The Structure and Generality of Spoken Route Instructions

Aasish Pappu and Alexander Rudnicky
Language Technologies Institute, Carnegie Mellon University
{aasish, air}@cs.cmu.edu

Abstract

A robust system that understands route instructions should be able to process instructions generated naturally by humans. Also desirable would be the ability to handle repairs and other modifications to existing instructions. To this end, we collected a corpus of spoken instructions (and modified instructions) produced by subjects provided with an origin and a destination. We found that instructions could be classified into four categories, depending on their intent such as imperative, feedback, or meta comment. We asked a different set of subjects to follow these instructions to determine the usefulness and comprehensibility of individual instructions. Finally, we constructed a semantic grammar and evaluated its coverage. To determine whether instruction-giving forms a predictable sub-language, we tested the grammar on three corpora collected by others and determined that this was largely the case. Our work suggests that predictable sub-languages may exist for well-defined tasks.

Index Terms: Robot Navigation, Spoken Instructions

1 Introduction

Generating and interpreting instructions is a topic of enduring interest. Cognitive psychologists have examined how people perceive spatial entities and structure route instructions (Daniel and Denis, 1998; Allen, 1997). Linguists and others have investigated how people articulate route instructions in conversation with people or agents (Eberhard et al., 2010; Gargett et al., 2010; Stoia et al., 2008; Marge and Rudnicky, 2010). Artificial intelligence researchers have shown that under supervised conditions autonomous agents can learn to interpret route instructions (Kollar et al., 2010; MacMahon et al., 2006; Matuszek et al., 2010; Bugmann et al., 2004; Chen and Mooney, 2010).

While the subject has been approached from different perspectives, it has been generally held that the language of directions is mostly limited and only parts of the vocabulary (such as location names) will vary from case to case. We are interested in being able to interpret natural directions, as might be given to a robot, and generating corresponding trajectory. But natural directions contain different types of information, some (more-or-less) easily interpreted (e.g., "go to the end of the hall") while others seem daunting (e.g., "walk past the abstract mural with birds"). So the question might actually be "is there enough interpretable data in human directions to support planning a usable trajectory?".

The language of instructions contains a variety of relevant propositions: a preface to a route, an imperative statement, or a description of a landmark. Previous work has proposed both coarse and fine-grained instruction taxonomies. (Bugmann et al., 2004) proposed a taxonomy of 15 primitive categories in a concrete "action" framework. In contrast, (Daniel and Denis, 1998) suggested a five-way categorization based on cognitive properties of instructions.

Instructions vary greatly and can include superfluous detail. (Denis et al., 1999) found that when people were asked to read and assess a set of instructions some of the instructions were deemed unnecessary and could be discarded. There is some evidence (Lovelace et al., 1999; Caduff and Timpf, 2008) that only the mention of significant landmarks along the route leads to better-quality instructions. Computational (rather than descriptive) approaches to this problem include: using sequence labeling approach to capture spatial relations, landmarks, and action verbs (Kollar et al., 2010), generating a frame structure for an instruction (MacMahon et al., 2006), or using statistical machine translation techniques to translate instructions into actions (Matuszek et al., 2010).

We describe a new instructions corpus, its analysis in terms of a taxonomy suitable for automated understanding and a verification that the instructions are in fact usable by humans. With a view to automating understanding, we also constructed a grammar capable of processing this language, and show that it provides good coverage

Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), pages 99–107,
Seoul, South Korea, 5-6 July 2012.

for both our corpus and three other corpora (Kollar et al., 2010; Marge and Rudnicky, 2010; Bugmann et al., 2004)

This paper is organized as following: Section 2 describes the corpus collection study. Then in Section 3, we discuss the taxonomy of route instructions. Section 4 focuses on which categories are important for navigation. In Section 5, we report our results and error analysis on parsing instructions from our corpus and three other corpora containing route instructions, followed by lessons learned and future work.

2 The Navagati[1] Corpus

We collected a corpus of spoken instructions describing how to get from one part of a large building complex to another. To ensure consistency we recruited individuals who were familiar with the environment and consequently could formulate such instructions without reference to maps or other materials. Since we are ultimately interested in how such instructions are edited, we also included conditions in which subjects were asked to modify their instructions in several ways. The corpus is publicly available[2].

