ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 34 (2015) 201–232

~~Computer~~
Speech
and
~~Language~~

The Roles and Recognition of Haptic-Ostensive Actions in Collaborative Multimodal Human-Human Dialogues

Lin Chen[a], Maria Javaid[b], Barbara Di Eugenio[a,1], Miloš Žefran[b]

[a] Natural Language Processing Lab, Department of Computer Science, University of Illinois at Chicago, Chicago, IL
[b] Robotics Lab, Department of Electrical and Computer Engineering, University of Illinois at Chicago, Chicago, IL

Abstract

The RoboHelper project has the goal of developing assistive robots for the elderly. One crucial component of such a robot is a multimodal dialogue architecture, since collaborative task-oriented human-human dialogue is inherently multimodal. In this paper, we focus on a specific type of interaction, Haptic-Ostensive (H-O) actions, that are pervasive in collaborative dialogue. H-O actions manipulate objects, but they also often perform a referring function.

We collected 20 collaborative task-oriented human-human dialogues between a helper and an elderly person in a realistic setting. To collect the haptic signals, we developed an unobtrusive sensory glove with pressure sensors. Multiple annotations were then conducted to build the *Find* corpus. Supervised machine learning was applied to these annotations in order to develop reference resolution and dialogue act classification modules. Both corpus analysis, and these two modules show that H-O actions play a crucial role in interaction: models that include H-O actions, and other extra-linguistic information such as pointing gestures, perform better.

For true human-robot interaction, all communicative intentions must of course be recognized in real time, not on the basis of annotated categories. To demonstrate that our corpus analysis is not an end in itself, but can inform actual human-robot interaction, the last part of our paper presents additional experiments on recognizing H-O actions from the haptic signals measured through the sensory glove. We show that even though pressure sensors are relatively imprecise and the data provided by the glove is noisy, the classification algorithms can successfully identify actions of interest within subjects.

Keywords: Haptic-Ostensive Actions, Multimodal Dialogues, Reference Resolution, Dialogue Act Classification

1. Introduction

In face-to-face human-human communication, interactions are spontaneously multimodal: interlocutors speak to each other, they use hand and body gestures, and eye gaze, to express their intentions. When people collaborate on physical or virtual tasks, they also act on objects, and on each other. Since these actions involve physical contact (touch), we will refer to them as haptic actions. It is widely held that interpreting the mix of audio-visual signals is essential to communicating [1]. Gestures and haptic actions are an important part of such dialogues. However, while the role of pointing gestures has been broadly explored [2, 3, 4], the role that haptic actions play in an interaction has not been investigated to the same extent.

[*] Corresponding author: Tel: +1 (312) 996-7566

Email addresses: linchen04@gmail.com (Lin Chen), mariajvd02@gmail.com (Maria Javaid), bdieugen@uic.edu (Barbara Di Eugenio), mzefran@uic.edu (Miloš Žefran)

Given the inherent multimodality of human-human collaborative dialogues, Human-Robot Interaction (HRI) needs to incorporate multimodality as well. A robot needs to be able to interpret and generate different kinds of gestures as well as haptic actions. Besides making the interaction natural, multimodal dialogues have other advantages [5]: multimodality reduces errors, makes the dialogue more robust because of *mutual disambiguation*, and it brings more efficient performance to various tasks.

In this paper, we focus on a specific type of interactions that has hardly been studied in the context of multimodal dialogues: **Haptic-Ostensive (H-O) actions** [6], haptic actions that manipulate objects and may simultaneously refer to the manipulated objects.[1] Our main contributions are two-fold:

1. **Corpus Collection and Analysis (Section 3)**. We present the ELDERLY-AT-HOME corpus of human-human collaborative dialogues, collected in the context of performing Activities of Daily Living (ADLs) [7]. ADLs are activities such as getting up from a bed or chair, getting dressed, preparing dinner etc., that are essential for a person to remain living independently. Our corpus analysis highlights the role of H-O actions. We believe this is the first extensive corpus analysis of H-O actions.[2]
2. **Multimodal Dialogue Processing (Section 4).** We then discuss two components of our dialogue processing model: the reference resolution module, and the dialogue act classifier. In both components, H-O actions and extra-linguistic information in general, play a crucial role.

Finally, we take our work one step further, and move beyond the realm of purely symbolic dialogue processing that is not connected to acting in the real world. Our reference resolution model and the dialogue act classifier include H-O actions, however those models are based on *humanly annotated* H-O actions. Needless to say, to make further progress in HRI, H-O actions need to be recognized from the haptic signals, namely, the dialogue processing work needs to be *grounded* in the real world. Section 5 presents our initial results in this regard. The experiments on recognizing H-O actions from haptic signals we describe are based on additional data we collected in the same realistic setting of the ELDERLY-AT-HOME corpus. We show how the models derived from this additional dataset can productively be used to recognize a large subset of the H-O actions relevant to the ELDERLY-AT-HOME corpus.

The experiments described in Section 5 support our contention that the extensive corpus analysis and modeling we did is not an end in itself, but has concrete potential for HRI, and not only in the sense that robots should take into account H-O actions. The fact that some H-O actions can be automatically recognized from human data is significant for a number of reasons. First, in building a corpus like the ELDERLY-AT-HOME corpus, manual annotation consumes significant amount of time and limits the size of the corpus. Automatic recognition makes the process much faster and the corpus potentially larger (as demonstrated by [8], that we will discuss below). HRI is largely driven by our understanding of human-human communication and corpora are the primary tool through which such a communication can be studied. Second, the prevailing paradigm for HRI, learning from demonstration [9, 10, 11], is heavily data driven. Devices like the glove and the automatic recognition algorithms we will discuss will play an important role in the process, even though they will not necessarily be used for everyday interaction.

2. Related Work

Research on spoken dialogue systems has been progressing for at least forty years, and many systems exist, from prototypes to commercial strength (please see [12] for a recent overview). Very early on, researchers realized that modalities other than speech play very important roles in interaction, and started integrating them into their computational models [13, 14, 15]. It is impossible to do justice to this large body of work within the space constraints of a paper. We will therefore focus on research that has addressed two important tasks within multimodal dialogue processing, referential expressions interpretation, and dialogue act recognition. Since our focus is on H-O actions, we will more broadly also review literature on the role of haptics in human communication, and its modeling within HRI.

Before we proceed, a clarification as concerns terminology. We will study two main types of extra-linguistic communication. The first is *deictic gestures* which *are generally understood as 'pointing gestures' that indicate real, implied or imaginary persons, objects, directions, etc., and are strongly related to their environment or 'gesture*

[1] Definition of *ostensive*: " Directly or clearly demonstrative" (http://www.oxforddictionaries.com/us/definition/american_english/ostensive).

[2] In due course, we will release the ELDERLY-AT-HOME corpus, which we believe will constitute a substantial contribution to the community.

space'.[3] Deictic (or pointing) gestures are normally conceived as not involving contact with the object pointed to. In contrast, *haptic actions*, the other important type of the multimodal interaction that we study, involve contact with an object as their inherent feature. H-O actions are a subset of haptic actions, which manipulate objects (as opposed to purely touching an object, e.g. pointing by making contact on a touch screen).[4]

Finally, we will talk about an *utterance* as *an uninterrupted chain of spoken or written language*[5] – namely, we will take as utterance the whole spoken contribution of one speaker before the floor moves to the other speaker (an utterance can then comprise zero, one or more units that can be termed sentences from a syntactic point of view).

Also note that whereas the models we will focus on in this section all rely on humanly annotated data, we will discuss annotation schemas later on in Section 3.2.

2.1. Multimodal Reference Resolution

Reference resolution aims at resolving *referring expressions* to the entities that they refer to (*referents*). The rationale for this task is that it is essential for many NLP applications to understand which entities are talked about, and that the same entity can be referred to through different linguistic referring expressions [16]. In fact, reference resolution for text is a standard module for many open source NLP toolkits: OpenNLP,[6] NLTK,[7] and GATE. [8]

Referring expressions in multimodal dialogue systems are more challenging to resolve, but also richer. Firstly, the utterances may be informal, ungrammatical, disfluent or incomplete; secondly, people spontaneously use hand gestures, body gestures, eye gaze and actions to interact with other people. To wit, in the following example from our corpus, the referent for *that one* is the pot that HEL(per) takes out from the cabinet; the referent for *there* is the cabinet which ELD(erly) is pointing to:

```
ELD: Can you get me a pot?
HEL: (opens cabinet, takes out pot)
ELD: Not that one, try over there (points to different cabinet)
```

As observed by many ([2, 3, 4, 17] *inter alia*), in a multimodal corpus, the antecedents of linguistic referring expressions are often introduced via gestures. Crucially from a computational point of view, including hand gestures information improves the performance of the reference resolution module [18, 19]. Sources of information different from hand gestures are important as well. For example, [20, 21, 22] integrated eye gaze into the reference resolution task. Subsequent experimental results showed a significant improvement when eye gaze was included in the feature set.

Researchers have mostly focused on the impact of *deictic gestures* on reference resolution. As mentioned, deictic gestures are normally understood as not involving contact with the object pointed to; however, a range of other interactions exist, where not only is an object touched, but manipulated in some fashion as well. In their groundbreaking work, Foster et al. [6] explored the role of **Haptic-Ostensive (H-O) actions** in a collaborative task.[9] They noted that these actions manipulate objects and can also be used to refer to the objects ostensively. They reported that about 36% of the initial linguistic references in their corpus (476 of 1333) were accompanied by such H-O actions. In the short excerpt from our corpus included above, both opening the cabinet and taking out the pot count as H-O actions, because they both introduce a potential referent, the cabinet and the pot, into the dialogue context.

We build on Foster et al. work [6]. We are the first to explore H-O actions in a completely natural setting – Foster et al.'s experiments take place either in a virtual shared space where two human subjects solve a tangram problem; or with one human interacting with a humanoid robot to solve a building block construction task. We are also the first to integrate H-O actions in computational models of both referring expressions and dialogue act recognition. Please note that whereas the role of haptic actions has been previously explored ([15, 24, 25] *inter alia*), these gestures are confined to making contact with a touch screen with pointing or "lasso" gestures.

[3] http://multimodalityglossary.wordpress.com/gesture/

[4] In reality, haptic actions are gestures and pointing gestures are actions. However, we use the term *action* only for haptics and the term *gesture* only for pointing gestures to keep with current terminology in linguistics and robotics, and to keep the references to the two distinct.

[5] http://www.oxforddictionaries.com/us/definition/american_english/utterance

[6] http://incubator.apache.org/opennlp/

[7] http://www.nltk.org/

[8] https://gate.ac.uk/

[9] But see [23] as concerns incorporating the action of picking up an object into reference resolution.

2.2. *Multimodal Dialogue Act Recognition*

Another component of many dialogue systems is Dialogue Act recognition, which attempts to uncover the speaker's intention behind an utterance. The notion of a Dialogue Act (DA) [26] is derived from the philosophical notion of a speech act [27]. In computational terms, DA recognition is often cast in terms of DA classification, in which a speaker's utterance is labelled with one (or more) of the possible DAs. As for reference resolution, DA classification in a multimodal environment is more challenging. First, an utterance may not directly follow another spoken utterance, but rather a pointing gesture or a haptic action. Likewise, the next move is not necessarily a spoken utterance, it can be a pointing gesture or a H-O action only, or a multimodal utterance. Third, when people use gestures and actions together with utterances, the utterances become shorter, hence the textual context that has been used to advantage in models of written text is impoverished.

Several corpora tagged for dialogue acts have been produced ([28, 29, 30, 31] *inter alia*), and a large body of research has addressed the problem of recognizing DAs from linguistic form, including [32, 33, 34, 35]. Some of these models also include acoustic features, for example prosody, and show that DA recognition improves when such features are included [36, 37, 38].

As for reference resolution, we focus on models of DA recognition that include extra-linguistic features, which have more recently started to attract interest. For example, [39] discovered that integrating facial expression could significantly improve the recognition of several dialogue act categories, whereas [40] showed that automatically recognized postural features may help to disambiguate DAs.

Not many researchers have explored the impact of gestures on DA classification per se, however some papers explore the role of gestures for recognizing intentions to e.g. perform specific actions in the domain.[10] For example, in [41] deictic gestures are used to recognize users' intentions to perform actions (some linguistics, some domain) in a visual interior decoration application.

Other researchers have investigated additional extra-linguistic features, in particular features of the context. Semi-automatic DA classification was applied in [8] for situated social agents in a restaurant game: they trained a classifier using 2% of the data, and used the trained model to annotate the remaining 98% of the data. The final model showed strong predictive power and included the role of the person (waitress or customer), their general posture (sitting or standing), who they are facing, and their location. As we will see, a triad of extra-linguistic features which includes pointing gestures, H-O actions, and location will be included in our best DA classification models. The study in [8] is notable for a second reason: they further investigated whether [p. 157] *dialogue acts are useful building blocks for learning patterns of interleaved utterances and physical actions*. It is our goal as well, to ultimately develop a model that can decide "what to do next", be it an utterance, possibly accompanied by a gesture, or an action in the domain.

2.3. *Haptic Recognition*

The literature on the role of haptics and touch in human communication can be broadly divided into two categories. The first studies communication through touch devices and primarily focuses on pointing and touch gestures that can be detected by such devices. Design of touch interfaces is studied in e.g. [42, 43, 44, 45], and the integration with speech has been addressed in e.g. [15, 24, 25] as mentioned earlier. Vibro-tactile displays provide an interesting extension of this work to bi-directional haptic communication (see e.g. [46]). We again emphasize that the focus in this work are haptic actions in general, and more specifically H-O actions. They involve touch, but they are distinct from gestures used with touch displays.

The second body of work studies the role of touch in social interaction. It is well documented that humans communicate emotions through touch [47]. Several groups have developed artificial skin that can sense such interactions and proposed various approaches to interpret them [48, 49, 50, 51, 52]. Robot companions that provide social interaction have been proposed by [53, 54]. Our work is different since we are interested in collaborative rather than social interactions.

Limited options exist for collecting data on actions that involve touch. Pressure or other sensors mounted either on a hand/glove or embedded in objects have been used to recognize haptic information. For example, in [55], a sensory glove has been used to recognize the type of grasp used by the subject. In [56], finger joint angles of a robotic hand

[10] There is of course a vast literature on recognizing human motion. However, here we are specifically focusing on deictic gestures and H-O actions.

are used to identify a set of 25 different objects, and in [57] the information from 45 pressure sensors mounted on Lucs Haptic Hand II is used to classify six objects of different shapes. An array of pressure sensors has been used in [58] to recognize a set of toys. In [59, 60], an object with embedded sensors was used as a user interface to virtual environments. Some of these studies rely on various machine learning techniques for recognition, but none of them studies H-O actions.

