NOUN PHRASE GENERATION FOR SITUATED DIALOGS

DISSERTATION

Presented in Partial Fulfillment of the Requirements for
the Degree Doctor of Philosophy in the
Graduate School of The Ohio State University

By

Laura Stoia, B.S., M.S.

* * * * *

The Ohio State University

2007

Dissertation Committee:

Donna K. Byron, Adviser
Eric Fosler-Lussier
Michael White

Approved by

[Signature]

Adviser
Graduate Program in
Computer Science and Engineering
Modern dialog systems present significant challenges for developing algorithms that can link appropriate linguistic behavior with information from the world surrounding the user, like the person’s position and orientation, or information about the visibility of objects in the world.

Some examples of the challenges that need to be overcome by these systems are:
(1) referring expressions: anchoring the descriptions in the world and using appropriate noun phrase forms to describe the objects while taking into account both dialog history and spatial relationships with the user
(2) synchronization problems: when is the system’s turn to speak and how can the system take advantage of the user’s possibility of movement to facilitate descriptions
(3) evaluation: algorithms for situated dialog systems are by nature very hard to evaluate due to the necessity to either have a complete system running or re-create the entire visual, spatial and dialog properties of the context to be presented for evaluation by human judges.

Existing algorithms for deciding the form of a noun phrase employed in a particular context are typically designed to operate on the discourse context, with little influence from other factors in the world. There has been research on creating the first mention of an object as a description that utilizes spatial relations. However, there are only a few studies that deal with subsequent references or other phenomena
like deixis or ambiguous descriptions made felicitous by the context (spatial, task or discourse related) and how the situated aspect of the environment influences the production/synchronization of the referring expressions.

This thesis is focusing on integrating information from the perceived world into the generation of noun phrases for a dialog system and dealing with the mobility/change of visual context of the user in producing appropriate speech. We are interested in the different ways to employ particular noun phrase forms to signal proximal marking, deixis or other more subtle factors, such as politeness or obligation to complete the current task. The thesis is proposing to study the contextual factors that favor the production of a description and the particular form chosen. It will further analyze synchronization problems and the interaction between producing a description or taking advantage of the user’s mobility. The findings will be incorporated in an on-line generation algorithm for referring expressions to objects in the perceived world. The final goal of this thesis is to propose an on-line approach to the generation of referring expressions that takes into account the user’s spatial relation to the world and takes advantage of his mobility. We will study the conditions that enable the usage of particular proximal markings (*this* vs. *that*) in a corpus of spontaneous dialogs and we will present findings that apply to noun phrase generation algorithms.

Finally, we will evaluate the algorithms developed by collecting human rankings of the output of an offline generation algorithm and a number of different rankings like naturalness, ease of understanding, friendliness, etc. for an implemented real-time generation system.
To my family ...
I would like to thank my adviser, prof. Donna K. Byron, for providing support and guidance during my Ph.D years. She has always been very patient with my struggles and contributed to my professional and personal growth.

I would also like to thank prof. Eric Fosler-Lussier, my co-adviser, and prof. Michael White which provided valuable feedback on my research and thesis writing. I have been very lucky to find such great people at The Ohio State University.

I would also like to thank the CL discussion group members, who always survived my presentations and gave me useful input.

Thanks to Mariana Sharp, my friend and fellow graduate student, who has been there for me all the way through our many years of grad school. Also, thanks to Adela Lechintan, for taking me swimming when I needed a break. My friends Tianfang Xu, Darla Shockley and Crystal Nakatsu also deserve my thanks for making my graduate school years much more fun.

I am grateful to my sister and my parents for being there for me and offering me their encouragement.
VITA

September 2002 – present ....................... Graduate Research Assistant/Teaching Associate, The Ohio State University
April 2007 – June 2007 ......................... Research Intern, Robert Bosch Research and Technology Center, Palo Alto, CA
March 2007 ....................................... M.S. Computer Science & Engineering, The Ohio State University
June 2005 – September 2005 ................. Research Intern, Northwestern University, Evanston, IL
June 2002 ........................................ B.S. Computer Science, Politehnica University of Bucharest
March 7, 1979 ................................. Born – Bucharest, Romania