2.1 Participants and Procedure

We recruited subjects who were both fluent English speakers and were also familiar with the environment (a university building complex). Subjects were told to imagine that they had encountered a visitor, not familiar with the campus, at a specific location (in front of elevators on a particular floor) who needed instructions to a specific location, a café two buildings away.

For each set of instructions, subjects were asked to think about the route and their instructions, then record them as a single monologue. Subjects sat in front of a computer and wore a close-talking microphone. Initially no map was provided and they were expected to rely on their memory. In subsequent tasks they were shown a floor-plan indicating a specific location of the visitor and asked to modify their instructions. Speech was transcribed using Amazon Mechanical Turk, shown to be a reliable resource for spoken language transcription (Marge et al., 2010). Transcriptions were normalized to standardize spellings (e.g., building names).

2.2 Design

Previous works have focused on eliciting route instructions between multiple pairs of locations. There is a general agreement that the structure of instructions did not vary with the increase in number of start-end location pairs. However previous works have not looked at how instructions would be modified under different situations.

We were interested in two general cases: normal instructions (**Simple** scenario) and repairing existing instructions (**Repair** scenario). Each scenario included three tasks, as described below.

We selected two locations that could be walked between without necessarily going outside. However the subjects were free to to give instructions for a route of their choice between a location pair. The first location (A) was in front of an elevator on the seventh floor of Gates Hillman Center, the second location (B) was a cafe on the fifth floor of Wean Hall. The expected pathway included changes in floor, direction and passing through a different building. It required reasonably detailed instructions.

In the **Simple** scenario, subjects were asked to generate three variants, as follows: (1) instructions for $A \rightarrow B$; (2) for $B \rightarrow A$; and (3) a simplified version of (2).

The motivation behind (2) is to learn whether people would make references about the parts of the route that were previously traversed in the opposite direction. In the case of (3), we were interested in the degree of instruction reuse and the condensation strategy. We explicitly told the subject "Imagine that the visitor found your instructions confusing. They asked you to simplify the instructions. How would you do that?"

The **Repair** scenario was designed to probe how a subject would alter their instructions in response to complications. Subjects were asked to modify their intial Simple instructions ($A \rightarrow B$) to cope with: (1) visitor missing a landmark and takes a wrong turn; (2) an obstruction (construction) blocking the original path; and (3) the visitor getting lost and ends up in an unknown part of the (middle) building. For each case, the subject was given a map (as in figure 1) that marked the visitor's location and had to get the visitor back on track.

Figure 1: Map of the construction area (marked as star)

The tasks in this scenario were designed to see whether people modify directions differently when three different situations are presented. Precisely, we want to know if

[1] Sanskrit root for Navigation meaning "to travel by boat"

[2] http://tts.speech.cs.cmu.edu/apappu/navagati/

there is any difference in the discourse structure and verbosity of the directions.

2.3 Analysis

Nine subjects performed 6 tasks each, producing 54 sets of instructions, for a total of 65 minutes of speech. Please note that other corpora in the route instructions domain have similiar scale (see Figure 5(a)). The transcriptions were segmented semi-automatically into atomic units corresponding to instruction steps. For example, the instruction "Go left, then turn right" was segmented into: "go left", and "then turn right" based on bigram heuristics. We compiled a list of most frequent bigrams and trigrams in the corpus e.g., "and then", "after that" etc. The transcriptions were segmented at the bigram/trigram boundaries and were manually verified for the correctness of a segment. The Simple scenario generated 552 instructions, the Repair part contained 382 instructions, a total of 934. The vocabulary has 508 types and 7937 tokens. Table 1 summarizes the factors measured in both the scenarios. Only two (marked by *) differed between scenarios (t-test at $p < 0.05$). We examined acoustic properties (for example mean pitch) but did not find any significant differences across scenario type.