3. The ELDERLY-AT-HOME corpus

3.1. Data Collection

The RoboHelper Project [61], as the larger context of this research, focuses on the elderly care domain. Its ultimate goal is that of building assistive robots that can help the elderly remain living independently at home. Such a robot needs to be able to help the elderly person perform Activities of Daily Living (ADLs) [7], which include getting up from a bed or chair, getting dressed, preparing dinner etc. These activities all require physical interaction between the elderly and the robot. To develop our models we had to collect a new corpus since in this domain no multimodal dialogue corpus exists which includes haptic data relevant to the interaction.[11]

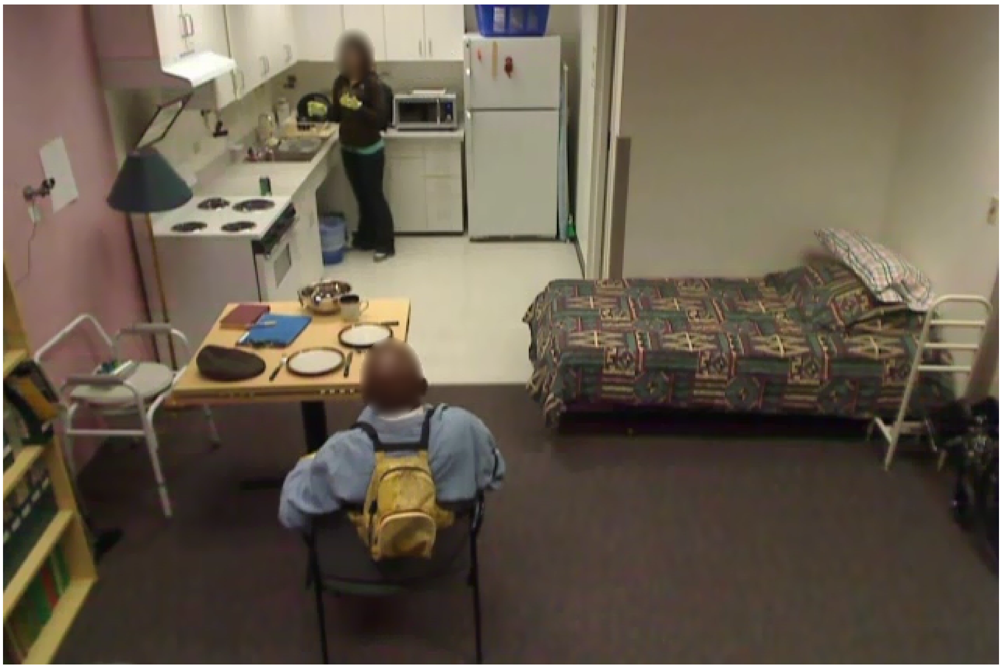


Figure 1. Data Collection Experiments

The data collection experiments were conducted in a fully functional studio apartment at Rush University in Chicago. Each experiment involved interaction between an ELDer and a HELper. The experiments focused on four ADLs: ambulating; getting up from a bed or a chair; putting on shoes; and preparing dinner. Undoubtedly, preparing dinner is the most complex ADL, since it comprises many subtasks, such as finding pots, opening cans and containers, putting pots on a stove, cooking pasta, draining the pot, setting up the table, and cleaning the table afterwards. Two students in gerontological nursing played the role of the HELper, both in pilot experiments and the experiments with real subjects. In five pilot experiments, two faculty members played the role of the elderly person. In the fifteen real experiments, the ELDer resided in an assisted living facility and was transported to the apartment mentioned above.

[11] Multimodal corpora that target ADLs such as CMU-MMAC [62] or CAET [63] are rich with signals of different sorts, e.g. motion capture or depth information. However, they are either limited as concerns the kinds of activities they focus on and/or do not include haptic signals.

All elderly subjects were highly functioning at a cognitive level and did not have any major physical impairment. Figure 1 shows a video frame from the experiment recordings. We[12] equipped the room with seven cameras to ensure multiple points of view. During the experiments, both participants were wearing a microphone, and a data glove on their dominant hand to collect haptics data.

We employed two gloves: the X-IST commercial glove, developed by noDNA and worn by the elderly subject, and the glove that we developed, worn by the helper. During pre-pilot experiments (not included in the corpus), we realized that the X-IST glove was not providing us with enough signals: this glove includes pressure sensors only on its fingertips, but when a subject grasps an object, the fingertips actually don't really touch the object, or at least, not with enough force that a signal can be transmitted. An excellent overview of various other sensory gloves can be found in [64]. They are primarily targeted at motion capture, none of them is suitable for collecting haptics data. Hence we developed our own glove by attaching FlexiForce pressure sensors (Tekscan, USA) to a cotton glove (we developed only one because fabricating the glove required considerable effort). These sensors are thin and light, and favorably compare to other similar sensors in terms of precision and linearity [65, 66]. The sensors were placed on every segment of each finger except for the middle segments of the thumb and the pinkie. We also placed four pressure sensors on the palm. In total, 17 pressure sensors were attached to the glove (Figure 2). Additionally, a 6 degree of freedom inertial measurement unit (ITG3200/ADXL345, SparkFun Electronics, USA) was used to capture hand tilt and acceleration. It was attached to the back of the palm. The glove is connected to a processor box based on Arduino Mega microcontroller [67] through two 20 wire cables. All the electronics is placed in a small backpack that the subject wears during the experiments (please see Figure 1). After reading sensor values, the microcontroller transmits the data wirelessly to the computer using an Xbee module. The computer receives the glove data through a USB XBee receiver module and stores it in a file with a time stamp for each data sample. The data is sampled at around 70Hz.

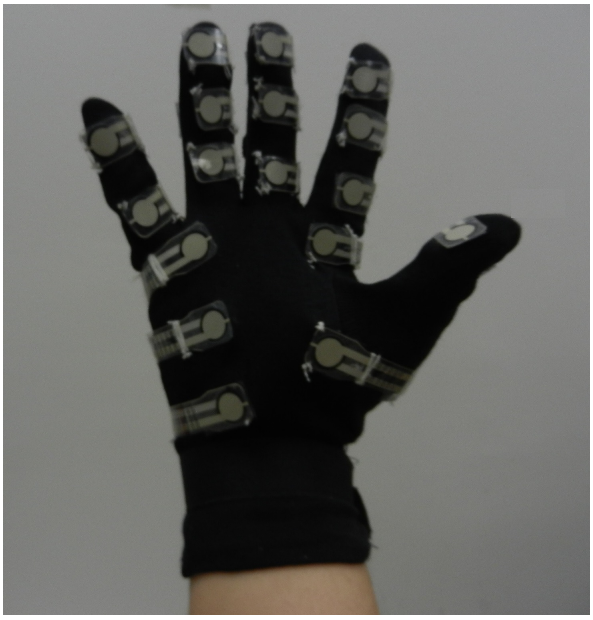


Figure 2. The sensory glove developed in-house

The size of the collected video data is shown in Table 1. The number of subjects refers to the number of different ELD's and does not include the HEL. The five pilot experiments were included though, since the interactions in those pilot experiments did not measurably differ from those with the real subjects. Usually one experiment lasted about 50 minutes (recording started after informed consent and after the microphones and data gloves were put on). Further,

[12] "We" refers to all the project members of the RoboHelper project, not just to the authors of this paper, since data collection was the result of a collective effort from all [61].

we eliminated irrelevant content such as interruptions, e.g., by the person who accompanied the elderly subjects, and further explanations of the tasks. This resulted in about 15 minutes of what we call *effective* data for each subject; the effective data comprises 4782 utterances (see Table 1) (recall that an utterance corresponds to one speaker's spoken turn, not to one single sentence).

Subjects	**Raw(Mins)**	**Effective(Mins)**	**Utterances**
20	482	301	4782

Table 1. RoboHelper Distributional Statistics

The Find SubCorpus. As observed in the experiments, one type of activity was common to the majority of the subtasks preparing dinner consists of: the subjects were frequently involved in finding and retrieving different types of objects (pots, silverware, pasta) from all kinds of containers (drawers, cabinets, fridge). Such tasks were labeled as "*Find*" tasks and are the focus of this work. Apart from their frequency, another reason for focusing on the "*Find*" tasks is to avoid possible safety issues. Tasks such as helping the elderly person ambulate or putting on shoes would require the robot to directly touch the elderly; while these issues are of much interest to the HRI community, they are outside the scope of the work discussed in this paper.

We define a *Find* task as a continuous time span during which the two subjects are collaborating on finding objects. A *Find* task could be initiated by a request from ELD or by a suggestion from HEL, and ends when they found what they wanted, or (rarely), gave up. We manually annotated for the beginning and end of a *Find* task (see track *FindTask* in Figure 4, to be discussed shortly). 137 *Find* tasks were extracted. The vast majority of them succeeded, with only four failing: after several rounds of searching, the subjects still could not find what they wanted, and they agreed to use an object other than what they had planned to use. The total length of the 137 *Find* tasks is about 1 hour and 24 minutes. This subcorpus comprises 1516 utterances, which include 6593 words (see Table 2). Collectively, the 137 *Find* tasks account for about 28% of the total length of our effective dialogues, and for about 32% of the total number of utterances. It appears that these subtasks elicit slightly more interaction than other tasks.

Figure 3 shows a *Find* task example, where the subjects are collaboratively searching for a spoon in the kitchen. The interaction has been annotated for pointing gestures, H-O actions and their targets, as we will describe next. In the following, we will use Utt_i to refer to the spoken utterance in one turn, and M_i to refer in general to what occurs in one turn of one participant. Whereas for a given i we sometimes have only either spoken utterance or pointing gesture or H-O action, utterance and either pointing gesture or H-O action can co-occur, for example for turns 1, 2, 4, 8, 12 in Figure 3. In Appendix B, we describe how we recognize whether a pointing gesture or H-O action is independent of or associated with a spoken utterance.

3.2. The annotation process

We used Anvil [68], a multimodal annotation research tool, to process the collected data. Figure 4 shows the annotation of the first 6 turns from Figure 3 (please note that the indices associated with utterances, pointing gestures and H-O actions in Figure 4 are internally generated and do not match those used in Figure 3). Several types of annotations were performed in different dimensions. Each dimension is implemented as a *track* in Anvil: these are the blue rows in Figure 4. Further, attributes are used to label various features of a track, and are shown as labels in rectangles on the track. The size of the rectangle corresponds to the time span of the attribute. One track was used to mark *Find* tasks. For a *Find* task, its start and end points, and the object which the subjects were trying to find during the task are marked (see track *FindTask* in Figure 4: the type of object looked for is *spoon*).

Two RoboHelper project members transcribed the spoken dialogues in the recorded experiment videos. The speech of ELD and HEL was separated into different tracks in Anvil, as shown by the *Utterance* track in Figure 4 which is subdivided into three: the HEL(per), the ELD(erly) and the mediator.[13] The speech was segmented into utterances

[13] Recall that in our discussion of effective dialogue, we eliminated interventions by third parties between tasks; however this track has been defined in case those interventions need to be ultimately annotated.

1	ELD	And there is a spoon down there, in the second drawer? *[Point(ELD,Drawer1)]*
2	HEL	*[Point(HEL,Drawer1)]* Down there?
3	ELD	Yes.
4	HEL	This? *[Touch(HEL,Drawer1)]*
5	ELD	Uh-huh.
6	HEL	*[Open(HEL,Drawer1)]*
7	ELD	A spoon.
8	HEL	*[Takeout(HEL,spoon1)]* Is this the spoon?
9	ELD	No, the second drawer.
10	HEL	*[Close(HEL,Drawer1),Open(HEL,Drawer2)]*
11	ELD	Yes, there it is.
12	HEL	*[Takeout(HEL,spoon2)]* This one?
13	ELD	Yes, uh-huh.
14	HEL	OK.

Figure 3. Find Task Example

(turns) according to pauses. Due to reasons, such as mumbling, speech not clear, and the words the subjects used, etc., the transcribing process took more than three hours for each subject.

Two types of annotations were performed on the transcribed data: linguistic as concerns referring expressions, their antecedents, and dialogue acts; extra-linguistic features in the form of pointing gestures, H-O actions and miscellaneous features such as subject location. To annotate extra-linguistic features, annotators looked at the videos, and for H-O actions, also at the haptic data recordings (when available, please see Section 5).

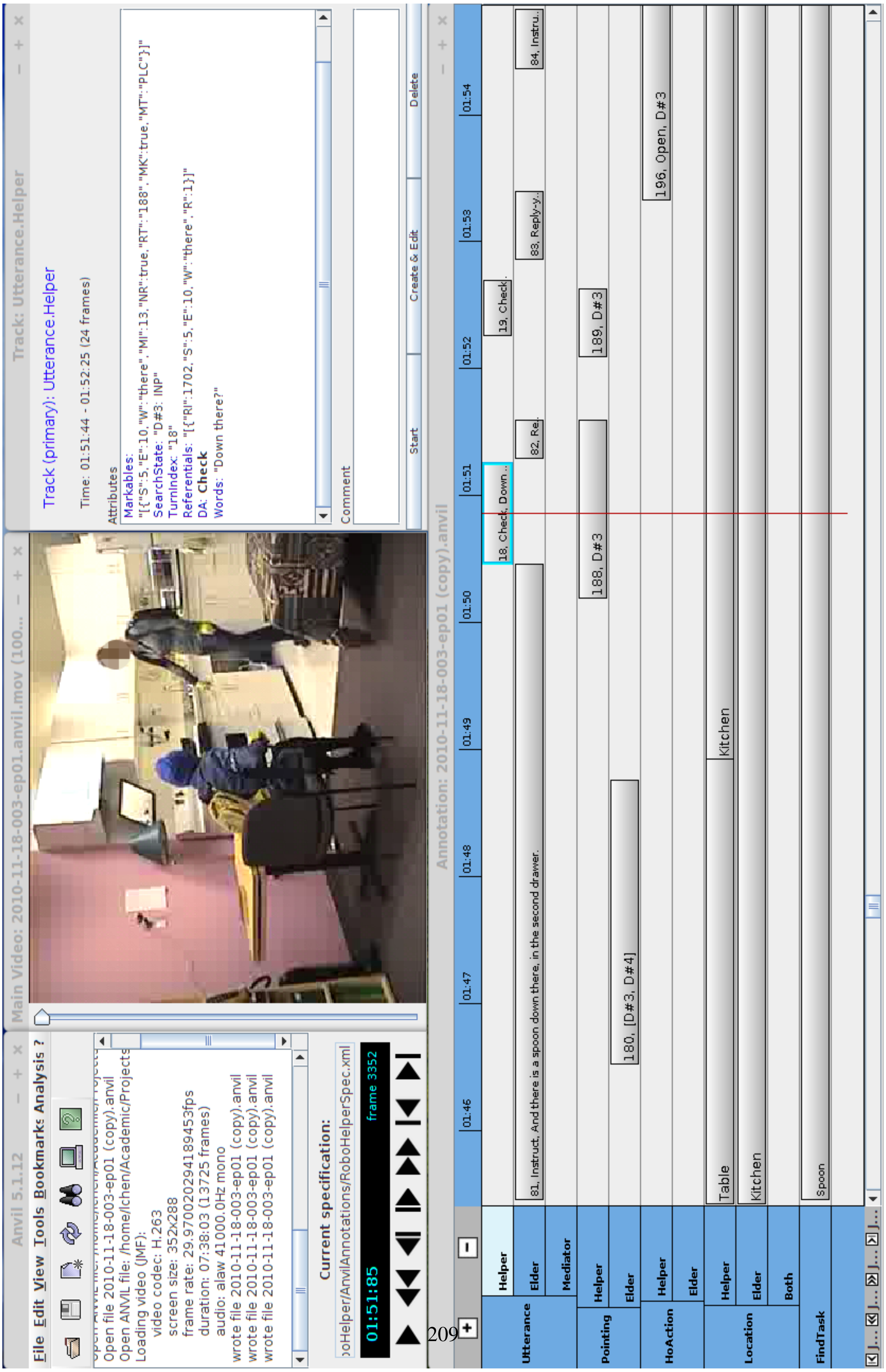
It is not possible to fully explain all the components of the Anvil interface within the space constraints of a paper. We will just note that annotations such as *[D#3, D#4]* refer to the objects pointed to, or manipulated; and that the top right window shows the current item being annotated: turn 2 by the helper (*Down there?*) and its features, which include markables, referential expressions and DA (Dialogue Act), to be discussed in Section 3.2.1.