PUBLICATIONS

Research Publications


FIELDS OF STUDY

Major Field: Computer Science and Engineering

Studies in:

Artificial Intelligence Prof. Donna K. Byron
Prof. Eric Fosler-Lussier
Prof. Michael White
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td>Dedication</td>
<td>iv</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>v</td>
</tr>
<tr>
<td>Vita</td>
<td>vi</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xi</td>
</tr>
<tr>
<td>List of Tables</td>
<td>xiv</td>
</tr>
<tr>
<td>Chapters:</td>
<td></td>
</tr>
<tr>
<td>1. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Thesis Organization</td>
<td>3</td>
</tr>
<tr>
<td>1.2 Contributions</td>
<td>3</td>
</tr>
<tr>
<td>2. Background and Motivation</td>
<td>7</td>
</tr>
<tr>
<td>2.1 Generation in Dialog Systems</td>
<td>7</td>
</tr>
<tr>
<td>2.2 Generation of Referring Expressions (GRE)</td>
<td>9</td>
</tr>
<tr>
<td>2.3 Previous Work on GRE for Situated Environments</td>
<td>14</td>
</tr>
<tr>
<td>2.4 Challenges in Producing Situated REs</td>
<td>19</td>
</tr>
<tr>
<td>2.5 Challenges for GRE Evaluation</td>
<td>22</td>
</tr>
<tr>
<td>2.6 GRE in Our Work</td>
<td>24</td>
</tr>
</tbody>
</table>
3. The Quake Situated Instruction Giving Corpus ........................................... 25
   3.1 Why Data Collection was Necessary .................................................... 25
   3.2 Description of Data Collection Procedure: Pilot Study in Human-Human Direction Giving ................................................................. 27
   3.3 Data Preparation: Transcriptions and Annotations .............................. 30
   3.4 Referring Expressions: Corpus Distribution ....................................... 31

4. Modeling GRE from human data ................................................................. 33
   4.1 Planning RE in a Dynamically Changing Context ............................... 33
      4.1.1 Deciding When a Description Should be Produced: Resulting Algorithm and Implications for GRE ............................................................. 35
      4.1.2 Feature Interaction ........................................................................ 40
      4.1.3 Discussion and Implications for GRE ............................................ 42
   4.2 Statistical Approaches to GRE ............................................................... 43
      4.2.1 Modeling GRE as a Statistical Problem ....................................... 43
      4.2.2 Deriving a Decision Tree model .................................................... 46
      4.2.3 Analyzing the Resulting Algorithm: A Human Based Evaluation .......... 48
      4.2.4 Deriving a Maximum Entropy Model ............................................ 54

5. A Study of Demonstratives and Proximity Marking .................................. 56
   5.1 Previous Research on Proximity Marking .......................................... 57
   5.2 New Factors that Determine Proximity Marking in Instruction Giving Dialogs ................................................................. 61

6. Human-Computer Interaction: An Implemented Instruction Giving System 65
   6.1 Human-Computer Interaction in Virtual Environments ....................... 65
   6.2 The GRE Algorithm Employed in the Basic System ........................... 69
      6.2.1 Salience ......................................................................................... 72
      6.2.2 Frames of Reference ...................................................................... 73
      6.2.3 Landmarks .................................................................................... 74
      6.2.4 Distractors .................................................................................... 79
      6.2.5 Spatial Relations .......................................................................... 82
      6.2.6 Realization .................................................................................... 87
   6.3 Comparison between Corpus Distributions and the GRE Algorithm Output .................................................................................. 89
      6.3.1 Corpus Properties of **First Locate** Expressions .......................... 90
      6.3.2 Comparison ................................................................................... 93
6.4 Grounding ......................................................... 102
  6.4.1 Tracking User Position and Interpreting Actions .......... 103
  6.4.2 Timing Issues in a Situated Natural Language System ... 103

7. An Experimental Study in Human-Computer Interaction ........ 106
  7.1 Testing the System’s Referring Behavior in Relation to the Context:
    Study Design .................................................. 106
  7.2 Study Materials .............................................. 109
  7.3 Results Obtained from Human Ratings and System Logs ...... 110
    7.3.1 Qualitative Metrics: User Survey Results ............... 110
    7.3.2 Quantitative Metrics: Obtained Automatically from the System’s Log ........................................ 113
  7.4 Discussion .................................................... 118