Table 1: Simple vs Repair Scenario

Factors	Simple	Repair
# Tokens	4461	3476
# Types	351	375
# Instructions	552	382
# Words-per-Instruction*	7.5	8.0
# Landmarks	450	314
# Motion Verbs*	775	506
# Spatial Prepositions	61	60
# Filler Phrases	414	380

We can compare language similarity across scenarios by comparing the perplexity of text in the two scenarios. If the instructions and repairs are similar, we would expect that a model built from one scenario should be able to capture data from the other scenario. We randomly divided data from each scenario into training (70%) and testing data (30%). We built a trigram language model (LM) smoothed with absolute discounting using the CMU-SLM toolkit (Rosenfield, 1995). Then, we computed the perplexity on testing data from each scenario against each model. From Table 2, Simple-LM has lower perplexity compared to Repair-LM on the test sets. The perplexity of Simple-LM on Repair-Test is slightly higher when compared to Simple-Test. This could be due to the lexical diversity of the Repair scenario or simply to the smaller sample size. Table 1 (row 1) indicates that the data in Repair scenario is smaller than data in Simple scenario. To explore the lexical diversity of these two scenarios we conducted a qualitative analysis of the instructions from both the scenarios.

In Task 1 of the Simple scenario, we only observed a sequence of instructions. However in Task 2 of Simple Scenario, we noticed references to instructions from Task 1 via words like "remember", "same route", etc. This suggests that instructions may be considered in context of previous exchanges and that this history should normally be available for interpretation purposes. In Task 3 of the Simple scenario, 7 out of 9 subjects simply repeated the instructions from Task 2 while the rest provided a different version of the same instructions. We did not observe any other qualitative differences across three tasks in the Simple scenario.

In Task 1 of the Repair scenario, all but one subject gave instructions that returned the visitor to the missed landmark, instead of bypassing the landmark. In Task 2, the obstruction on the path could be negotiated through a shorter or longer detour. But only 4 out of 9 participants suggested the shorter detour. In Task 3, we did not observe anything different from Task 2. Despite the difference in the situations, the language of repair was found to be quite similar. The structure of the delivery was organized as follows: (1) Subjects introduced the situation of the visitor; (2) then modified the instructions according to the situation. Introduction of the situation was different in each task, (e.g., "you are facing the workers" vs "looks like you are near office spaces" vs "if you have missed the atrium you took a wrong turn"). But the modification or repair of the instructions was similar across the situations. The repaired instructions are sequences of instructions with a few cautionary statements inserted between instructions. We believe that subjects added cautionary statements in order to warn the visitor from going off-the-route. We observed that 6.3% of the repaired instructions were *cautionary* statements; we did not observe cautionary statements in the original Simple scenario. In order to see the effect of these cautionary statements we removed them from both training and testing sets of the Repair scenario, then built a trigram LM using this condensed training data (Repair–w/o-cautionLM). Table 2 shows that perplexity drops when cautionary statements are excluded from the repair scenario, indicating that Simple and Repair scenarios are similar except for these cautionary statements.

3 Taxonomy of Route Instructions

Taxonomies have been proposed in the past. Daniel and Denis (1998) proposed a taxonomy that reflected attributes of spatial cognition and included 5 classes: (1) Imperatives; (2) Imperatives referring a landmark; (3) Introduction of a landmark without an action; (4) Non-spatial description of landmarks and (5) Meta comments.

Table 2: Perplexity of Simple/Repair Language Models

LM/Test	Simple-Test	Repair-Test	Repair -w/o-caution
Simple-LM	29.6	36.5	30.3
Repair-LM	37.4	37.3	35.6
Repair -w/o-cautionLM	31.9	37.6	26.8

Bugmann et al. (2004) suggested 15 primitive (robot-executable) actions. We present a hierarchical instruction taxonomy that takes into account both cognitive properties and the needs of robot navigation. This taxonomy is based on 934 route instruction monologues. It should be noted that this taxonomy is not based on dialog acts but rather takes the intent of the instruction into the account.

3.1 Categories

We segmented the spoken instructions using a criterion that split individual actions and observations. Our taxonomy is roughly comparable to that of (Daniel and Denis, 1998) but differs in the treatment of landmarks because the mention of the landmarks in an instruction can be of two types: contextual mention and positional mention. Contextual Mention means when a landmark in the surroundings but it is not on the path. On the other hand, positional mention requires the landmark to be on the path. In our taxonomy, contextual mention becomes Advisory instruction and positional mention is called Grounding instruction. The taxonomy has four major categories that subsume 18 sub-categories; these are given in Table 3.

