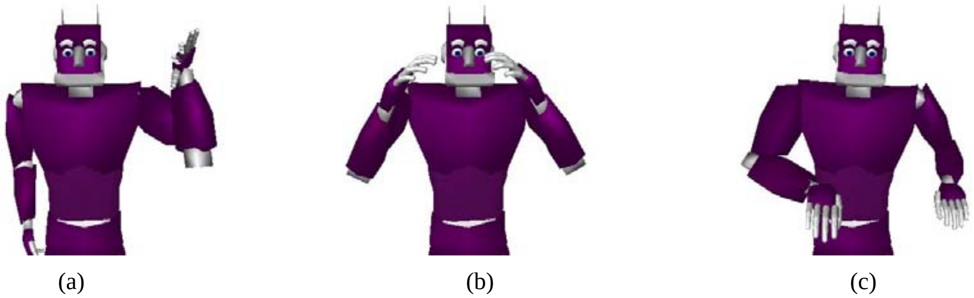


Fig. 10: NUMACK, our ECA, producing (a) a route perspective gesture, (b) a non-locating gesture, (c) a survey perspective gesture.