8. Future work ...................................................... 119

9. Conclusions ...................................................... 124

Appendices:

A. Data Collection Materials ....................................... 127

B. Materials Used in the Human Evaluation ....................... 134

C. Human-Computer Interaction Study Materials ................. 135

D. Example Transcript of a Human-Computer Interaction ....... 141

E. All Possible Turns of the Interaction System ................ 143

F. Example Transcript from the Human Corpus .................. 146

Bibliography .......................................................... 154
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 A dialog system’s architecture</td>
<td>7</td>
</tr>
<tr>
<td>2.2 A simple micro-planner architecture [Reiter and Dale, 2000, page 123]</td>
<td>8</td>
</tr>
<tr>
<td>3.1 Example objects from the experiment world</td>
<td>28</td>
</tr>
<tr>
<td>3.2 Sample dialog fragment and accompanying video frame (Session 4, 28 min 5 sec)</td>
<td>29</td>
</tr>
<tr>
<td>4.1 An example sequence with adjustment of context (Session 13, 13 min 16 sec)</td>
<td>34</td>
</tr>
<tr>
<td>4.2 Sample dialog fragment (Session 6, 7 min 29 sec)</td>
<td>35</td>
</tr>
<tr>
<td>4.3 The micro-planner architecture observed in corpus</td>
<td>36</td>
</tr>
<tr>
<td>4.4 An example configuration with spatial context features identified (the target object of the referring expression is B4)</td>
<td>37</td>
</tr>
<tr>
<td>4.5 The decision tree obtained</td>
<td>39</td>
</tr>
<tr>
<td>4.6 NP frame slot values and examples</td>
<td>46</td>
</tr>
<tr>
<td>4.7 The Anvil software tool used for judging</td>
<td>49</td>
</tr>
<tr>
<td>4.8 A closeup of the selection tool</td>
<td>49</td>
</tr>
<tr>
<td>5.1 Proximity marking compared to absolute distance</td>
<td>61</td>
</tr>
<tr>
<td>5.2 Proximity marking compared to the time of the event</td>
<td>61</td>
</tr>
</tbody>
</table>
5.3 Proximity marking compared to partner completing action . . . . . . . . 62

6.1 System’s physical setting ....................................................... 66

6.2 System’s interaction loop ....................................................... 67

6.3 The Locative Incremental Algorithm from [Kelleher and Kruijff, 2006] 72

6.4 Overhead view of a room in our VR world. ............................. 77

6.5 Example of felicitous mention of an object that is not visible: the cabinet on the right ....................................................... 80

6.6 Enlarged acceptance areas from [Moratz and Tenbrink, 2006] ............................. 84

6.7 Prototypical directions from [Moratz and Tenbrink, 2006] ............................. 84

6.8 The expression generated using the angle to the center of the bounding box will be the door on your left which does not correspond to our intuition ....................................................... 86

6.9 The expression generated using the angle to the center of the bounding box will be the door in front of you which does not correspond to our intuition ....................................................... 86

6.10 Examples from the corpus where people changed or waited to produce the spatial description ....................................................... 105

7.1 The referring strategy adopted by the system in the Immediate Description condition ....................................................... 108

7.2 The referring strategy adopted by the system in the Adjusted Context condition ....................................................... 108

8.1 Examples from the corpus for group lexicalization ............................. 120

8.2 Examples from the corpus where the property needs temporal indexing 121

A.1 The instructions received by the Direction Follower ............................. 127