For instance, "You want to take a right" belongs to the Imperative category. "You will see a black door" is an Advisory instruction about the surroundings. "You are on the first floor" denotes Grounding. "Your destination is located in another building and you will walk across three buildings in this route" gives an overview of the route, a Meta Comment. From Figure 2, we see that majority of the route instructions are Imperative.

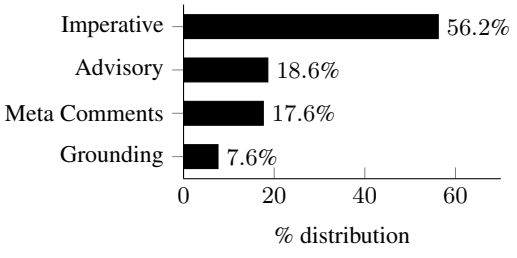

Figure 2: First Tier Instruction Categories

3.1.1 Imperative Instructions

Imperative instructions are executable and can result in physical displacement. We identified seven subcategories of Imperatives that distinguish different contexts (e.g., going along a corridor, changing floors via elevator or stairs, or going to a specific location).

Imperative instructions can also include *preconditions* or *postconditions*. The order of their execution varies based on the directionality of the condition between two instructions. `Continue` is interesting because it can have *travel-distance* and *travel-direction* arguments, or even no arguments. In the latter case the follower continues an action (e.g., "keep walking"), until some unspecified condition ends it.

3.1.2 Advisory Instructions

While giving route instructions people mention landmarks along the route as feedback to the direction-follower. Some of these landmarks are not part of the path but do serve as waypoints for the follower (e.g., "you will see a hallway right there"). We observe that landmarks are distinct either functionally and/or physically. For example, a hallway is both functionally and physically different from an elevator but only physically different from a door because both function as an instrument (or path) to get from one place to another. Based on this distinction, we divided advisory instructions into five sub-categories depending on the type of landmark mentioned in the instruction (see Table 3).

Compound locations (see Table 3) are closely located but physically distinct. They may constitute part-whole relationships e.g., "TV screen with a motion sensor". We observed that compound locations are used to disambiguate when multiple instances of a landmark type are present e.g., "chair near the elevator vs "chair near the hallway".

3.1.3 Grounding Instructions

Grounding instructions report absolute position. These instructions indicate current view or location as opposed to future view or location (indicated through advisory instructions). These instructions constitute a landmark name similar to advisory instructions and also follow the distinction between the type of landmark mentioned in the instruction (see Table 3).

3.1.4 Meta Comments

Meta comments are non-executable instructions added to route instructions. People often make these comments at the beginning of instructions and sometimes in between two imperative statements e.g., a precautionary statement. In our corpus we found meta-comments in two situations: (1) Preface or introduction of the route; (2) Caution against a (metaphorical) pitfall in the route.

Category	SubCategory	Distribution	Example
Imperative	Leave-Location	2.3%	Exit the building; Come out of the room
	Follow-Path	7.0%	Walk along the corridor; go across the bridge
	Floor-Transition	11.2%	Take the elevator to fourth floor; Take the stairs to the fifth
	Turn	24.2%	Turn left
	Go-To	27.2%	Walk to the elevators
	Continue	28.0%	Keep going straight for few steps
Advisory	Floor-Level	5.4%	You will see fourth floor of other building
	Floor-Transition	12.2%	You will see elevators
	Compound-Location	13.4%	You will see a hallway to the right of elevators
	End-of-Pathway	21.5%	You will see end of the hallway
	Landmark	47.5%	You will see a TV screen
Grounding	Compound-Location	5.9%	You are on a hallway right next to the elevators
	End-of-Pathway	8.2%	You are on the bridge leading to other building
	Floor-Level	42.4%	You are on fourth floor of the building
	Landmark	43.5%	You are on standing near TV screen
Meta Comments	Caution	14.7%	You can find it immediately; Don't go that side
	Miscellaneous	36.0%	Let me guide you through it; I guess a simpler way would be
	Preface	49.3%	I will guide you to the cafe in that building

Table 3: Taxonomy of Categories with Examples

Both the example instructions and the distribution of the subcategories are given in Table 3.