Table 2 shows the counts of different events divided by type of participant (words were extracted automatically from the transcribed dialogues, gestures and H-O actions were annotated as discussed below). It is apparent that: a) pointing gestures and H-O actions were frequently used. Their total corresponds to 61% of the number of spoken utterances. b) ELD performed 66% of pointing gestures, and HEL performed 97.5% of H-O actions.

	ELD	**HEL**	**Total**
Utterances	756	760	1516
Words	3612	2981	6593
Pointing	219	113	332
H-O Actions	15	582	597

Table 2. Multimodal Event Statistics in the *Find* Corpus

Before we proceed further, we should clarify why, rather than using any existing annotation scheme, we developed our own, albeit relying on previous efforts such as [69] for referential expressions and MapTask for dialogue acts [28]. Indeed rich multimodal annotation schemes exist, for example MUMIN [70], the Behavior MarkUp Language [71, 72] or the ISO 24617-2 standard [73]. The issue is that multimodality is an extremely complex phenomenon, and understandably, none of these schemes is all encompassing. For example none of the just referenced schemes include reference annotations of the kind we need here (although gestures can be semantically linked to their referents in MUMIN [4]). As far as gestures are concerned, even if these gesture annotation schemes are extremely rich as concerns annotating e.g. the shapes of the hands or other functional features, as far as we know, they do not address


Anvil 5.1.12
File Edit View Tools Bookmarks Analysis ?
Main Video: 2010-11-18-003-ep01.anvil.mov (100...
Track: Utterance.Helper
Annotation: 2010-11-18-003-ep01 (copy).anvil

```
Open file 2010-11-18-003-ep01 (copy).anvil
Open ANVIL file: /home/lchen/Academic/Projects
Loading video (JMF):
    video codec: H.263
    screen size: 352x288
    frame rate: 29.9700200294189453fps
    duration: 07:38:03 (13725 frames)
    audio: alaw 41000.0Hz mono
wrote file 2010-11-18-003-ep01 (copy).anvil
wrote file 2010-11-18-003-ep01 (copy).anvil
wrote file 2010-11-18-003-ep01 (copy).anvil
```

Current specification:
oHelper/AnvilAnnotations/RoboHelperSpec.xml

01:51:85 frame 3352

Track (primary): Utterance.Helper

Time: 01:51:44 - 01:52:25 (24 frames)

Attributes
Markables: "[{"S":5,"E":10,"W":"there","MI":13,"NR":true,"RT":"188","MK":true,"MT":"PLC"}]"
SearchState: "D#3: INP"
TurnIndex: "18"
Referentials: "[{"RI":1702,"S":5,"E":10,"W":"there","R":1}]"
DA: **Check**
Words: "Down there?"

Comment

Figure 4. The Anvil annotation interface (as applied to Figure 3)

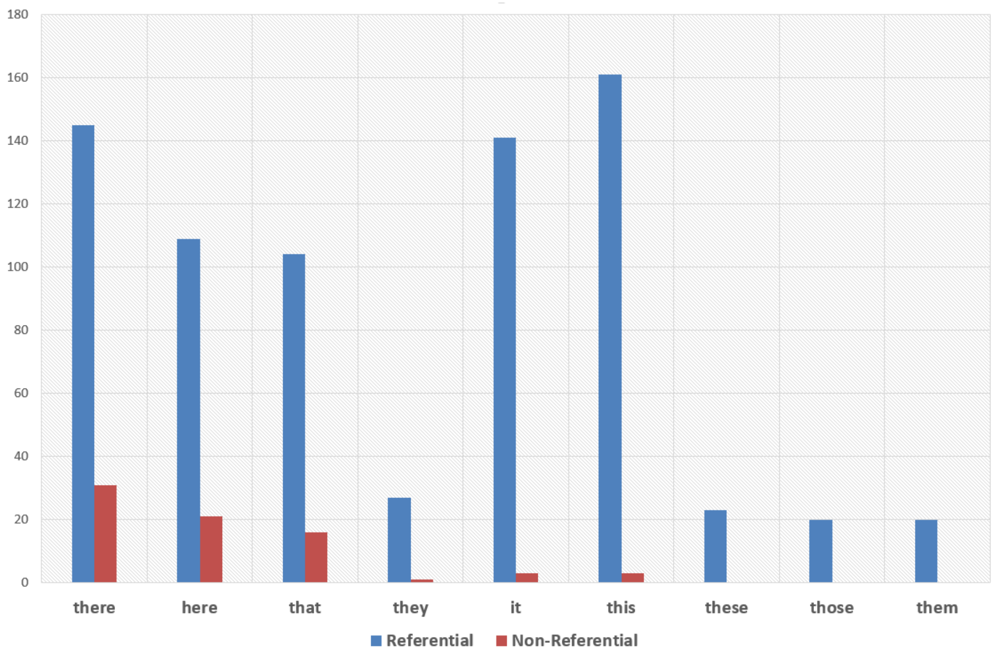


Figure 5. Third Person Pronouns and Deictics Distributions

manipulative / referential function for haptic interaction (please see [74] for a very recent review of annotation schemes that focus on gestures). As far as dialogue acts are concerned, some schemes such as DiaML (part of the ISO 24617-2 standard [73]) decompose utterances along a variety of dimensions, for example communicative function versus realm (e.g. *task* vs *social obligation*). We still resort to a DA annotation that consists of primitive labels for several reasons: first, annotating each utterance along several dimensions adds to the burden the annotators face, and it would have required more resources that we could allocate to the coding task; second, we wished to compare our work with much recent work that uses machine learning to develop DA classifiers, where DA labels are considered as primitive units. Additionally, DA inventories like MapTask were specifically developed for task-oriented dialogues, as opposed to say DiaML or the older DAMSL-Switchboard [31], which attempt to cover all types of conversations.

3.2.1. Linguistic Annotations

Coreference Annotation was done in order to support reference resolution, which as discussed in Section 2.1, aims at resolving *referring expressions* to the entities that they refer to (*referents*). In essence, *Coreference Annotation* concerns linking a referring expression to its closest referent in the data.

Our choices were based on the following distributional considerations. The subjects in our experiments frequently used third person pronouns (*it, them, they*) and deictics (*here, there, this, that, these, those*). In the *Find* corpus, 827 out of 6593 words (12.5%) are third person pronouns and deictics. The distributions for different third person pronouns and deictics are shown in Figure 5. Note that deictics can both be used as pronouns (e.g., all usages of *this* in Figure 3) or as adjectives, as in *that big red bowl*. Figure 5 does not distinguish between deictics used as pronouns or as adjectives, since they need to be resolved in both cases. However, Figure 5 distinguishes referential usage from non referential usage. This is necessary because pronouns and deictics are also used pleonastically: e.g. in *Yes, there it is* (Utt_{11} in Figure 3), *there* is not referential, but *it* is referential.

The coreference annotations followed a similar approach to [69], and included two phases of annotation: markable annotations and coreference annotations proper. Simply put, markable annotation labels potential referential expressions, and potential antecedents for referential expressions. In our case, markables comprise certain phrases (see below), and all targets of pointing gestures and H-O actions (to be described in the next section). While the markable annotations were semi-automatic, the coreference annotations were completely manual.

Type	Count
OBJ (Object)	762
PLC (Place)	616
NEITHER	262

Table 3. Markable Type Counts

Type	Count
Pronouns and Deictics	827
Resolved Pronouns and Deictics	757
Textual Antecedent	218
Pointing Gesture Antecedent	266
H-O Antecedent	273

Table 4. Multimodal Coreference Annotation Statistics

To annotate for textual markables, the shallow parser from Apache OpenNLP Tools was used to chunk the utterances into phrases. Most of the Noun Phrases (NPs) and Adverb Phrases (ADVPs) were automatically marked as markables. NPs or ADVPs which contain first and second person pronouns, or questions words like "what", "which" and "where" were excluded.

After the automatic step, for each utterance, the accuracy of the automatically generated markable candidates was checked. If any markable was missed, a markable would be inserted using the same notation format, with an auto-increment markable index. Textual markables were also annotated for type, according to the distinction between PLACE (unmovable objects) and OBJECT (movable objects), to be discussed more fully below. The counts for different types of markables are shown in Table 3.

If a markable is a pronoun or a deictic, and hence a referential expression, then the annotators marked its coreference target, i.e. the closest markable to which it refers.

The statistics for coreference annotations in the *Find* corpus are shown in Table 4. Note that for only 757/827 (92%) were the annotators able to determine an antecedent. Interestingly, 71% of those 757 pronouns or deictics referred to antecedents that had been introduced *exclusively* by pointing gestures or H-O actions (about 35% each). This is an important point: pointing gestures and H-O actions need to be taken into account by any reference resolution model, since if we neglect them, we would not be able to take into account their targets as potential referents for referential expressions.

Dialogue Act Annotation. To understand how multimodality can affect dialogue, utterances were labeled using DA labels. We adapted the DA inventory used by in the MapTask corpus [28]. [14] As we mentioned earlier, whereas other DA inventories are available, we chose the MapTask inventory because it is simpler, and was developed for task-oriented collaborative tasks.

The MapTask original set includes six initiating move labels and four response move labels.[15] Initiating move labels include: *instruct*, *explain*, *check* (when a speaker asks their interlocutor for confirmation), *align* (which checks the attention or agreement of the partner), *query-yn*, *query-w*. Response move labels include: *acknowledge*, *reply-y*, *reply-n*, *reply-w*.

However, the MapTask label inventory of DAs does not cover utterances that are used to respond to gestures and actions, such as Utt_{11} in Figure 3. Hence, we devised three more tags, which apply **only** to statements that follow a move composed exclusively of a pointing gesture or an H-O action, like M_{10}: **state-y**, a statement which conveys "yes", such as Utt_{11} in Figure 3; **state-n**, a statement which conveys "no", e.g., if Utt_{11} had been *Wait, try the third drawer.*; **state**, still a statement, but not conveying acceptance or rejection, e.g., *So we got the soup.*

[14] http://groups.inf.ed.ac.uk/maptask/interface/expl.html#moves

[15] A twelfth pre-initiating move, *ready*, did not occur in our data; neither did a fifth response move, *clarify*.

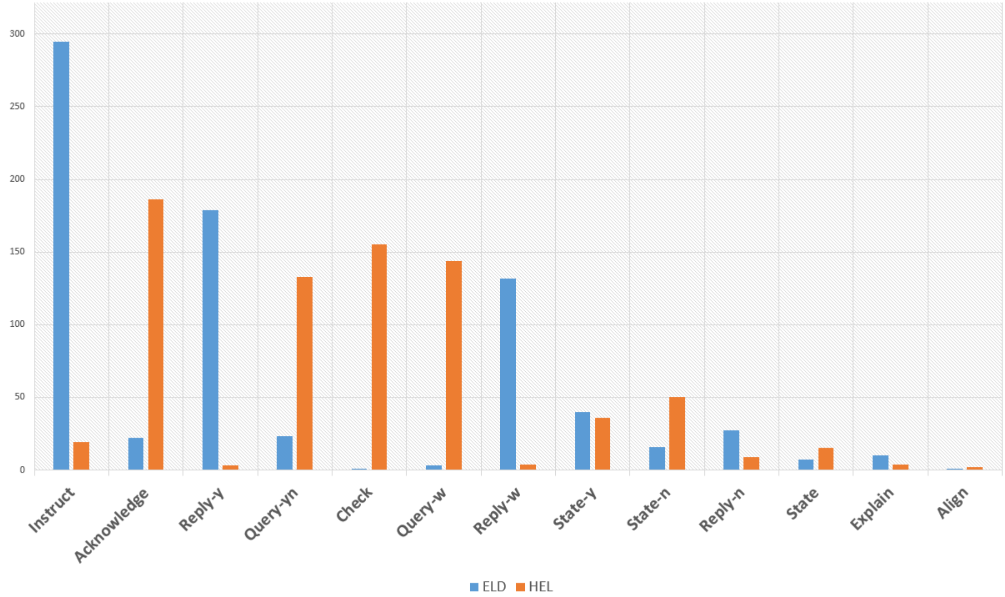


Figure 6. Dialogue Act Distribution, per Actor

Hence, the new DAs {*state-y*, *state-n*, *state*} are used to tag responses to actions, and the original DAs in {*reply-y*, *reply-n*, *reply-w*} are used to tag responses to utterances. The distributions of DAs which appear in the *Find* corpus is provided in Figure 6. The most frequent DA is *Instruct*, which accounts for 20.7% of the total DAs. As the figure shows, for almost every dialogue act, the frequency of usage is quite different according to the actor (ELDerly person or HELper).

3.2.2. Extra-Linguistic Annotations

We annotate for pointing gestures, H-O actions and subjects' locations.

Pointing Gestures are defined as hand gestures (of the left or right hand) that indicate the targets of the gestures, and for which there is no physical contact between the hand and the targets. The latter is the main distinction between pointing gestures and H-O actions, since it is a defining feature of H-O actions that there is physical contact with the manipulated objects. Among pointing gestures, we did not include head movements, or other body part movements, which can also be used to indicate targets, since such occurrences are extremely rare in our corpus.

For a pointing gesture, two attributes are marked: the time span (indicated by the width of the corresponding rectangle), and the target. The time span of a pointing gesture starts when the subject initiates the hand movement; and ends when the subject starts to draw back the hand. To mark the physical targets of pointing gestures, a referring index system was designed as follows.

- A compile time referring index is assigned to each object which can be considered as a PLACE, i.e., that belongs to one of the following two categories:
  - Objects with fixed locations, like cabinets, drawers, microwaves, fridges, etc. For example, in Figure 4, all the indices of the form *D#N* where *N* is an integer, refer to different drawers.
  - Objects which can be moved, but usually are not moved during the experiments, like the table
- A run time referring index is assigned for each object which is movable, and has many different instances, such as cups, glasses, etc. The run time index is given during the annotation process, according to the object type and the order of appearance in the experiment context. An example of the run time index is "Cup#1", which means the first cup appearing in the experiment.