xii
A.2 The instructions received by the Direction Giver .................. 128
A.3 The Quake controls received by all participants (in data collection and in the human-computer interaction study) .................. 129
A.4 The task descriptions received by the Direction Giver ............. 130
A.5 The descriptions of buttons functionality and contents of cabinets received by the Direction Giver .......................... 131
A.6 The map of the world received by the Direction Giver ............. 132
A.7 The survey at the end of the data collection ......................... 133
B.1 The instructions received by the human evaluators ................. 134
C.1 The instructions received by the human participants .............. 136
C.2 Information received by the human participants: possible tasks ... 137
C.3 System related survey questions received by the human participants (page 1) .................................................. 138
C.4 System related survey questions received by the human participants (page 2) .................................................. 139
C.5 General survey questions received by the human participants (received at the end) ............................................. 140
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>General survey questions and average ratings</td>
<td>30</td>
</tr>
<tr>
<td>3.2</td>
<td>Distribution of <strong>Determiner</strong> and <strong>Head</strong> values in the corpus</td>
<td>32</td>
</tr>
<tr>
<td>4.1</td>
<td>Detailed performance of the decision tree</td>
<td>38</td>
</tr>
<tr>
<td>4.2</td>
<td>Detailed performance of the maximum entropy model</td>
<td>42</td>
</tr>
<tr>
<td>4.3</td>
<td>The maximum entropy model feature’s weights</td>
<td>42</td>
</tr>
<tr>
<td>4.4</td>
<td>The context features used by the algorithm</td>
<td>44</td>
</tr>
<tr>
<td>4.5</td>
<td>Testing independence of the slot values</td>
<td>47</td>
</tr>
<tr>
<td>4.6</td>
<td>Results of human judging</td>
<td>50</td>
</tr>
<tr>
<td>6.1</td>
<td>The cognitive load associated with object properties [Kelleher and Kruijff, 2006]</td>
<td>82</td>
</tr>
<tr>
<td>6.2</td>
<td>The spatial relationships/landmarks employed in our system</td>
<td>83</td>
</tr>
<tr>
<td>6.3</td>
<td>Distribution of <strong>Determiner</strong>, <strong>Head</strong> and <strong>Modifier</strong> values in the <strong>First Locate</strong> expressions</td>
<td>90</td>
</tr>
<tr>
<td>6.4</td>
<td>The distribution of spatial/non spatial descriptions</td>
<td>90</td>
</tr>
<tr>
<td>6.5</td>
<td>The modifiers for the non-spatial relations</td>
<td>91</td>
</tr>
<tr>
<td>6.6</td>
<td>The types of spatial relations observed in the corpus</td>
<td>92</td>
</tr>
</tbody>
</table>
6.7 Detailed counts of spatial relations observed in the corpus 92
6.8 The landmarks used in the corpus 93
6.9 Examples of system output judged wrong by exact matching 95
6.10 The correctness values for different features obtained by the GRE algorithm 96
6.11 The content selection performance of the GRE algorithm 100
6.12 The distribution of spatial/non spatial descriptions 101
6.13 The types of spatial relations observed in the corpus 101
6.14 The landmarks used by the GRE algorithm 102
7.1 The order that the tasks were presented to the human subjects 109
7.2 General survey questions 110
7.3 The System related survey questions and the average ratings obtained in the two conditions (significant differences are indicated in bold) 111
7.4 The participant’s choice of system 113
7.5 Average time spent per task 114
7.6 Task completion per task (each task was presented 16 times) 115
7.7 The different types of referents and their counts in the two systems 116
7.8 The average time it took to ground different types of referents in the two systems 117
CHAPTER 1

INTRODUCTION

This thesis presents the collection of a corpus of human dialogs in a situated environment and its use in developing computational models that mimic the human behavior observed in this data. It contains the design, implementation and evaluation of a natural language instruction giving system that is aware of the user’s movement in the world and plans its language accordingly. This thesis reports on an interdisciplinary study that draws on research in linguistics and computer science (including artificial intelligence, machine learning and software engineering), as it presents results obtained in an offline study of human language behavior, results of human evaluations and the architecture and design of an implemented system. We are trying to connect these areas, with the goal of advancing the development of situated human-computer language interaction.

The field of Natural Language Processing (NLP) is concerned with developing applications related to the automatic processing of data encoded in natural language, either textual or spoken. Several examples of subfields of NLP are machine translation, speech recognition, text summarization and dialog systems. For certain applications, the communication that is modeled happens in concrete situations and software agents acting in these domains need to incorporate world attributes in their
algorithms. Understanding communication in real world situations is of vital impor-
tance for the progress of software agents that can be co-located with the user in an
environment.