The language of meta comments is more diverse than that of the other three categories. If we build trigram language models for each category and measure the perplexity on a held-out set from same category the perplexity is relatively high for Meta (49.6) compared to other categories (Advisory: 19.5; Imperative: 18.5; Grounding: 11.4). This suggests that automatic understanding of meta comments might be problematic, consequently it would be useful to determine the ralative utility of different instruction categories. The next section describes at attempt to do this.

4 Which Instructions are Relevant?

Given a variety of information present in a set of route instructions, we wanted to investigate whether all that information is relevant for navigation. In order to find that out we devised a user study asking people to follow instructions collected in our previous study. (Daniel and Denis, 1998) conducted a similar study where they asked subjects to read a set of instructions and strike-off instructions with too much or too little information. However, people may or may not feel the same when they follow (physically navigate) these instructions. Therefore, in our study the experimenter read instructions (of varying amount of detail) to the subjects while they physically navigated through the environment.

4.1 Participants and Procedure

We chose 5 out of the 9 instruction sets, spoken by different subjects (of average length 26.8 instructions per set) from Task 1 of the Simple scenario discussed above. We did not use the others because they contained few instructions (average of 13.5) and provided fewer instances of instructions in different categories. Also, we did not use instructions from Repair Scenario because those instructions dependent on a scenario and a set of instructions that were already provided to the direction follower.

Our set of instructions included the full set, a set with only imperatives and additional sets adding only one of the remaining categories to the imperative set (see Table 4), producing 25 distinct sets of instructions. Additionally, building names and the destination name (transcribed in the instructions) were anonymized to avoid revealing the destination or the "heading" at the early stage of the route.

We recruited 25 subjects, each doing one variant of the instructions. In the session, the experimenter read one instruction at a time to the subject and walked behind the subject as they proceeded. Subjects were asked to say "done" when ready for the next instruction; they were allowed to ask the experimenter to repeat instructions but otherwise were on their own. The experimenter kept track of how and where a subject got lost on their way to destination. (No systematic effects were observed, but see below.) At the end subjects were handed the entire set of instructions and were asked to mark which instructions were difficult to follow and which were redundant. Remaining instructions were deemed to be useful and interpretable.

Table 4: Variants of an Instruction Set

Variant	Imperative	Advisory	Grounding	Meta
Imp	✓			
Imp+Adv	✓	✓		
Imp+Grnd	✓		✓	
Imp+Meta	✓			✓
Entire Set	✓	✓	✓	✓

Category/Variant	Imp	Imp+Grnd	Imp+Meta	Imp+Adv	Entire Set
Diff-Imp	11	10	12	9	12
Diff-Adv	0	10	5	10	10
Diff-Grnd	0	0	13	0	0
Diff-Meta	4	15	12	4	4
Diff-All	6	9	11	7	9

Category/Variant	Imp	Imp+Grnd	Imp+Meta	Imp+Adv	Entire Set
Redun-Imp	5	8	12	11	8
Redun-Adv	5	10	19	10	29
Redun-Grnd	20	13	47	53	27
Redun-Meta	19	31	65	23	50
Redun-All	9	13	26	17	21

Figure 3: What percent of instructions are Difficult (Diff) or Redundant (Redun)? On the `left`: Darker is Difficult `right`: Darker is More Redundant Instructions

4.2 Analysis

Except for one subject, everybody reached the destination. Subjects found Imperative and Advisory instructions more useful compared to Grounding instructions and Meta comments, irrespective of the instruction-set they followed (see Figure 3). Figure 3(a) shows percentage of category-wise difficult instructions in each variant of an instruction set and 3(b) shows percentage of category-wise redundant instructions in each variant of an instruction set. For e.g., Diff-Imp/Imp+Meta means that 12% of imperative-instructions are difficult in the Imperative+Meta variant.

16 out 25 Subjects got lost at least once i.e., they misinterpreted an instruction, followed along wrong path, then they realized inconsistencies with spatial information and the following instruction, and finally recovered from the misinterpreted instruction. A subject lost thrice in the entire experiment who misunderstood one instruction twice and another instruction once. The subject was lost at an intersection of three hallways and only one of them leads towards the destination. This instruction did not have sufficient information about the next heading. All subjects who recovered from misinterpretation informed that landmark's attributes such as number of floors in a building (if building is the landmark) and the spatial orientation of the landmark helped them in recovery.