The most common type of pointing gesture has one identifiable target, which is usually indicated by a short time stable hand pointing – see *D#3* in the track *Pointing/Helper* in Figure 4. In some cases, pointing gestures have more than one identifiable target. In this case, a set of referents is created, such as *[D#3, D#4]* in the track *Pointing/Elderly*. This means the subject was rapidly moving the target from Drawer#3 to Drawer#4. Finally, there are pointing gestures for which potential targets are not identifiable: '*' is then used as the referring index. For example, the notation "#*" means the subject initiated a pointing gesture towards a totally unidentifiable target.

Haptic-Ostensive Actions. As mentioned, during the collaborative tasks which we are interested in, subjects physically interact with objects. Those physical contact actions often perform a referring function as well, by either adding new entities to the discourse model or referring to an already established referent. For example, in Figure 3, the action [Touch(HEL, Drawer1)] that accompanies Utt_4 disambiguates *This* by referring to Drawer1, tantamount to a pointing gesture; conversely, the action [Takeout(HEL, spoon1)] associated with Utt_8 establishes a referent for spoon1. Following [6], we label Haptic-Ostensive (H-O) those actions that involve physical contact with an object, and that can at the same time perform a referring function. Note that target objects here exclude the partner's body parts, as when HEL helps ELD get up from a chair.

No existing work that we know of identifies types of H-O actions. Hence, we had to define our own categories, based on the following two principles: (1) The H-O types must be grounded in our data, namely, the definitions are empirically based: these H-O actions are frequently observed in the corpus. (2) They are within the scope of what we envisioned could be recognized from the haptic signals. As we will see in Sec. 5, the two sets of actions (the annotated types and the recognized types) do coalesce to a certain degree. The five H-O action types we defined are:

- *Touch*: When the subject only touches the targets, no immediate further actions are performed
- *Grasp-Show*: A grasp occurs when the subject takes out or picks up an object, holds it stably for a short period of time, and (in the judgement of the annotator) intentionally shows it to the other subject
- *Grasp-No-Show*: This occurs when the subject takes out or picks up an object but (in the judgement of the annotator) the subject does not intentionally show it to the other subject
- *Open*: The open process starts when the subject has physical contact with the handle of the fridge, a cabinet or a drawer, and starts to pull; ends when the physical contact is off
- *Close*: The close action starts when the subject has physical contact with the handle of the fridge, a cabinet or a drawer, and starts to push; it ends when the physical contact is off

For H-O action annotation, both the video and the haptic data recordings (when available) were inspected by the annotators. Three attributes are marked: time span, target and action type. The time span starts when the action is initiated and ends when the action ends. The "Target" attribute is similar to the "Target" attribute in pointing gesture annotation – see examples in the *HoAction/Helper* track in Figure 4. Since H-O actions are more accurate than pointing gestures [6], the targets are all identifiable. Table 5 shows the distribution of H-O action types. Not surprisingly given the nature of the task, ELD performs very few H-O actions.

	ELD	**HEL**	**Total**
Touch	3	33	36
Grasp-Show	0	71	71
Grasp-No-Show	4	79	83
Open	5	227	232
Close	3	172	175
Total	15	582	597

Table 5. H-O actions types

Category	Kappa
Coreference	0.700
Dialogue Act	0.789
H-O actions	0.703
Pointing Gestures	0.751

Table 6. Kappa values for different annotation

Subject Location Annotations. It is intuitive that subjects' different locations will trigger different dialogue behaviors. For instance, if two subjects are together, and one subject points to a close target, there will be no ambiguity for the other subject; hence no clarification about the pointing gesture will be raised.

Because of the absence of 3-D modeling for the room, the subject location annotation is only based on experiment videos. The room was divided into four areas: table, bed, kitchen, and lounge. Among those four areas, table and kitchen areas were the most frequently used areas.

Three tracks were used in Anvil for the subject location annotation: one track for each subject, and if both subjects appeared in the same location, a "Both" track was used (see Figure 4).

3.3. Intercoder Agreement

All the annotations just discussed rely on humans to make subtle distinctions. To test the reliability of human annotations, the standard approach is to have multiple coders annotate the same subset of the corpus, and then compare the annotations in the double coded part to calculate the agreement. To evaluate intercoder agreement, the kappa coefficient of agreement, κ, has become the de facto standard for tagging tasks in recent years [75, 76, 77]. Cohen's Kappa [78] and Fleiss' Kappa [79] are two main methods for calculating κ. In this research, we used Cohen's Kappa.

15% of our data was double coded for pointing gestures, H-O actions and coreference, namely the dialogues from three pairs of subjects, or 22 *Find* subtasks. Table 6 shows the κ scores for the main categories we annotated. All values are in the range $[0.70, 0.79]$. These values are taken to exhibit a substantial agreement level [80], and given the complexity of the tasks, they are in fact quite reasonable.

4. Haptic-Ostensive Actions in Multimodal Dialogue Processing

We will now illustrate the roles that H-O actions, and extra-linguistic information in general, play in processing multimodal dialogues of the kind we collected. We first note that we subscribe to an information state approach to dialogue processing [81, 82, 83, 84]. In this approach, the *information state* is an abstraction of the dialogue context and the current state and intentions of the speakers. Crucially, such models include [85, p. 839] *a set of update rules that update the information state as dialogue acts are interpreted and [...] rules to generate dialogue acts.*

Figure 7 shows the general architecture of our dialogue processor. The inputs (utterances, pointing gestures, H-O actions) are an abstraction of what a speech processing module, a vision module and a haptic recognition module would return: for the work described in this paper, they are the transcribed utterance and the annotated pointing gestures / H-O actions. Our information state (*search state*) is much simplified with respect to general models such as [82], and only reflects the state of the objects and locations that ELD and HEL are collaboratively searching for (recall that we focus on *Find* tasks). Heuristic update rules update such state on the basis of the results of the coreference resolution and dialogue act classifier (we do not have room to present them in this paper). The *Next* diamond is a placeholder for the module that would decide what to do next, and generate a spoken utterance and / or a pointing gesture / H-O action. This module is left for future work.

We will now focus on the two main modules of our dialogue processor, coreference resolution and dialogue act classification. Our goal is to specifically highlight the contributions that H-O actions and other extra-linguistic features play in these tasks.

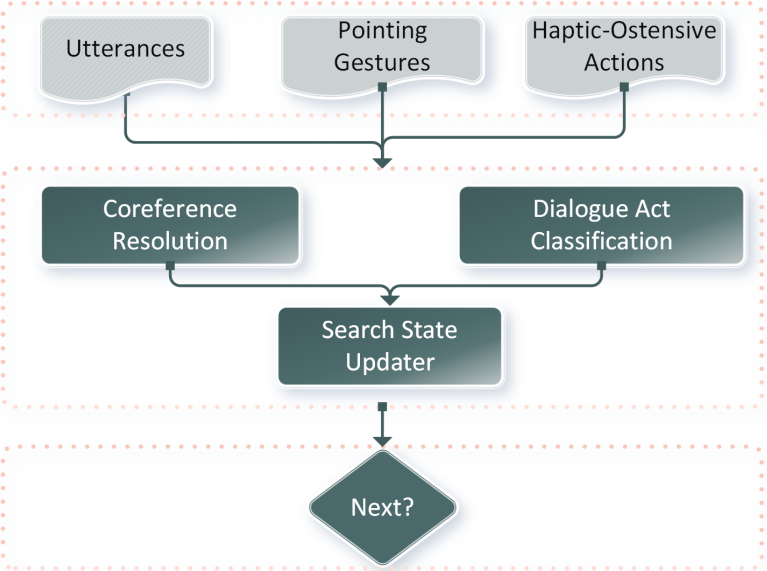


Figure 7. The Architecture of the Multimodal Dialogue Manager

4.1. Multimodal Coreference Resolution

Our multimodal coreference resolution system contains four modules: a referential classifier, a candidate pair generator, a candidate pair classifier, and a heuristic referent decider. Given a third person pronoun or deictic P,

1. the referential classifier decides whether P is referential;
2. the candidate pair generator generates potential coreference candidate pairs $< r_i, P >$ for P;
3. the candidate pair classifier labels each candidate pair $< r_i, P >$ as either *True* or *False*; namely, if $< r_i, P >$ is classified as true, r_i is a true referent for P;
4. if module 3 results in more than one coreference pair classified as *True*, the referent decider chooses one according to heuristic criteria.

Modules 1 and 3 were developed via supervised learning algorithms. Since our main interest was the role played by multimodal features, and not, uncovering why a specific supervised method performs better than others, we experimented with supervised learning algorithms that have been successfully deployed in similar language applications, specifically Maximum Entropy (MaxEnt), SVM and Decision Trees. In all cases, MaxEnt performs the best, according to more than one metric, and in a variety of settings, as we will detail below.

In the following we will discuss Modules 2-4. As concerns the first module, it decides whether P is referential or not. Recall that *it, there* can be used pleonastically: e.g. in *Yes, there it is* (Utt_{11} in Figure 3), *there* is not referential, but *it* is referential. The MaxEnt referential classifier achieves F-score=0.954 and will not be discussed further. This classifier uses lexical information about P (the word itself and its POS tag) and its surrounding context (word and POS tag bigrams and trigrams).

4.1.1. The coreference candidate pair generator

This module generates potential coreference candidate pairs $< r_i, P >$ for P. To measure the impact of pointing gestures and H-O actions for coreference resolution, we experimented with three different types of coreference candidate pair generators.

- *Textual only:* For each P, we create pairs by pairing P with all the compatible textual markables within six spoken turns (current turn plus five previous turns, skipping moves consisting only of pointing gestures or H-O actions – please refer to Appendix B for a description of how we apportion moves). [16] For example for *P=it* in Utt_{11} in Figure 3, all compatible textual markables from the current utterance and back to Utt_4 are included (moves M_{10} and M_6 consist of only H-O actions). This results in three candidate pairs: [<*A spoon, it*> (from Utt_7), < *the spoon, it* > (from Utt_8), <*the second drawer, it*> (from Utt_9)].

- *Pointing Gesture/H-O actions (extra-linguistic):* For each *P*, we create pairs by pairing it with the targets of all compatible pointing gestures, or with the targets of all compatible H-O actions, or with the targets of both. The compatible antecedents are the targets of pointing gestures or H-O actions that end within 4 seconds of the start of the current utterance. [17] E.g. for H-O only, for *P=it* in Utt_{11}, this results in: [$< spoon1, it >$ (from M_8), $< Drawer1, it >$ (from M_{10}), $< Drawer2, it >$ (from M_{10})]. Note that these referents come from the actions in moves M_{10} and M_8, not from any spoken words. That's why Drawer1 is included, even if it is never spoken of.
- *Combined:* For each *P*, we create pairs by pairing it with both textual and pointing gesture/H-O markables. Namely, the pairs generated under *Combined* are the union of the pairs generated by *Textual only* and *Pointing Gesture/H-O Action*.

4.1.2. The Candidate Pair classifier

We implemented the candidate pair classification model with the following algorithms: MaxEnt, Decision Tree, and Support Vector Machine (SVM). MaxEnt models were built using the MaxEnt[18] package from the Apache OpenNLP package. Decision Tree and SVM were run via J48 from Weka [86], and LibSVM [87], respectively. All of the results reported below were calculated using ten-fold cross validation.

Recall that our datapoints are pairs $< r_i, P >$, where r_i is the markable that is a potential antecedent for the pronoun/deictic *P*; the classifier will perform a binary classification, namely, whether $< r_i, P >$ is a true pair, i.e., whether r_i is a true antecedent for *P*. In the following, *U* represents the utterance containing *P*, e.g. Utt_{11} in our running example; whereas *ANT* represents either the utterance that contains r_i, or the pointing/H-O action of which r_i is the target.

Table 7 summarizes the features that the Candidate Pair classifier uses. These features depend on the type of r_i. Some features are common to both textual and extra-linguistic antecedents (leftmost column in Table 7), while others depend on the type. Hence, if r_i is textual, $< r_i, P >$ will have all the features listed in the left and center columns in Table 7; if it is the target of a pointing gestures, the pair will have the features in the leftmost column, plus *Object Agreement* in the rightmost column; finally, if it is the target of a H-O action, the pair will have all the features in the leftmost column, plus both features in the rightmost column. Full description of these features is contained in Appendix A.

Independent from r_i type	r_i type = Textual	r_i type = Extra-linguistic
Time distance	Lexical	Object agreement
Markable type agreement	Markable Distance	H-O type
Actor agreement	Turn distance	
Number agreement	Sentence distance	

Table 7. Features used by the Candidate Pair classifier, according to type of antecedent r_i

4.1.3. The heuristic referent decider

It is the last step in the coreference resolution process for *P*, and it was derived based on observations of the data and on experimentation. Its steps are:

1. Given a pronoun/deictic *P*, if the coreference pair classifier returns only one positive coreference pair for *P* $< r_i, P >$, return the markable r_i in that pair as the referent.

[16] The history length of 5 turns was determined experimentally.

[17] Again the value 4 seconds was determined experimentally.

[18] http://maxent.sourceforge.net

Pair Generator	Algorithm	Precision	Recall	F-Score	Accuracy	Resolved (out of 757)
Textual only	J48	0.0	0.0	0.0	0.951	0
	SVM	0.0	0.0	0.0	0.951	0
	MaxEnt	0.411	0.279	**0.332**	0.946	**35**
Pointing only	J48	0.854	0.896	**0.875**	0.910	**222**
	SVM	0.835	0.815	0.825	0.879	200
	MaxEnt	0.837	0.888	0.862	0.900	220
H-O only	J48	0.751	0.643	**0.693**	0.847	**148**
	SVM	0.662	0.374	0.478	0.780	82
	MaxEnt	0.719	0.0.6	0.654	0.829	135
Pointing + H-O	J48	0.772	0.722	0.747	0.850	329
	SVM	0.745	0.689	0.716	0.833	299
	MaxEnt	0.798	0.785	**0.792**	0.873	**359**
Combined	J48	0.610	0.409	0.489	0.882	236
	SVM	0.742	0.520	0.612	0.908	289
	MaxEnt	0.732	0.670	**0.700**	0.920	**390**

Table 8. Coreference Model Results

2. If there are multiple positive coreference pairs for *P*, we divide those positive pairs into two subsets, SAME and OTHER. SAME comprises those coreference pairs where the speaker of *P* is the SAME as the speaker of ANT, or as the actor of the pointing gesture / H-O action of which r_i is the target. OTHER contains the pairs associated with the other speaker. If SAME is not empty, choose SAME, otherwise OTHER.
3. If the chosen set contains pointing gesture/H-O action pairs, we choose the target of the *closest* pointing gesture/H-O action pair. Otherwise, we choose the markable of the *closest* textual pair.

4.1.4. Experimental Results

Table 8 shows the results for experiments with different combinations of algorithms and coreference candidate generators. In the coreference model experiments, we only generated candidate pairs for the 757 third person pronouns and deictics which were annotated as resolvable (please see Table 4).