Intelligent software agents are being developed for a wide variety of domains to aid
humans in performing numerous tasks. Examples of systems where the computer is
helping with a navigation related task include hand-held tourist information portals
[Johnston et al., 2002], campus tour guides [Yang et al., 1999, Long et al., 1996,
Striegnitz et al., 2005], direction-giving avatars for visitors to a building [Cassell et al.,
2002, Chou et al., 2005, Theune et al., 2007], in-car driving direction systems [Dale
et al., 2003, Wahlster et al., 2001], and pedestrian navigation systems [Muller, 2002].
These applications present an exciting and challenging new frontier for dialog agents,
since attributes of the real-world setting must be combined with other contextual
factors for the agent to communicate successfully.

The particular setting for the research presented in this thesis is a scenario in which
the system provides directions incrementally to a mobile user who is following the
instructions as they are produced. This task requires that real-time instructions based
on dynamic local context variables (such as visibility or distance to reference points)
are produced by one participant who monitors his partner’s progress in carrying out
the task. In referring to items in the setting, human speakers produce a wide variety
of noun phrase forms, including descriptions that are headed by a common noun
and employ a definite, indefinite, or demonstrative determiner, one anaphors, and
pronouns such as it, this and that. The goal of our current work is to study this
entire space of variation and the different factors that contribute in deciding upon a
particular form: discourse history, spatial context, visual context, proximity marking, etc. and to validate our algorithms against human behavior.

1.1 Thesis Organization

The thesis is organized as follows: Chapter 2 introduces the problem of noun phrase (NP) generation and previous research in generating and evaluating referring expressions (REs) and incorporating extra-linguistic features in the generation process. The challenges that remain unsolved in generating referring expressions for situated dialogs are also highlighted in this chapter.

Chapter 3 describes the data used in our studies and the process through which it has been collected and annotated. Chapter 4 will present some preliminary results and how to incorporate those in a general algorithm for NP production in situated environments. In this chapter we will describe both the training of machine learning algorithms modeling the human behavior observed in the corpus and a human-based evaluation of the algorithms. Chapter 5 will present our findings related to factors that influence the choice of proximal marking in various noun phrases. We will describe the development of a direction-giving system in Chapter 6 and will present the results of a study in human-computer interaction in Chapter 7.

Chapter 8 will present remaining issues of our studies, and we will draw the final conclusions in Chapter 9.

1.2 Contributions

This thesis asks the question if we are able to develop natural language techniques specific to systems that interact with a user in dynamic physical settings
(and overcome some of the challenges offered by these systems that will be presented in Section 2.4) and shows that corpus derived techniques enable the development of novel algorithms for noun phrase generation that closely mimic human behavior. The specific contributions of this work are as follows:

- A corpus of human-human dialogs in a situated environment was collected. We utilized a virtual world interaction model that allows us to obtain detailed spatial information about the user and other objects in the world. Noun phrases referring to objects in the world were annotated and situated features were extracted for statistical modeling. The resulting corpus contains spatial information synchronized with the collected dialogs. To our knowledge, this corpus is the first one to contain this type of information, and we hope it will become a useful resource for the natural language processing community.

- This corpus was used to inform various algorithms:
  - to develop an algorithm that learns when a RE was produced in the human corpus and what context attributes influenced this decision.
  - to devise a set of spatial, visual and dialog history features and developed an algorithm for deciding which noun phrase form should be used in a particular context.

These studies advance the field of statistical language generation by providing information about the importance of different situated features and the performance of different machine learning techniques in trying to resolve NP generation problems. The results obtained from analyzing the human-produced corpus
also provide interesting insights from a cognitive perspective on the referring process in situated environments.

- We studied proximity marking in a situated configuration by analyzing the corpus distribution of temporal, spatial and pragmatic features in relation to linguistic markers of proximity, and showed that proximity assignment is influenced by factors not identified previously, such as obligation to complete a task. This is the first corpus study in proximity marking that employs spontaneous speech and synchronized positional information to validate its claims.

- We implemented a dialog system that acts as Direction Giver to test GRE in human-computer interaction.

  - We defined the RE behavior by implementing the state of the art incremental algorithm for GRE in visual worlds and integrating it with the output obtained from the virtual reality engine used in our experiments.

  - We defined a new strategy for generating referring expressions in a situated environment, as a dynamic task, where the system can change the user context to facilitate referring to physical objects through a command directed to the user to move in a better context.