Instructions that lacked spatial orientation were found to be particularly difficult to follow. Subjects found a few of the imperative and advisory instructions difficult to follow. While following these difficult instructions, people realized that they got lost and asked the experimenter to repeat the instructions. Examples of difficult instructions and the people's complaint on that instruction are as follows:

- *So you kind of cross the atrium* **Complaint:** participants reported that they were not sure how far they had to walk across the atrium.
- *Go beside the handrails till the other end of this building* **Complaint:** no absolute destination, multiple hallways at the end of handrails
- *Just walk down the hallway exit the building* **Complaint:** multiple exits to the building
- *After you get off the elevator, take a left and then left again* **Complaint:** more than one left confused the subjects
- *You can see the building just in front of you* **Complaint:** there were three buildings standing in front and the target building was slightly to the left.
- *You will see the corridor that you want to take* **Complaint:** there were two corridors and the orientation was unspecified in the instruction

5 Understanding Experiments

The Navagati (NAV) corpus instructions were divided into training set (henceforth abbreviated as NAV-train) and testing set (abbreviated as NAV-test) of size 654 (of 6 subjects) and 280 (of 3 subjects). The training set was used to create a grammar based on the taxonomy described in Section 3.

5.1 Grammar

A domain-specific grammar was written to cover most frequent phrases from the training set using the Phoenix (Ward, 1991) format. Phoenix grammars specify a hierarchy of target concepts and is suited to parsing spontaneous speech. The resulting grammar produced correct and complete parses on 78% of the training data (NAV-train). The remaining training instances were not included due to unusual phrasing and disfluencies. The concepts in the grammar are listed in the Table 5.

5.1.1 Managing Variable Vocabulary

Concepts such as Locations, Pathways and Adjectives-of-Location use vocabulary that is specific to an environment, and the vocabulary of these concepts will change

Corpus	#Instr	Words/Instr	Environmnt	Modality	H/R-H/R
NAV	934	9	UnivCampus	Speech	Human-Human
MIT	684	15	UnivCampus	Written	Human-Human
IBL	769	8	ModelCity	Speech	Human-Robot
TTALK	1619	7	OpenSpace	Speech	Human-Robot

Figure 4: (a) Nature of the Corpora

LiftingDevic	PathWays	Landmarks	Adjectives
0.029	0.046	0.169	0.13
0.045	0.016	0.163	0.062
n.a.	0.039	0.076	0.13
n.a.	0.027	0.01	0.039

(b) Type-Token Ratio of Concepts across Corpora

Table 5: Higher level and Leaf node Concepts in Grammar

Category Concepts	**Examples**
Imperative	GoToPlace, Turn, etc
Conditional Imperative	Move_Until_X where X is a condition
Advisory Instructions	You_Will_See_Location
Grounding Instructions	You_are_at_Location
Auxillary Concepts	**Examples**
Locations	buildings, other landmarks on the route
Adjectives-of-Locations	large, open, black, small etc.
Pathways	hallway, corridor, bridge, doors, etc.
LiftingDevice	elevator, staircase, stairwell, etc.
Spatial Relations	behind, above, on right, on left, etc.
Numbers	turn-angles, distance, etc.
Ordinals	first, second as in floor numbers
Filler phrases	you may want to; you are gonna; etc.

with surroundings. We used an off-the-shelf part-of-speech tagger (Toutanova et al., 2003) on NAV-train to identify "location-based" nouns and adjectives. These were added to the grammar as instances of their respective concepts.

5.2 Parsing NAV Instructions

A parse can fall into one of the following categories: 1) *Complete*: clean and correct parse with all concepts and actions mentioned in the instruction. 2) *Incomplete*: If some arguments for an action are missing. 3) *Misparse*: no usable parse produced for an instruction.

Table 6 shows that 87% of the instructions from the NAV corpus (excluding meta comments) are parsed correctly. Correct parses were produced for 89% of Imperatives, 87% of Advisory and 73% of Grounding instructions. Meta comments were excluded because they do not constitute any valid actions and can be ignored. Nevertheless 20% of the meta comments produced a valid parse (i.e. unintended action).