We used several metrics to measure the performance of our reference resolution process. The standard measures of precision, recall, F-Score and accuracy are computed with respect to their performance on the generated coreference pairs – namely, these measures are computed on the output of Module 3, the candidate pair classifier, which for a single pronoun *P*, can return more than one candidate pair as *True*. The *Resolved* column instead is computed with respect to the output of Module 4, after the heuristic decider has chosen one single referent among all the possible ones that the candidate pair classifier has returned. Note that the *True/False* class distribution is very unbalanced, since the vast majority of candidate pairs are labelled as *False* in the training set. Hence, in Table 8, whereas accuracy has very high values, it is F-score and *Resolved* that provide a more meaningful measure of performance. The reader should also keep in mind that the numbers of candidate pairs to classify changes across the five pair generators; hence the only measure that is computed with respect to the same number of datapoints is *Resolved*, since the number of deictics and pronouns *P* is constant (757). As a consequence, it is not meaningful to compare precision, recall, F-score and accuracy across generators.

In terms of F-Measure, MaxEnt performs significantly better ($p \leq .005$, using χ^2) than the other two algorithms (Decision Tree and Support Vector Machine), for most coreference candidate pair generators. The MaxEnt model resolves more pronouns/deictics, for the *Textual only; Pointing + H-O; Combined* generators.

As mentioned, our main question is to ascertain whether multimodal information, including H-O actions, is useful to resolve referential expressions. Consider the row *Textual only* and the next group of three rows, pertaining to extra-linguistic information. First, by just glancing through the results, it is clear that candidate pairs generated exclusively via extra-linguistic antecedents are conducive to much better results than candidate pairs generated exclusively via textual antecedents. When we consider extra-linguistic antecedents only, across all algorithms more pronouns are resolved when only *Pointing* antecedents are used as opposed to only *H-O* antecedents. However, it is when we pair pronouns/deictics with markables which are targets of pointing gestures or H-O actions, that the performance of all algorithms improve; MaxEnt achieves the best result, by correctly resolving almost half of the referential pronouns and deictics (47.4%, 359/757). This shows the importance of using pointing gestures and H-O actions in multimodal, task-oriented corpora like the *Find* corpus.

In contrast, when we only generate textual candidate pairs, J48 and SVM simply classify every pair to “False”, which results in very high accuracy (0.951). As mentioned, the unbalanced training set is responsible for this high accuracy, since 95.1% of the coreference pairs are negative examples. Since all of the pairs were classified as “False”, there are no true-positive classifications, which in turn results in 0.0% precision and 0.0% recall. Whereas MaxEnt does a little better, and successfully resolves 32 pronouns/deictics, still, this is barely 4% of the total pronouns.

When we *Combine* the three types of antecedents together, clearly, more candidate pairs will be generated with respect to *Textual only* or *Pointing/H-O*. MaxEnt obtains the best result across all settings, for a total of 390/757 resolved pronouns (51%). Since the *Combined* model gives us the best results, it is the model used by the *Coreference Resolution* component within our dialogue manager (see Figure 7).

In conclusion, we have shown that extra-linguistic information by itself can correctly solve almost half of the pronouns; that textual information by itself is not helpful, but that when it is combined with non-textual information, it results in a small improvement, of about 3.7%. As concerns the importance of including H-O actions, we have shown that by including markables which are targets of H-O actions in addition to those which are targets of pointing gestures, we increase the number of **Resolved** pronouns by 50% (359 as opposed to 222).

4.2. Multimodal Dialogue Act classification with Haptic-Ostensive Actions

Recall that in our framework, the DA classification problem means, to assign one of the thirteen DA labels we discussed in Section 3.2.1 to the current spoken utterance U_i. We also use the notion of *previous move* M_{i-1}, which can be another utterance, a pointing gesture, or a H-O action. At the same time, the current utterance U_i may be accompanied by a pointing gesture or by a H-O action. Appendix B describes the algorithm we developed to distinguish between pointing gestures / H-O actions that constitute independent moves, or are associated with an utterance.

As for reference resolution, all the DA classification models we present are machine learning models. We will again show that H-O actions, and also the structure of the dialogue, play a critical role. All of the experiments shared the same features, which include features of the individual utterance U_i (both linguistic and extra-linguistic), and features of the larger context: the dialogue history (what transpired before U_i) and the dialogue game to which U_i belongs (a correlate of the hierarchical structure of dialogue). Table 9 summarizes the type of features we are considering. Both lexico-syntactic and meta features still comprise more elementary features, to be described in Appendix C.

Utterance		Dialogue Structure	
Linguistic	Extra-Linguistic	Dialogue History	Dialogue Game
Lexico-syntactic	Pointing gesture H-O action Location Meta	Type of previous move M_{i-1} M_{i-1} actor Same actor as M_{i-1}	Dialogue Game

Table 9. Different kinds of features used by the DA classifier

Classes of features commonly used in the literature are linguistic features of the utterance, and dialogue history. Extra-linguistic features (other than standard meta features such as actor of the utterance or time length) and dialogue game have not been used as extensively; H-O actions have never been used, as far as we know. Appendix C includes the precise definition for all features, including subfeatures comprising utterance and meta features; here we only briefly discuss the computation of the *Dialogue game feature (DG)*, that models the hierarchical dialogue structure. Some previous research on DA classification has shown that hierarchical dialogue structure encoded via the notion of conversational games [88] significantly improves DA classification [32, 38, 89].

To extract the DG feature without explicit annotation, we rest on an approximation to the definition of Dialogue Game from MapTask. In MapTask, a dialogue game was annotated as a sequence of utterances that starts with an

Features	**MaxEnt**	**NB**	**J48**
LX-SYN (Linguistic)	.641	.453	.450
PT (Pointing)	.228	.212	.212
H-O (Haptic-Ostensive)	.272	.243	.212
LO (Location)	.281	.259	.265
META (Meta)	.342	.417	.392
DG (Dialogue Game)	.189	.212	.212
DH (Dialogue History)	.291	.284	.294
LX-SYN+PT	.646	.453	.450
LX-SYN+PT+H-O	.644	.456	.449
LX-SYN+PT+H-O+LO	.666*	.523*	.536*
LX-SYN+PT+H-O+LO+META	.680	.563	.568
LX-SYN+PT+H-O+LO+META+DG	.710**	.591**	.607**
LX-SYN+PT+H-O+LO+META+DG+DH	**.746†**	**.606**	**.627**

Table 10. Dialogue Act Classification Accuracy

initiating move U_1 and encompassed all utterances $U_2...U_n$ up until the purpose of the game has been fulfilled, or abandoned. In MapTask, initiating moves are *instruct, explain, check, align, query-yn, query-w*; the Dialogue Game label is equated with the DA label of U_1, for all utterances $U_1, U_2...U_n$ belonging to the specific Dialogue Game.

To compute the DG label, we use a just-in-time approach. For each Find subtask dialogue, we maintain a stack for dialogue games. After an utterance is classified by our classifier as an initiating move, we assume the dialogue has entered a new dialogue game, and push its DA label as the dialogue game to the top of the stack. Hence, when the next utterance U_i is being classified for DA, the DG feature value used is the closest preceding initiating DA. The DG label will not change until a new initiating move is found.

4.2.1. Experimental results

We employed supervised learning approaches, specifically: Maximum Entropy (MaxEnt), Naive Bayes (NB), and J48 (Decision Tree). These algorithms are widely used for DA classification. Naive Bayes often provides a more telling baseline than the simple majority class [38, 40, 90, 91]. We used the same packages we had used for coreference resolution, MaxEnt from OpenNLP, and J48 from Weka; also the package for Naive Bayes came from Weka. SVM which we used for coreference resolution, is inherently a binary rather than a multiclass classifier, and hence, it is not frequently used for DA classification.

To evaluate feature effectiveness, we evaluated 512 feature combinations, namely, all feature combinations for the seven feature groups from Table 9: lexico-syntactic (LX-SYN), pointing gesture (PT), H-O action (H-O), location (LO), meta (META), dialogue game (DG), dialogue history (DH).[19] It would be impossible to report all results for 512 experiments. Similar to [40], we chose to report the results in Table 10, by showing the contributions of single feature groups, and then incremental feature combinations by adding features to the most effective individual feature group, LX-SYN. Please note that the omitted 499 combinations did not shed any additional light on the problem.

In Table 10, bold font indicates the feature set giving best performance for each column. The diacritics ***, **** and † indicate significant improvements (always assessed with χ^2). Specifically, ***** indicates significant improvement with respect to the results obtained with *LX-SYN* only; ****** indicates the next significant improvement; † indicates the next significant improvement, which is obtained with MaxEnt after adding all feature groups together. The majority baseline, which always assigns the most frequent tag *INSTRUCT* to every utterance, had an accuracy of 20.7% (see Figure 6).

Overall, when all the features came into play, MaxEnt performed the best, and it significantly outperformed NB and J48 ($p<.005$). Concerning features, contrary to what was observed for referring expressions, linguistic features

[19] We preferred to exhaustively try out all combinations, at least at the feature group level, rather than performing feature selection.

(LX-SYN) were the most predictive as a feature type used by itself; however, when pointing gesture (PT), H-O features (H-O) and location features (LO) were added together to linguistic features, we noticed a significant performance improvement for all the models. None of the pointing gesture, H-O action and location features alone significantly improved the results, but all three together did. This confirmed the finding of [8, 40] that extra-linguistic features help DA classification. Since, as mentioned, we ran all possible feature combinations, we carefully examined the combinations where exactly one of PT, H-O and Location feature group is ablated from the full experiment feature set (LX-SYN+PT+H-O+LO+META+DG+DH). No significant difference was discovered.

As concerns the contribution of dialogue structure features, when the dialogue game feature (DG) was added to the models, performance increased significantly for all the models: MaxEnt (p<.025), NB (p<.05), and J48 (p<.005). This confirmed previous findings, including those by our group [32], that dialogue game features (DG) play a very important role in DA classification, even via the simple approximation we used. At last, dialogue history (DH) features further improved the performance for the MaxEnt model (p<.005), and resulted in the best performance.

4.3. Summary of reference resolution / dialogue act classification experiments

The main hypothesis we evaluated via developing models of reference resolution and of dialogue act recognition was whether extra-linguistic features, and specifically H-O actions, play a significant role in these processes. Our practical goal is of course to eventually deploy the best models we found in an assistive robot.

First, we have shown that in both cases extra-linguistic features provide crucial information that greatly improves the models. For reference resolution, by including the targets of pointing gestures and H-O actions among the potential referents, and using features of these targets, almost half of pronouns are correctly resolved, as opposed to barely 4% with only textual antecedents. For dialogue act classification, adding extra-linguistic features (PT+H-O+LOC) provides significant improvement as well.

Whereas we are not the first to show that extra-linguistic features help, as far as we know we are the first to include H-O actions, as arising in a real-world domain, in these models. As specifically concerns H-O actions, our results show that they are effective, either on their own or in concert with other extra-linguistic features. For co-reference resolution, even by themselves they help resolution: 20% of pronouns/ deictic are correctly resolved when using only targets of H-O actions as antecedents, which is double the rate with only textual antecedents, even if it is lower than when using only targets of pointing gestures. As already mentioned, models that include both pointing gestures and H-O actions solve almost 50% of pronouns. For dialogue act classification, whereas we do not have evidence to conclude whether one extra-linguistic feature is more important than the others, it is only when all three are together that the best result is obtained.

All the experiments we have presented so far do include extra-linguistic features, but as we made clear, they were run on annotated features – namely, neither pointing gestures had been automatically recognized via a vision module, nor had H-O actions been automatically recognized from the haptic signals. It is clear that, for recognition to truly occur in a physical domain, true vision and haptic recognition are necessary. Whereas within RoboHelper vision efforts focused more on recognition of faces and objects[20] [92, 93] and are not relevant to the current discussion, we now turn to the experiments we performed to start tackling the issue of recognizing H-O actions from the haptic signals. As we noted earlier, the experiments on H-O action recognition we will discuss serve as proof-of-concept that grounding multimodal dialogue processing in true signals is feasible.

5. Recognizing Haptic-Ostensive actions

In the previous sections, we discussed analyses and models that highlight the role played by H-O actions. The input to those models are H-O actions *annotated* by humans on the basis of the video stream and the haptic data. Clearly, in order to use our models in HRI, we need to recognize H-O actions from the haptic signals. But *grounding* the high-level, Natural Language component of the RoboHelper system we envision, in the physical world, is not the only reason why recognizing H-O actions from haptic signals is necessary. To start with, among various annotations, H-O actions require the most time. It is obvious that the task of annotating the corpus would be greatly simplified

[20] But not the objects in our videos.

if H-O actions can be automatically recognized in the haptic data stream. Being able to recognize H-O actions from haptic data would also directly enable learning by demonstration, a commonly used approach in HRI [9, 10, 11].

Two clarifications before we proceed. First, in our data collection, 97.5% of the H-O actions are performed by the HELper, as shown in Table 5. This is a consequence of the specific type of task we focused on in the study presented in this paper. In a realistic HRI scenario, the HELper would be the robot, and it is the H-O actions of the ELDer that would need to be recognized. We claim that this does not diminish the value of our work. The methodology that is presented in this paper is valid regardless where the recognition system resides. However, in any long-term deployment of a robot, the number of H-O actions performed by the ELDer in these tasks would be significant, especially because a robot should not just provide convenience for the ELDer but actually stretch their capabilities [94]. Also, while working in the kitchen, there are a number of other activities such as a handover task or a collaborative manipulation (think of a pot of water collaboratively moved to the stove) where the H-O actions are equally performed by the ELDer. We have started to study these actions that we call *haptic interaction*, but we have no room to discuss them in this paper.

Second, in an actual deployment of RoboHelper, it would clearly not be practical for a human to wear the sensory glove to transmit haptic data to the robot. Some features of H-O actions could be detected using vision, although clearly occlusions are challenging for vision; alternatively, embedded sensors in instrumented objects can be used, as we have done in [95]; neither of these approaches though, can distinguish actually touching an object from being close to it. However, as already mentioned earlier, a sensory glove is one of a few options to measure haptic data when collecting a multimodal corpus. Second, a glove could clearly be used in collecting data for learning by demonstration. That is why a methodology for automatically recognizing H-O actions from the glove data is directly useful for HRI.

Now turning to the haptic recognition experiments, our methodology to automatically recognize H-O actions would preferably be demonstrated directly on the haptic data collected via the glove in our *Find* tasks. Unfortunately, this was not possible because the collected haptic data turned out to be corrupt in some cases. A posteriori, we identified several problems. In some cases, the communication was lost between the data collection module and the computer used to store the data (the speech and video data was not affected). In other cases, the pressure sensor data was corrupted due to an outer glove that was worn over the fabricated glove to help with the computer vision algorithms (see the yellow gloves on the helper's hands in Figure 1). We thus had to collect additional data to develop the automatic recognition algorithms for H-O actions. We collected data in the same mock apartment at Rush University where we had collected the ELDERLY- AT-HOME corpus. Machine learning techniques were applied to the data to demonstrate that H-O actions can be successfully recognized.

The additional data we collected and the models we developed based on it, are informed by the following observations, related to the five H-O actions we had annotated for (*Touch, Grasp-Show, Grasp-No-Show, Open, Close*).