- We evaluated experimentally the new GRE strategy in a human-computer study and collected quantitative and qualitative measurements of the different systems’ performances. These results contain information from the human perceived difference between the two systems, but also quantitative information collected by the game engine about the average times it took people to identify the correct referents, the time it took to complete the tasks, the success
rates obtained by each system, etc. This is the first study in natural language generation of referring expressions in situated environments to report this kind of detailed quantitative and qualitative information resulting from the human interaction with the system.
CHAPTER 2

BACKGROUND AND MOTIVATION

2.1 Generation in Dialog Systems

There are many proposed architectures for dialog systems, e.g. [Levin et al., 2000, Ferguson and Allen, 1998], but all share some commonalities about the necessary modules. A simple architecture of a dialog system is presented in Figure 2.1. The system receives inputs (speech or typed natural language, but it can also receive other types of information) from the user which it needs to understand, and then decide on its response (this decision is dictated by the Dialog Manager). The response is translated from the system specific format into an output that is presented to the user (text or speech, but there are also systems that present images coordinated with speech, or other outputs) by the Generation component.

Figure 2.1: A dialog system’s architecture
The Generation component of a dialog system is the part that translates the internal representation of the system (for example, a logical form) into textual or spoken output. The Referring Expression (RE) generation module has to decide on the content to be expressed in describing an entity (e.g. what properties should be included in the expression, if a pronoun should be used, what type of article is appropriate, etc). In a dialog system, we have identifiers (IDs) that represent concrete objects in the world, and the job of the GRE module is to produce a felicitous string of words (a noun phrase, NP) that uniquely refers to that entity given the current context of the dialog.

![Figure 2.2: A simple micro-planner architecture](http://example.com/image.png)

**Figure 2.2:** A simple micro-planner architecture [Reiter and Dale, 2000, page 123]

In their book that describes the generation process, Reiter and Dale [Reiter and Dale, 2000] identify multiple stages in the process of generating a dialog turn (see Figure 2.2\(^1\)). The system receives some input message that it needs to transform

\(^1\)this stage is called micro-planning and it does not plan the content of the system’s turn, but it deals with transforming a given input message into phrase specifications
into a phrase. The input message is a language independent representation that is lexicalized (by using the appropriate words to express the concepts), and then sent to the Aggregation module. This module decides how much content should be packed into each utterance to be produced by the system. The Referring Expression Generation module substitutes any system internal representations of entities with referring expressions. The output from this phase is sent to a realization component that ensures that the resulting phrase is grammatically and morphologically correct and may also add prosody labels if the output will be spoken.

Another approach to generation, presented in [Stone and Doran, 1997, Stone and Webber, 1998], integrates the generation of referring expression into the larger problem of generating sentences as descriptions of objects, actions and events, by treating both syntax and semantics simultaneously. While this technique ensures textual economy and offers the possibility of producing efficient description by taking advantage of semantic contributions and requirements of the output, it necessitates a rich modeling of the world and there are speed concerns about integrating it in a real time dialog system.

2.2 Generation of Referring Expressions (GRE)

This section describes background in GRE. The problem of GRE is generating a distinguishing expression, e.g. an expression that describes the target object unambiguously from the set of other possible objects that are available for reference, called distractors. The act of referring has been long studied in computational linguistics. One of the first definitions of referring which we will adopt for our research appears in [Appelt and Kronfeld, 1987]:
An agent is referring when he has a mental representation of what he believes to be a particular object, and he intends the hearer to come to have a mental representation of the same object, at least in part through the use of a noun phrase that is intended to be a linguistic representation of that object.

Appelt and Kronfeld have identified two perspectives of referring, which have been treated separately in the literature: *internal* vs. *external*. The internal perspective determines constraints under which a pronoun or a definite description may be anaphorically linked to another NP and it concentrates on subsequent mentions of an entity during the dialog. The external perspective determines the relations between a particular object and a NP and it concentrates on identifying the concrete object through the words used (usually a first mention of the object). Appelt and Kronfeld propose a formal model in terms of goals and beliefs, but do not describe the actual form of the NP to be employed or the way perception and literal content should be combined.