5.3 Grammar Generality

The results for the NAV corpus seem encouraging but it would be useful to know whether the NAV grammar generalizes to other directions scenarios. We selected three corpora to examine this question: MIT (Kollar et al., 2010), IBL[3] (Bugmann et al., 2004) and TTALK[4] (Marge and Rudnicky, 2010). All were navigation scenarios but were collected in a variety of settings (see Figure 4(a)). Corpus vocabularies were normalized using the process described in 5.1.1 and location specific nouns and adjectives added to the grammar. Punctuation was removed. Figure 4(b) shows the type-token ratios for "variable" concepts. There are more landmarks and adjectives (that tag along landmarks) in NAV and MIT compared to IBL and fewest in TTALK corpus (a closed space with two robots). Since, IBL and TTALK do not involve extensive navigation inside the buildings there are no instances of the elevator concept. However, IBL corpus has "exits, roads, streets" in the city environment which were included in the PathWay concept.

5.4 Performance across Corpora

We randomly sampled 300 instructions from each of the three corpora (MIT, IBL and TTALK) and evaluated their parses against manually-created parses. Table 6) shows results for each type of parse (Complete, Incomplete, or Misparse). Meta comments were excluded, as discussed earlier. The NAV grammar appears portable to three other corpora. As shown in Category-Accuracy of Table 6 Imperatives and Advisory instructions are well-parsed by the grammar. In TTALK corpus, there are very few landmark names but there are certain unusual sentences e.g., "she to the rear left hand wall of the room" causing lower accuracy in Advisory instructions. We noticed that MIT corpus had longer description of the landmarks, leading to lower accuracy for Grounding. From Table 6 11% to 16% of Imperative instructions fail to get parsed across the corpora. We consider these failures/errors below.

5.5 Error Analysis

We found six situations that produced incomplete and misparsed instructions: (1) Underspecified arguments; (2) Unusual or unobserved phrases; (2) False-starts and ungrammatical language; (3) Uncovered words; (4) Prolonged description of landmarks within an instruction;

[3]http://www.tech.plym.ac.uk/soc/staff/guidbugm/ibl/readme1.html
[4]http://www.cs.cmu.edu/~robotnavcps/

Table 6: Parse Results

Parse Results	NAV	MIT	IBL	TTALK
# Instructions	280	300	300	300
% Complete	87%	78.8%	83.8%	83.4%
% Incomplete	3.1%	17%	6.6%	3.7%
% Misparse	9.8%	4.1%	9.5%	13%
Category Accuracy				
Imperative	89%	89.4%	86.5%	84.7%
Advisory	87%	93.4%	87.4%	60%
Grounding	73%	62%	100%	100%

(5) Coreferences; 6) Non-specific instructions (eg. either take the right hallway or the left hallway).

5.5.1 Incomplete and Misparsed Instructions

Out-of-Vocabulary (OOV) words were responsible for the majority of incomplete parses across all the corpora; many were singletons. Unusual phrases such as "as if you are doubling back on yourself" caused incomplete parses. We also observed lengthy descriptions in instructions in the MIT corpus, leading to incomplete parses. This corpus was unusual in that it is composed of written, as opposed to spoken, instructions.

Misparsed instructions were caused due to both ungrammatical phrases and OOV words. Ungrammatical instructions contained either missed key content words like verbs or false starts. These instructions did contain meaningful fragments but they did not form a coherent utterance e.g., "onto a roundabout".

We note that incomplete or otherwise non-understandable utterancess can in principle be recovered through clarification dialog (see e.g., (Bohus and Rudnicky, 2005). Direction giving should perhaps not be limited to monologue delivery.