- It would be very easy to recognize *Touch* if HEL or ELD were not touching anything unless they are physically interacting with an object; in this case, one would expect a clear increase in the pressure data on mostly the fingertip of the index finger and possibly, on the fingertips of middle and ring fingers. However, we observed that the participants' hands were usually resting on a part of their body or some surface, when they were not performing an H-O action. Furthermore, no matter the glove, even when the hand is idle the sensor readings may be nonzero if the glove is folded. Hence, *Touch* per se is hard to recognize in the present scenario.
- As concerns H-O action recognition, *Grasp-Show* and *Grasp-No-Show* are the same: the pressure data does not vary depending upon whether one subject is explicitly showing the grasped object to the other subject or not.
- The annotation scheme does not take into account one important dimension of variation of the pressure signals, namely, the object that is grasped, or open, or closed.

Hence, whereas we will not recognize *Touch* per se, given the difficulties just discussed, we will distinguish a separate *Idle-Hand* class; we will collapse *Grasp-Show* and *Grasp-No-Show* to a single class *Grasp*; and, for the purposes of the experiments we will discuss here, we will not maintain the distinction between *Open* and *Close* (in additional experiments, we found that *Open* and *Close* can be distinguished on the basis of the measurements we obtained via a six degree-of-freedom inertial measurement unit attached to the back of the data glove [96]).

Apparently, we are left with three classes of H-O actions to be recognized. However, the task is more challenging than distinguishing between three different action types, since these actions differ according to the objects they are applied to. In our experiments, this will result in seven different classes of H-O actions (please see Tables 11 through 15).

Note that this subdivision nicely dovetails with the coreference resolution component of the dialogue manager, since H-O actions have annotated targets, as we discussed in Section 3.2.2. Clearly, for a realistic HRI scenario, seven action types are still too few. The work we present here is just the beginning of further modeling of the many other haptic actions that the ELDERLY-AT-HOME corpus also includes. Some of those haptic actions pertain to the haptic interactions we discussed at the beginning of this section; other naturally arise in our data when the HELper has to directly touch the ELDer to help him/her ambulate, put on a sweater, or his/her shoes. It would not have been prudent, or realistic, to include those other actions in this initial analysis, since issues of safety arise that are beyond the scope of the work discussed in this paper.

5.1. NatVal and ForcedRep Datasets

Since the haptic data collected during the *Find* tasks turned out to be corrupt in some cases, we performed an additional data collection, driven by two goals: (1) to collect naturalistic data as similar as possible to the ELDERLY-AT-HOME corpus (same environment, same tasks) with coherent haptic signals, on which to evaluate the recognition algorithms; and (2) to collect additional haptic data that would involve multiple repetitions of the same action in a naturalistic setting. The problem with completely naturalistic data is that in practice the same haptic action is never repeated: even when subjects open or close the same cabinet, their body position with respect to the cabinet subtly varies from one instance to the next, and affects the haptic signals. Hence, haptic actions in the naturalistic setting are rather sparse, and such variability poses a challenge for automatic recognition. In particular, we need at least two instances of an action in order to perform recognition experiment so that one instance can be used for training, while the other is used as a test data (please note that each action is a time series of samples). After our subjects were done with the experiments mirroring the ELDERLY-AT-HOME corpus, we thus asked them to repeat certain actions. As a result, the experiments produced two additional sets of data: the naturalistic data mirroring the ELDERLY-AT-HOME corpus, that we call *Naturalistic Validation* (NatVal) set; and data with forced repetitions, which we call *ForcedRep* set.

For our experiments, four pairs of subjects wear the same equipment as in the ELDERLY-AT-HOME data collection, and perform the same ADLs. Our subjects were young adults (UIC students); one subject played the role of the helper, and one the role of the elderly person. Additionally, at the end of the naturalistic tasks, we asked three helpers to grasp kitchen items and open/close cabinets/drawers multiple times. The helper wore the data glove we had developed, mirroring the ELDERLY-AT-HOME data. Video of the experiments was recorded through the cameras we had installed in the mock apartment. This video was then synchronized with the glove data using the time stamps of the glove data. Since the haptic data was not uniformly sampled, we subsampled it at 50Hz after removing outliers. The data was subsequently filtered with a moving average filter using a 1 second window to remove noise.

Table 11 presents the frequency distribution of actions among the four HELper subjects (HEL1 through HEL4), in the *NatVal* set. Table 12 presents the frequency distribution of H-O actions in the *ForcedRep* set. Note that HEL1 is not included in Table 12 because this subject was not asked to repeat actions.

Table 11. Frequency distribution of actions in the *NatVal* set

	HEL 1	HEL 2	HEL 3	HEL 4	Total
Open/Close Cabinet	13	13	24	17	**67**
Open/Close Drawer	16	0	17	4	**37**
Open/Close Fridge	0	4	0	0	**4**
Grasp Plate	3	4	2	8	**17**
Grasp Pot	8	6	5	4	**23**
Grasp Small Items	3	6	4	4	**17**
Idle Hand	8	4	8	2	**22**
Total	**51**	**33**	**60**	**39**	**187**

From a haptic data point of view, a *Grasp* action and *Open/Close* actions are inherently different: the former is static while the latter are dynamic. In other words, while after contact with the object has been established, the different samples that are part of a *Grasp* action are very similar (the sensor readings don't change much), *Open/Close*

Table 12. Frequency distribution of actions in the *ForcedRep* set

	HEL 2	HEL 3	HEL 4	Total
Open/Close Cabinet	3	0	15	**18**
Open/Close Drawer	20	11	8	**39**
Open/Close Fridge	6	0	0	**6**
Grasp Plate	11	23	9	**43**
Grasp Pot	0	0	31	**31**
Grasp Small Items	8	14	0	**22**
Idle Hand	0	4	30	**34**
Total	**48**	**52**	**93**	**193**

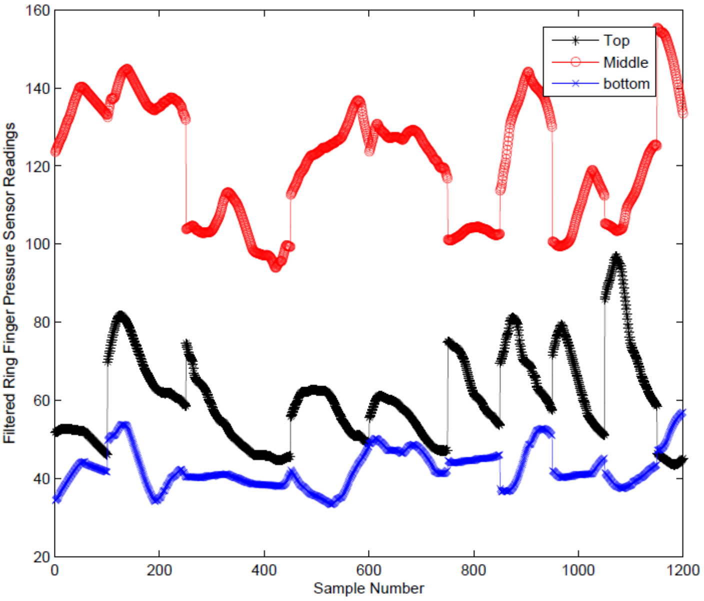


Figure 8. Top, Middle, Bottom Sensors on the Ring Finger: Readings During Opening of a Cabinet

are better seen as a time varying sequence of pressure sensor readings, as shown in Figure 8 (Top, Middle, Bottom in Figure 8 refer to the sensors on each of the three phalanxes of the ring finger, see the glove in Figure 2). Hence, we experimented with two supervised learning algorithms for time sequences, RISq [97] and Dynamic Time Warping (DTW) [98]. Since the results of the two algorithms were not statistically significantly different, we only report DTW results here.

5.2. H-O action recognition experiments

We performed three classification experiments, which are progressively more challenging: (1) using cross-validation on *ForcedRep* set; (2) using cross-validation on *NatVal* set; and (3) using the *ForcedRep* set to train a model, which was in turn used to classify the *NatVal* set. In all cases, the classification was performed within a subject. In other words, the training data and the test data in each experiment came from the same subject. Note that in the experiments, the *Grasp* actions are executed in different configurations, for example *Grasp Plate* can be executed with one hand, with one hand and another person, with both hands, with both hands and another person, or grasping three plates with both hands: for the purpose of this paper, these are treated as the same action.

1. Cross-validation on ForcedRep set. Table 13 shows the recognition results for the DTW algorithm when both training data and test data came from the *ForcedRep* set. We used between 50% and 90% of the data for training (depending on whether we had at least 10 instances of the action or not), and the rest as test data. Training and test data came from the same subject. The experiments were repeated 10 times or until each data combination could be used for testing, whichever was lower. While not shown in Table 13, opening and closing actions (Cabinet, Drawer and Fridge) are at times confused with each other but they are well separated from the other actions. In this case, accuracy is very high since 94.8% of the actions are classified into the right group.

Table 13. Confusion Matrix for Cross-Validation on the *ForcedRep* set

	Open/ Close Cabinet	**Open/ Close drawer**	**Open/ Close Fridge**	**Grasp Plate**	**Grasp Pot**	**Grasp Small Items**	**Idle Hand**	**Total**
Open/Close Cabinet	**18**	0	0	0	0	0	0	**18**
Open/Close Drawer	0	**38**	0	0	1	0	0	**39**
Open/Close Fridge	0	0	**6**	0	0	0	0	**6**
Grasp Plate	0	0	0	**42**	1	0	0	**43**
Grasp Pot	0	0	0	0	**31**	0	0	**31**
Grasp Small Items	5	0	0	0	3	**44**	0	**52**
Idle Hand	0	0	0	0	0	0	**4**	**4**
Total	**23**	**38**	**6**	**42**	**36**	**44**	**4**	**193**

2. *Cross-validation on NatVal set.* Table 14 shows the recognition frequencies for the DTW algorithm when both training data and test data came from the *NatVal* set. As above, we used between 50% and 90% of the data for training, and the rest as test data. The experiments were repeated 10 times or until each data combination could be used for testing, whichever was lower. As above, opening and closing of a cabinet were at times confused with each other but these actions are well separated from the other actions; the same is true for opening/closing of drawers. Overall 68.4% of the total of 187 H-O actions are correctly classified to the right group. In our earlier work, we also demonstrated successful classification of different H-O actions within one group (for example, for manipulation of planar objects, distinguishing grasping a plate alone or with another person) [99].

Table 14. Confusion Matrix for Cross-Validation on the *NatVal* set

	Open/ Close Cabinet	**Open/ Close Drawer**	**Open/ Close Fridge**	**Grasp Plate**	**Grasp Pot**	**Grasp Small Items**	**Idle Hand**	**Total**
Open/Close Cabinet	**53**	5	3	0	1	0	5	**67**
Open/Close Drawer	2	**28**	0	0	0	3	4	**37**
Open/Close Fridge	0	0	**4**	0	0	0	0	**4**
Grasp Plate	4	2	0	**6**	0	3	2	**17**
Grasp Pot	4	1	0	1	**17**	0	0	**23**
Grasp Small Items	5	2	0	2	2	**5**	1	**17**
Idle Hand	1	4	0	1	0	1	**15**	**22**
Total	**69**	**42**	**4**	**13**	**20**	**12**	**25**	**187**

3. *Training across sets: ForcedRep used to recognize NatVal.* In our final experiment we used the data from the *ForcedRep* set as training data and the data from the *NatVal* set as testing data. In a sense, this is the most logical experiment: we collected repeated instances of the actions of interest performed in the natural environment; we then use those actions to classify the actions occurring during unstructured interaction, which introduces much greater variability.

The recognition frequencies are shown in Table 15. Note that as before, the training data and the test data were always from the same subject. The number of actions is reduced to 102 from the original 183 in Table 11 because

ForcedRep data was not collected for HEL1; and, for HEL3, repeated *Open/Close Cabinet* data was not collected. In this case, 58.8% of the 102 H-O actions are classified as belonging to the right group. Not surprisingly, performance is worse than for cross-validation on the *ForcedRep* set, but not too much lower than the performance for cross-validation on the *NatVal* set.

Table 15. Confusion Matrix when training on *ForcedRep*, but testing on *NatVal*

	Open/ Close Cabinet	**Open/ Close drawer**	**Open/ Close Fridge**	**Grasp Plate**	**Grasp Pot**	**Grasp Small Items**	**Idle Hand**	**Total**
Open/Close Cabinet	**20**	4	0	5	1	0	0	**30**
Open/Close Drawer	0	**26**	0	1	0	3	0	**30**
Open/Close Fridge	0	1	**3**	0	0	0	0	**4**
Grasp Plate	0	5	0	**4**	3	5	0	**17**
Grasp Pot	0	0	0	0	**2**	2	0	**4**
Grasp Small Items	0	2	0	0	0	**5**	0	**7**
Idle Hand	0	7	0	0	1	2	**0**	**10**
Total	**20**	**45**	**3**	**10**	**7**	**17**	**0**	**102**

6. Conclusions and Future work

This paper described research to study the roles of multimodal events, including pointing gestures and Haptic-Ostensive actions, in collaborative task oriented multimodal human-human dialogues in the elderly care domain. We described the process of building the ELDERLY-AT-HOME corpus, including data collection and transcription, and annotations at many levels for its *Find* subcorpus. We ran reference resolution experiments and dialogue act classification experiments to investigate the roles of pointing gestures and H-O actions. The experiments show that H-O actions play a crucial role in interaction. To ground our corpus analysis and modeling in the physical world, we performed further data collection to study whether (some) H-O actions can be automatically recognized from the haptic data collected through a sensory data glove instrumented with pressure sensors. Machine learning experiments were conducted that showed that the H-O actions of interest can be recognized within subjects, even though pressure sensors are relatively imprecise and the data provided by the sensory glove is noisy.

The work has several direct implications for HRI. To start with, we demonstrated that pressure sensors can provide useful information about the haptic actions that are part of a multimodal interaction. Next, we showed that H-O actions, a subset of haptic actions, play a crucial role in reference resolution and dialogue act classification. In turn, the ELDERLY-AT-HOME corpus is one of the few multimodal corpora, if not the only one, that contains haptic data of the sort we describe: it will constitute a substantial contribution to the community when we release it. Finally, the algorithms we developed for automatic recognition of H-O actions from the haptic data can be directly used to expand the ELDERLY-AT-HOME corpus as new data is collected, and can provide the input for learning by demonstration in the assistive robotics domain.

However, several issues need to be resolved to make the work fully usable from an HRI perspective. From a dialogue management point of view, our current information state is very simple, and only reflects the state of the objects and locations that ELDer and HELper are collaboratively searching for. Additionally, we mentioned in Section 4 that *Next* in Figure 7 is a placeholder for the module that would decide what to do next, and generate a spoken utterance and / or a pointing gesture / H-O action. Some mileage can be obtained by addressing *what to do next* at a purely symbolic level, and by applying machine learning to this task too, on the sequences of turns that occur in our corpus. However, this module should clearly be informed by feedback from the robot as well, since tasks that are considered primitive at the dialogue level result in further decompositions – for example *Can you get me a pot* results in HEL moving towards the cabinets before opening any of them.

A crucial future step concerning H-O action recognition from the haptic signals, is to study recognition across subjects, when training data and test data do not come from the same individual. Additional data has been collected in a laboratory setting for this purpose. Further, as we discussed in Section 5, several other haptic actions exist in our data, that we have only started to explore. Some of them pertain to *haptic interaction*, when two agents collaboratively manipulate objects; some of them pertain to the HELper directly touching the ELDer, as when helping the ELDer with ambulating or with putting on clothing.

One intriguing issue is the fact that *Grasp-Show* and *Grasp-No-Show* cannot be distinguished from a haptic point of view, and given the current state of the art of vision algorithms, it is doubtful they may be distinguished visually. Further analysis should be conducted to assess whether in the annotated data, this distinction affects coreference. For example, is it the case that when a referential expression has the target of a *Grasp* action as referent, it is in fact *Grasp-Show*, rather than *Grasp-No-Show*? If that is the case, there may be cues in the surrounding context, for example the type of the previous move, and in the information state of the dialogue, that could allow us to recognize the true ostensive intention behind such actions.

Finally, and equally important is to test the developed methodology on a robotic platform. A preliminary implementation has been completed in ROS [100], including a real-time implementation of the H-O action recognition algorithms [101]; experiments with a Nao robot [102] are under development. An important question that will be addressed is whether H-O actions can inform coreference resolution and dialogue act classification even when they are recognized automatically and thus with lower accuracy.

Acknowledgements

This work was supported by award IIS-0905593 from the National Science Foundation (NSF). We gratefully acknowledge the contributions of the other members of RoboHelper, and specifically Jezekiel Ben-Arie, Mark Foreman, Simone Franzini, Shankaranand Jagadeesan, and Kai Ma. Additional funding was provided by NSF awards CNS-0910988 and CNS-1035914.

Appendix A. Coreference Pair classifier features

Coreference pairs with textual or extra-linguistic antecedents share a common set of features (the feature domain is in parenthesis) - ANT is the utterance where the textual markable r_i occurs, or the pointing gesture/H-O action of which r_i is the target.

- *Time distance (discrete, in seconds)*: The distance between the time spans of *U* and of *ANT*. If the two spans overlap, the distance is 0.
- *Actor agreement (boolean):* If the speaker of *U* and the speaker or actor of ANT are the same.
- *Markable type agreement (boolean):* If the markable type (OBJECT, PLACE, or NEITHER) of *P* is compatible with the type of r_i.
- *Number agreement (boolean)*: If the number of *P* is the same as that of r_i.

For coreference pairs with textual antecedents, we use the following additional features. ANT here is only the utterance where the textual markable r_i occurs.

- *Lexical (vector)*: The words and part-of-speech tags in r_i and in *P*, and whether they share (some of) the same words.
- *Utterance turn distance (integer)*: The turn distance between *U* and *ANT*; if *U* and *ANT* are in the same turn, the distance is 0
- *Sentence distance (integer)*: if *U* and *ANT* are in the same turn, the distance between the two sentences where r_i and *P* respectively appear
- *Markable distance (integer)*: how many markables occur between r_i and *P*.

For coreference pairs with extra-linguistic antecedents, we used the following additional features. Recall that in this case, *ANT* is only the pointing gesture/H-O action of which r_i is the target.

- *Object agreement (boolean)*: If the deictic P is contained in a phrase, such as “this big blue bowl”, whether the head noun “bowl” matches the object type of r_i.

- *H-O Action type (discrete)*: If r_i is target of an H-O action (not of a pointing gesture), the type of the H-O action (*Touch, Grasp-Show, Grasp-No-Show, Open, Close*).

Appendix B. Algorithm to compute moves

As noted in Section 4.2, a pointing gesture or H-O action may be independent moves or may accompany a spoken utterance U_i. We developed an algorithm that distinguishes between the two cases. In turn, the algorithm makes use of additional annotations on pointing gestures / H-O actions that we will briefly describe here. Note that just using timespans is not sufficient. It is not necessarily the case that utterance U is associated with gesture / H-O action G if their timespans overlap.

First, we assign to each utterance, pointing gesture and H-O action a unique event index, so that we can refer to these events with their indices. For pointing gestures and H-O actions, we define two more attributes: “associates” and “follows”. If a pointing gesture or H-O action is associated with an utterance, the “associates” value will be the index of that utterance; by default, the “associates” value is empty. If a pointing gesture or H-O action independently follows an utterance, the “follows” value will be that utterance’s index. E.g., H-O action 6 (“Open”) in Figure 3 is marked with “follows [5]”. For utterances, we only mark the “follows” attribute. If an utterance directly follows a pointing gesture or H-O action, we use the index of the pointing gesture or H-O action as the “follows” value. By default, the “follows” attribute of an utterance is empty. It means that an utterance follows the immediate previous utterance. The annotation of “associates” and “follows” was done by our coders by looking at the video, and based on their human judgement.

We define a *move* as any combination of related utterances, pointing gestures and H-O actions, performed by the same subject. On the basis of the event relation annotations, we can compute the dialogue’s move flow using the following algorithm.

1. Order all the utterances in a *Find* task session by the utterance start time
2. Until all the utterances are processed, for each unprocessed utterance u_i:
   (a) If u_i *follows* a pointing gesture or H-O action, that pointing gesture or H-O action forms a new *move* m_k; add m_k to the sequence before u_i
   (b) Find all the pointing gestures and H-O actions labelled as *associates* of u_i. These events form the *move* m_i together with u_i
   (c) Recursively find the events which follow the last generated *move*, together with all their associated events to form another *move*

The third and recursive step is necessary to properly apportion to moves, gestures that follow one another, without being associated with an utterance. For example consider Utt_i with a pointing gesture P *associated* with it, and immediately *followed* by an H-O action O. O would be recognized as a separate move by step 3.

Appendix C. Dialogue Act Classification features

Utterance features. Lexico-Syntactic (LX-SYN) are the most widely used features for DA classification [32, 36, 38, 39, 40, 103, 104, 90], and include:

- *Number of sentences (integer)* and *number of words (integer)* in the utterance. The Apache OpenNLP library [21] was used to detect sentences and tokenize them.

- *Lexical (vector)*: Unigrams of the words and part-of-speech tags in U_i. The words were processed using the morphology tool from the Stanford parser [105].

[21] http://opennlp.apache.org/

- *Lexical choice (two boolean features)*: Whether U_i contains WH words (e.g., *what, where*); whether U_i contains yes/no words (e.g., *yes, no, yeah, nope*).

- *Syntactic (vector)*: The top node and its first two child nodes from the sentence parse tree. If U_i contains multiple sentences, only the last sentence is used. Sentences are parsed using the Stanford parser.

Extra-linguistic features include:

Pointing gesture feature (PT) (boolean): whether the actor of U_i is making a pointing gesture.

Haptic-Ostensive feature (H-O) (discrete): the type of the H-O action performed by the speaker of U_i if any; null otherwise.

Location features (LO) (vector) includes the locations of the two actors, whether they were in the same location, and whether the actor of U_i changed the location during the utterance.

Meta features (META) are extracted from U_i's meta information. Previous research had shown that the utterance meta information such as the utterance speaker helped classify DAs [104, 91].

- The *actor* who spoke U_i (discrete, HEL or ELD)

- The *time length* of U_i (seconds)

- The *distance* of U_i from the beginning of the dialogue (integer)

Dialogue structure features. Dialogue history features (DH) model what happened before U_i [38, 32]. We encoded:

- The *actor* of the previous move M_{i-1} (discrete, HEL or ELD) – what precedes U_i can be a spoken utterance, but also a pointing gesture or a H-O action;

- Whether M_{i-1} has the *same actor* as U_i (boolean)

- The *type* of M_{i-1} (discrete): if it is an utterance, its DA label; if it is a H-O action, the type of H-O action; if it is a pointing gesture, a label that indicates this.

The computation of the *Dialogue game feature (DG)* was described in Section 4.2.

References

[1] A. Jaimes, N. Sebe, Multimodal human-computer interaction: A survey, Computer Vision and Image Understanding 108 (1-2) (2007) 116–134.

[2] A. Kehler, Cognitive Status and Form of Reference in Multimodal Human-Computer Interaction, in: AAAI 00, The 15th Annual Conference of the American Association for Artificial Intelligence, 2000, pp. 685–689.

[3] S. Goldin-Meadow, Hearing gesture: How our hands help us think, Harvard University Press, 2005.

[4] C. Navarretta, Anaphora and gestures in multimodal communication, in: Proceedings of the 8th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2011), Faro, Portugal, Edicoes Colibri, 2011, pp. 171–181.

[5] P. Cohen, D. McGee, Tangible multimodal interfaces for safety-critical applications, Communications of the ACM 47 (1) (2004) 41–46.

[6] M. Foster, E. Bard, M. Guhe, R. Hill, J. Oberlander, A. Knoll, The roles of haptic-ostensive referring expressions in cooperative, task-based human-robot dialogue, in: Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction, ACM, 2008, pp. 295–302.

[7] K. M. Krapp, The Gale Encyclopedia of Nursing & Allied Health, Gale Group, Inc. Chapter Activities of Daily Living Evaluation., 2002.

[8] J. Orkin, D. Roy, Semi-automated dialogue act classification for situated social agents in games, in: Agents for Games and Simulations II, Springer, 2011, pp. 148–162.

[9] P. Bakker, Y. Kuniyoshi, Robot see, robot do: An overview of robot imitation, in: AISB96 Workshop on Learning in Robots and Animals, 1996, pp. 3–11.

[10] A. Billard, S. Calinon, R. Dillmann, S. Schaal, Robot Programming by Demonstration, in: B. S. Prof, O. K. Prof (Eds.), Springer Handbook of Robotics, Springer Berlin Heidelberg, 2008, pp. 1371–1394.

[11] B. D. Argall, S. Chernova, M. Veloso, B. Browning, A survey of robot learning from demonstration, Robotics and Autonomous Systems 57 (5) (2009) 469–483.

[12] G. Tur, R. De Mori, Spoken language understanding: Systems for extracting semantic information from speech, John Wiley & Sons, 2011.

[13] R. A. Bolt, "Put-that-there": Voice and Gesture at the Graphics Interface, in: Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '80, ACM, New York, NY, USA, 1980, pp. 262–270.

[14] P. R. Cohen, M. Johnston, D. McGee, S. Oviatt, J. Pittman, I. Smith, L. Chen, J. Clow, Quickset: Multimodal interaction for distributed applications, in: Proceedings of the fifth ACM International Conference on Multimedia, ACM, 1997, pp. 31–40.

[15] M. Johnston, S. Bangalore, G. Vasireddy, A. Stent, P. Ehlen, M. Walker, S. Whittaker, P. Maloor, Match: An architecture for multimodal dialogue systems, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 2002, pp. 376–383.

[16] M. Poesio, S. Ponzetto, Y. Versley, Computational models of anaphora resolution: A survey, Linguistic Issues in Language Technology.

[17] F. Landragin, Visual perception, language and gesture: A model for their understanding in multimodal dialogue systems, Signal Processing 86 (12) (2006) 3578–3595.

[18] J. Eisenstein, R. Davis, Gesture Improves Coreference Resolution, in: Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers, 2006, pp. 37–40.

[19] T. Baldwin, J. Y. Chai, K. Kirchhoff, Communicative gestures in coreference identification in multiparty meetings, in: Proceedings of the 2009 International Conference on Multimodal Interfaces, ACM, 2009, pp. 211–218.

[20] Z. Prasov, J. Y. Chai, What's in a gaze?: the role of eye-gaze in reference resolution in multimodal conversational interfaces, in: Proceedings of the 13th international conference on Intelligent user interfaces, IUI '08, ACM, New York, NY, USA, 2008, pp. 20–29. doi:http://doi.acm.org/10.1145/1378773.1378777.

[21] R. Iida, M. Yasuhara, T. Tokunaga, Multi-modal reference resolution in situated dialogue by integrating linguistic and extra-linguistic clues., in: IJCNLP, 2011, pp. 84–92.

[22] C. Liu, R. Fang, J. Y. Chai, Shared gaze in situated referential grounding: An empirical study, in: Eye Gaze in Intelligent User Interfaces, Springer, 2013, pp. 23–39.

[23] R. Wasinger, A. Krüger, O. Jacobs, Integrating intra and extra gestures into a mobile and multimodal shopping assistant, in: Pervasive Computing, Springer, 2005, pp. 297–314.

[24] V. Samek-Lodovici, C. Strapparava, Identifying Noun Phrase References: the Topic Module of the AlFresco System, in: Proceedings of ECAI 90, Ninth European Conference on Artificial Intelligence, 1990, pp. 573–578.

[25] F. Landragin, The role of gesture in multimodal referring actions, in: The Fourth IEEE International Conference on Multimodal Interfaces, IEEE Computer Society, Pittsburgh, PA, USA, 2002, pp. 173–178. doi:http://doi.ieeecomputersociety.org/10.1109/ICMI.2002.1166988.

[26] H. Bunt, Context and dialogue control, Think Quarterly 3 (1) (1994) 19–31.

[27] J. Austin, How to do things with words, Harvard University Press, 1962.

[28] A. H. Anderson, M. Bader, E. G. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. S. Thompson, R. Weinert, The HCRC Map Task Corpus, Language and Speech 34 (4) (1991) 351–366.

[29] L. Levin, A. Thymé-Gobbel, A. Lavie, K. Ries, K. Zechner, A discourse coding scheme for conversational Spanish, in: Fifth International Conference on Spoken Language Processing, 1998, pp. 2335–2338.

[30] D. Graff, A. Canavan, G. Zipperlen, Switchboard-2 Phase I, LDC 99S79–http://www.ldc.upenn.edu/Catalog (1998).

[31] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, H. Carvey, The ICSI Meeting Recorder Dialog Act (MRDA) Corpus, in: Proceedings of 5th SIGdial Workshop on Discourse and Dialogue, Cambridge, MA, 2004, pp. 97–100.

[32] B. Di Eugenio, Z. Xie, R. Serafin, Dialogue act classification, higher order dialogue structure, and instance-based learning, Dialogue & Discourse 1 (2) (2010) 1–24.

[33] N. Webb, M. Ferguson, Automatic extraction of cue phrases for cross-corpus dialogue act classification, in: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Association for Computational Linguistics, 2010, pp. 1310–1317.

[34] A. Ezen-Can, K. E. Boyer, Unsupervised classification of student dialogue acts with query-likelihood clustering, in: International Conference on Educational Data Mining, 2013, pp. 20–27.

[35] V. Ashok, Y. Borodin, S. Stoyanchev, I. Ramakrishnan, Dialogue act modeling for non-visual web access, in: Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), Association for Computational Linguistics, Philadelphia, PA, U.S.A., 2014, pp. 123–132.

[36] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, M. Meteer, Dialogue act modeling for automatic tagging and recognition of conversational speech, Computational Linguistics 26 (3) (2000) 339–373.

[37] J. Ang, Y. Liu, E. Shriberg, Automatic dialog act segmentation and classification in multiparty meetings, in: Proc. ICASSP, Vol. 1, 2005, pp. 1061–1064.

[38] V. Sridhar, S. Bangalore, S. Narayanan, Combining lexical, syntactic and prosodic cues for improved online dialog act tagging, Computer Speech & Language 23 (4) (2009) 407–422.

[39] K. Boyer, J. Grafsgaard, E. Ha, R. Phillips, J. Lester, An affect-enriched dialogue act classification model for task-oriented dialogue, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, Association for Computational Linguistics, 2011, pp. 1190–1199.