Table 7: Error Analysis for Incomplete and Misparsed instructions

Incomplete	NAV	MIT	IBL	TTALK
# Incomplete Instructions	8	49	19	10
MissingArgs	50%	8%	0%	0%
UnusualPhrases	0%	28%	35%	60%
Lengthy Descriptions	0%	20.4%	0%	0%
Coreferences	0%	0%	20.2%	0%
Non-concrete phrases	3%	2%	5%	0%
OOVs	47%	41.6%	39.8%	40%
Misparse				
# Misparse Instructions	25	12	27	39
Ungrammatical phrases	24%	44%	16%	10%
OOVs	76%	66%	84%	90%

6 Conclusion

To better understand the structure of instructions and to investigate how these might be automatically processed, we collected a corpus of spoken instructions. We found that instructions can be organized in terms of a straighforward two-level taxonomy. We examined the information contents of different components and found that that the Imperative and Advisory categories appear to be the most relevant, though our subjects had little difficulty dealing with instructions composed of only Imperatives; physical context would seem to matter.

We found that it was possible to design a grammar that reasonably covered the information-carrying instructions in a set of instructions. And that a grammar built from our corpus generalized quite well to corpora collected under different circumstances.

Our study suggests that robust instruction-understanding systems can be implemented and, other than the challenge of dealing with location-specific data, can be deployed in different environments. We believe that this study also highlights the importance of dialog-based clarification and the need for strategies that can recognize and capture out-of-vocabulary words. These capabilities are being incorporated into a robot navigation system that can take instructions from humans.

References

G. Allen. 1997. From knowledge to words to wayfinding: Issues in the production and comprehension of route directions. *Spatial Information Theory A Theoretical Basis for GIS*, pages 363–372.

D. Bohus and A.I. Rudnicky. 2005. Sorry, i didn't catch that!-an investigation of non-understanding errors and recovery strategies. In *6th SIGdial Workshop on Discourse and Dialogue*.

G. Bugmann, E. Klein, S. Lauria, and T. Kyriacou. 2004. Corpus-based robotics: A route instruction example. *Intelligent Autonomous Systems 8*.

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

D.L. Chen and R.J. Mooney. 2010. Learning to interpret natural language navigation instructions from observations. *Journal of Artificial Intelligence Research*, 37:397–435.

M.P. Daniel and M. Denis. 1998. Spatial descriptions as navigational aids: A cognitive analysis of route directions. *Kognitionswissenschaft*, 7(1):45–52.

M. Denis, F. Pazzaglia, C. Cornoldi, and L. Bertolo. 1999. Spatial discourse and navigation: An analysis of route directions in the city of venice. *Applied Cognitive Psychology*, 13(2):145–174.

K. Eberhard, H. Nicholson, S. Kubler, S. Gundersen, and M. Scheutz. 2010. The indiana .cooperative remote search task.(crest) corpus. In *Proc. of LREC*, volume 10.

A. Gargett, K. Garoufi, A. Koller, and K. Striegnitz. 2010. The give-2 corpus of giving instructions in virtual environments. In *Proc. of LREC*.

T. Kollar, S. Tellex, D. Roy, and N. Roy. 2010. Toward understanding natural language directions. In *Proceeding of the 5th ACM/IEEE HRI*. ACM.

K. Lovelace, M. Hegarty, and D. Montello. 1999. Elements of good route directions in familiar and unfamiliar environments. *Spatial information theory. Cognitive and computational foundations of geographic information science*, pages 751–751.

M. MacMahon, B. Stankiewicz, and B. Kuipers. 2006. Walk the talk: Connecting language, knowledge, and action in route instructions. *Def*, 2(6):4.

M. Marge and A.I. Rudnicky. 2010. Comparing spoken language route instructions for robots across environment representations. In *SIGDIAL*.

M. Marge, S. Banerjee, and A.I. Rudnicky. 2010. Using the amazon mechanical turk for transcription of spoken language. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 5270–5273. IEEE.

C. Matuszek, D. Fox, and K. Koscher. 2010. Following directions using statistical machine translation. In *Proceeding of the 5th ACM/IEEE international conference on Human-robot interaction*, pages 251–258. ACM.

R. Rosenfield. 1995. The cmu statistical language modeling toolkit and its use in the 1994 arpa csr evaluation.

L. Stoia, D.M. Shockley, D.K. Byron, and E. Fosler-Lussier. 2008. Scare: A situated corpus with annotated referring expressions. In *LREC 2008*.

K. Toutanova, D. Klein, C.D. Manning, and Y. Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics.

W. Ward. 1991. Understanding spontaneous speech: the phoenix system. In *ICASSP*. IEEE.