[40] E. Y. Ha, J. F. Grafsgaard, C. Mitchell, K. E. Boyer, J. C. Lester, Combining verbal and nonverbal features to overcome the "information gap" in task-oriented dialogue, in: Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Association for Computational Linguistics, Seoul, South Korea, 2012, pp. 247–256.

[41] S. Qu, J. Y. Chai, Beyond attention: the role of deictic gesture in intention recognition in multimodal conversational interfaces, in: Proceedings of the 13th International Conference on Intelligent User Interfaces, 2008, pp. 237–246.

[42] S. Oviatt, P. Cohen, L. Wu, L. Duncan, B. Suhm, J. Bers, T. Holzman, T. Winograd, J. Landay, J. Larson, D. Ferro, Designing the User Interface for Multimodal Speech and Pen-Based Gesture Applications: State-of-the-Art Systems and Future Research Directions, Human-Computer Interaction 15 (4) (2000) 263–322.

[43] J. Y. Han, Low-cost Multi-touch Sensing Through Frustrated Total Internal Reflection, in: Proceedings of the 18th Annual ACM Symposium on User Interface Software and Technology, UIST '05, ACM, New York, NY, USA, 2005, pp. 115–118.

[44] B. Buxton, Multi-touch systems that I have known and loved, Microsoft Research 56 (2007) 1–11.

[45] D. Wigdor, J. Fletcher, G. Morrison, Designing User Interfaces for Multi-touch and Gesture Devices, in: CHI '09 Extended Abstracts on Human Factors in Computing Systems, CHI EA '09, ACM, New York, NY, USA, 2009, pp. 2755–2758.

[46] J. Luk, J. Pasquero, S. Little, K. MacLean, V. Levesque, V. Hayward, A Role for Haptics in Mobile Interaction: Initial Design Using a Handheld Tactile Display Prototype, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '06, ACM,

New York, NY, USA, 2006, pp. 171–180.

[47] M. J. Hertenstein, D. Keltner, B. App, B. A. Bulleit, A. R. Jaskolka, Touch communicates distinct emotions, Emotion 6 (3) (2006) 528–533.

[48] T. Noda, H. Ishiguro, T. Miyashita, N. Hagita, Map acquisition and classification of haptic interaction using cross correlation between distributed tactile sensors on the whole body surface, in: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2007, pp. 1099–1105.

[49] P. Mittendorfer, G. Cheng, Self-organizing sensory-motor map for low-level touch reactions, in: 2011 11th IEEE-RAS International Conference on Humanoid Robots (Humanoids), 2011, pp. 59–66.

[50] M. Cooney, S. Nishio, H. Ishiguro, Recognizing affection for a touch-based interaction with a humanoid robot, in: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2012, pp. 1420–1427.

[51] D. Silvera-Tawil, D. Rye, M. Velonaki, Interpretation of Social Touch on an Artificial Arm Covered with an EIT-based Sensitive Skin, International Journal of Social Robotics 6 (4) (2014) 489–505.

[52] H. Barron-Gonzalez, T. Prescott, Discrimination of Social Tactile Gestures Using Biomimetic Skin, in: A. Natraj, S. Cameron, C. Melhuish, M. Witkowski (Eds.), Towards Autonomous Robotic Systems, Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2014, pp. 46–48.

[53] W. Stiehl, J. Lieberman, C. Breazeal, L. Basel, L. Lalla, M. Wolf, Design of a therapeutic robotic companion for relational, affective touch, in: (RO-MAN 2005: The 14th IEEE International Workshop on Robot and Human Interactive Communication, 2005, pp. 408–415.

[54] S. Yohanan, K. E. MacLean, The haptic creature project: Social human-robot interaction through affective touch, in: Proceedings of the AISB 2008 Symposium on the Reign of Catz & Dogs: The Second AISB Symposium on the Role of Virtual Creatures in a Computerised Society, Vol. 1, Citeseer, 2008, pp. 7–11.

[55] S. Ekvall, D. Kragic, Grasp recognition for programming by demonstration, in: Proceedings of the 2005 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2005, pp. 748–753.

[56] S. Ratnasingam, T. McGinnity, Object recognition based on tactile form perception, in: IEEE Workshop on Robotic Intelligence in Informationally Structured Space (RIiSS), IEEE, 2011, pp. 26–31.

[57] M. Johnsson, C. Balkenius, Haptic perception with a robotic hand, in: Proceedings of the Ninth Scandinavian Conference on Artificial Intelligence (SCAI 2006), Espoo, Finland, 2006, pp. 127–134.

[58] M. Schöpfer, M. Pardowitz, R. Haschke, H. Ritter, Identifying relevant tactile features for object identification, in: Towards Service Robots for Everyday Environments, Springer, 2012, pp. 417–430.

[59] B. T. Taylor, V. M. Bove Jr, Graspables: grasp-recognition as a user interface, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2009, pp. 917–926.

[60] R. Wimmer, Flyeye: grasp-sensitive surfaces using optical fiber, in: Proceedings of the fourth International Conference on Tangible, Embedded, and Embodied Interaction, ACM, 2010, pp. 245–248.

[61] B. Di Eugenio, M. Žefran, J. Ben-Arie, M. Foreman, L. Chen, S. Franzini, S. Jagadeesan, M. Javaid, K. Ma, Towards Effective Communication with Robotic Assistants for the Elderly: Integrating Speech, Vision and Haptics, in: Dialog with Robots, AAAI 2010 Fall Symposium, Arlington, VA, USA, 2010.

[62] F. De la Torre, J. Hodgins, J. Montano, S. Valcarcel, R. Forcada, J. Macey, Guide to the Carnegie Mellon University Multimodal Activity (CMU-MMAC) Database, Tech. Rep. CMU-RI-TR-08-22, Robotics Institute, Carnegie Mellon University (2009).

[63] M. Swift, G. Ferguson, L. Galescu, Y. Chu, C. Harman, H. Jung, I. Perera, Y. C. Song, J. Allen, H. Kautz, A multimodal corpus for integrated language and action, in: N. Calzolari, K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the LREC'12 Workshop: Multimodal Corpora: How Should Multimodal Corpora Deal with the Situation?, European Language Resources Association (ELRA), Istanbul, Turkey, 2012.

[64] L. Dipietro, A. Sabatini, P. Dario, A survey of glove-based systems and their applications, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 38 (4) (2008) 461–482.

[65] A. Hollinger, M. M. Wanderley, Evaluation of Commercial Force-Sensing Resistors, Tech. rep., School of Music, McGill University Montreal, QC, Canada (January 2006).

[66] C. Lebosse, B. Bayle, M. de Mathelin, P. Renaud, Nonlinear modeling of low cost force sensors, in: IEEE International Conference on Robotics and Automation, IEEE, 2008, pp. 3437–3442.

[67] Arduino-homepage (Feb. 23 2014).
URL `http://arduino.cc/`

[68] M. Kipp, Anvil-a generic annotation tool for multimodal dialogue, in: Proceedings of the 7th European Conference on Speech Communication and Technology, 2001, pp. 1367–1370.

[69] J. Eisenstein, R. Davis, Gesture features for coreference resolution, Machine Learning for Multimodal Interaction (2006) 154–165.

[70] J. Allwood, L. Cerrato, K. Jokinen, C. Navarretta, P. Paggio, The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena, Language Resources and Evaluation 41 (3-4) (2007) 273–287.

[71] S. Kopp, B. Krenn, S. Marsella, A. N. Marshall, C. Pelachaud, H. Pirker, K. R. Thórisson, H. Vilhjálmsson, Towards a common framework for multimodal generation: The behavior markup language, in: Intelligent Virtual Agents, Springer, 2006, pp. 205–217.

[72] K. Bergmann, S. Kopp, Gestural alignment in natural dialogue, in: Proceedings of the 34th Annual Conference of the Cognitive Science Society (CogSci 2012), 2012, pp. 1326–1331.

[73] H. Bunt, J. Alexandersson, J.-W. Choe, A. C. Fang, K. Hasida, V. Petukhova, A. Popescu-Belis, D. R. Traum, ISO 24617-2: A semantically-based standard for dialogue annotation, in: LREC, ELRA, 2012, pp. 430–437.

[74] P. Wagner, Z. Malisz, S. Kopp, Gesture and speech in interaction: An overview, Speech Communication 57 (2014) 209–232.

[75] K. Krippendorff, Content analysis: An introduction to its methodology, Sage Publications, Beverly Hills, CA, 1980.

[76] J. Carletta, S. Isard, G. Doherty-Sneddon, A. Isard, J. C. Kowtko, A. H. Anderson, The reliability of a dialogue structure coding scheme, Computational Linguistics 23 (1) (1997) 13–31.

[77] B. Di Eugenio, M. Glass, The kappa statistic: A second look, Computational linguistics 30 (1) (2004) 95–101.

[78] J. Cohen, A coefficient of agreement for nominal scales, Educational and Psychological Measurement 20 (1) (1960) 37–46.

[79] J. Fleiss, Measuring nominal scale agreement among many raters., Psychological Bulletin 76 (5) (1971) 378.
[80] T. Rietveld, R. van Hout, Statistical Techniques for the Study of Language and Language Behaviour, Mouton de Gruyter, Berlin - New York, 1993.
[81] S. Larsson, D. R. Traum, Information state and dialogue management in the TRINDI dialogue move engine toolkit, Natural Language Engineering 6 (3&4) (2000) 323–340.
[82] D. R. Traum, S. Larsson, The information state approach to dialogue management, in: Current and New Directions in Discourse and Dialogue, Springer, 2003, pp. 325–353.
[83] H. Buschmeier, S. Kopp, Unveiling the information state with a bayesian model of the listener, in: SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, 2011.
[84] D. Z. Hakkani-Tür, G. Tür, L. P. Heck, A. Fidler, A. Celikyilmaz, A discriminative classification-based approach to information state updates for a multi-domain dialog system, in: Interspeech, 2012.
[85] D. Jurafsky, J. H. Martin, Speech and Language Processing, 2nd Edition, Pearson Education – Prentice Hall, 2009.
[86] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, The WEKA data mining software: An update, SIGKDD Explorations 11 (1).
[87] C. Chang, C. Lin, LIBSVM: A Library for Support Vector machines, ACM Transactions on Intelligent Systems and Technology (TIST) 2 (3) (2011) 27.
[88] L. Carlson, Dialogue games: An approach to discourse analysis, D. Reidel Publishing Company, 1985.
[89] H. W. Hastie, M. Poesio, S. Isard, Automatically predicting dialogue structure using prosodic features, Speech Communication 36 (1–2) (2002) 63–79.
[90] S. N. Kim, L. Cavedon, T. Baldwin, Classifying dialogue acts in multi-party live chats, in: Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation, Faculty of Computer Science, Universitas Indonesia, Bali, Indonesia, 2012, pp. 463–472.
[91] E. Ivanovic, Automatic instant messaging dialogue using statistical models and dialogue acts, Master's thesis, University of Melbourne (2008).
[92] J. Ben-Arie, Method of recognition of human motion, vector sequences and speech, US Patent 7,366,645 (Apr. 29 2008).
[93] K. Ma, J. Ben-Arie, Multi-view multi-class object detection via exemplar compounding, in: IEEE-IAPR 21st International Conference on Pattern Recognition (ICPR 2012), Tsukuba, Japan, 2012, pp. 3256–3259.
[94] N. Folbre, Conceptualizing care, in: Frontiers in the Economics of Gender, Routledge, 2008, pp. 101–115.
[95] E. Noohi, M. Žefran, Quantitative measures of cooperation for a dyadic physical interaction task, in: 2014 14th IEEE-RAS International Conference on Humanoid Robots (Humanoids), 2014, pp. 469–474.
[96] M. Javaid, Communication through physical interaction: Robot assistants for the elderly, Ph.D. thesis, University of Illinois at Chicago (2014).
[97] S. Franzini, J. Ben-Arie, Speech recognition by indexing and sequencing, in: International Conference of Soft Computing and Pattern Recognition (SoCPaR), 2010, pp. 93–98.
[98] H. Sakoe, S. Chiba, Dynamic programming algorithm optimization for spoken word recognition, IEEE Transactions on Acoustics, Speech and Signal Processing, 26 (1) (1978) 43–49.
[99] M. Javaid, M. Žefran, B. Di Eugenio, Communication through physical interaction: A study of human collaborative manipulation of a planar object, in: RO-MAN 2014: The 23rd IEEE International Symposium on Robot and Human Interactive Communication, 2014, pp. 838–843.
[100] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, A. Y. Ng, ROS: an open-source robot operating system, in: ICRA Workshop on Open Source Software, 2009.
[101] U. Ahmed, Implementing the Robot Modules in ROS, Master's thesis, University of Illinois at Chicago (2014).
[102] D. Gouaillier, V. Hugel, P. Blazevic, C. Kilner, J. Monceaux, P. Lafourcade, B. Marnier, J. Serre, B. Maisonnier, Mechatronic design of NAO humanoid, in: ICRA '09, IEEE International Conference on Robotics and Automation, 2009, pp. 769–774.
[103] S. Bangalore, G. Di Fabbrizio, A. Stent, Learning the structure of task-driven human–human dialogs, IEEE Transactions on Audio, Speech, and Language Processing 16 (7) (2008) 1249–1259.
[104] S. N. Kim, L. Cavedon, T. Baldwin, Classifying dialogue acts in one-on-one live chats, in: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2010, pp. 862–871.
[105] M.-C. De Marneffe, C. D. Manning, The Stanford typed dependencies representation, in: Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation, 2008, pp. 1–8.

Lin Chen received his B.A. in Linguistics in 2004, and his M.A. in Computational Linguistics in 2007, both from Tsinghua University, Beijing, China. He received his Ph.D. in Computer Science in 2014 from the University of Illinois at Chicago, where he was a member of the Natural Language Processing laboratory. His research interests include multimodal dialogue systems, intelligent tutoring systems, and Information Extraction.

Maria Javaid received her B.Sc. and M.Sc. in Electrical Engineering in Dec. 2004 and July 2008 respectively from the University of Engineering and Technology, Lahore, Pakistan. She is currently a Ph.D. Candidate in Robotics in the Department of Electrical and Computer Engineering at the University of Illinois at Chicago, Chicago, IL. Her research interests include human robot interaction and haptics.

Barbara Di Eugenio (Ph.D., University of Pennsylvania, 1993) is currently a Professor of Computer Science at the University of Illinois at Chicago. She has published extensively in Natural Language Processing. Her research has been funded by the National Science Foundation (NSF), the Office of Naval Research, Motorola, Yahoo!, and the Qatar Research Foundation. She is a recipient of the NSF CAREER award (2002), and of the 2013 AWIS Chicago Innovator of the Year Award.

Miloš Žefran (Ph.D., University of Pennsylvania, 1996) is currently a Professor of Electrical and Computer Engineering at the University of Illinois at Chicago. His research interests include robotics, modeling and control of hybrid systems, human manipulation, rehabilitation. His research has been funded by the National Science Foundation (NSF), and he is a recipient of the NSF CAREER award (2001).