Algorithms have been developed for the two perspectives of GRE, with much more focus on the problem of determining an expression that is uniquely identifying and can be used as a first mention. The problem has been formalized as deciding what properties should be mentioned in the expression, what is the ordering in which properties should be added, and what is the computational cost of the process. Among the most popular techniques for GRE are the *Full Brevity Algorithm*, described in [Dale, 1989], and the *Incremental Algorithm* [Reiter and Dale, 1992, Dale, 1992]. The *Full Brevity Algorithm* constructs a description by selecting the minimal number of attributes to be added to the object description to make it distinguishable from the rest of the objects present in the world, while the *Incremental Algorithm* constructs a description by incrementally selecting attributes from an ordered list to be added to
the description to make it distinguishable. These algorithms have been adopted by many researchers and many extensions have been proposed. For example Krahmer and Theune proposed the use of salience information in ordering the properties to be tested for inclusion in the final description [Krahmer and Theune, 2002]; van Deemter extended the technique to expressions that involve boolean relations [van Deemter, 2002], and Krahmer et al. introduced a graph based representation of the GRE problem and an algorithm which improves efficiency and allows the use of cost functions which can be set to model one of the well accepted algorithms, such as the Full Brevity or Incremental, but also has the advantage of ease of integration with statistical approaches [Krahmer et al., 2003].

Most of these algorithms have not been tested against human spontaneously produced expressions and the influence of the situated world attributes or task at hand has not been formalized. In some recent studies [Gupta and Stent, 2005, Viethen and Dale, 2006] researchers have tried to validate these algorithms against human language production.

As opposed to generating a first mention, for the problem of generating subsequent referring expressions there has not been extensive research. The same algorithms designed for determining a reference that identifies an object might not apply for generating subsequent mentions that are anaphorically linked to other NPs. These algorithms do not usually define constraints under which pronouns or definite descriptions should be used, but only form the expression used to initially identify the referent. One of the early proposals for a unified algorithm for first and subsequent references was made in [McDonald, 1978]. McDonald’s approach is based on syntactic
and rhetorical constraints, which are translated in a series of conditions for determining the use of a pronoun or other NP form. For each of the entities in the discourse list, the features used to decide upon pronominalization are syntactic information (such as clause-index or depth, syntactic category), information about the strategies used to refer to the entity before and information about the distracting referents. A set of derived, descriptive features that characterize the relationship of an instance of a RE to the previous one are used as input to the heuristics. Still, this proposal lacks corpus validation and integration with dialog or situated features (an oracle is used to determine which elements are in the list of competing targets, and simulates the determination of focus). Syntactic information for spontaneous speech data is not easy to obtain automatically and the results are not very reliable.

In a system focused on the referring process in a situated context [Kelleher et al., 2005], the authors identify three possible types of reference: evoking, exphoric and anaphoric. We also encounter these in our corpus study, as the special exphoric type is specific to situated contexts. An exphoric use of a referring expression refers to an object that has entered the discourse model through perception, and it has not been mentioned before. Also, in [Byron, 2003] the author notes that in a virtual world, “users expect the system to have full perceptual knowledge of any graphical elements produced by it... [therefore] a visual history, analogous to the discourse history, must be accumulated”, so we expect exphoric expressions to be more frequent in our domain, where the users share visual input. In the literature, there is, to our knowledge, no study that concentrates on the process involved in generating exphoric references. The most related research is focused on coordinating pointing gestures

\footnote{a knowledge source that contains the information accumulated in the dialog so far}
with generated expressions in a situated environment [Kranstedt and Wachsmuth, 2005]. In [Kelleher et al., 2005], the authors identify these type of references, but the work is more concentrated on interpreting them.

Choosing the reference form has also been motivated by the notion of cognitive status, which was used for generation in [Gundel et al., 1993]. Grosz and Sidner’s theory of focus [Grosz and Sidner, 1986] was used in [Dale, 1988] to constrain the generation algorithm. Centering theory [Grosz et al., 1995] has been used for interpretation of pronouns, but its use in generation for speech data is marginal, mostly because obtaining a parse automatically for spoken data is difficult. In [Callaway and Lester, 2002] the necessity of Centering for generation algorithms is questioned. The authors present a simpler method based on count measures such as recency, sentential distance and nominal element distance that achieves good results on newspaper text and stories. One of the integrations of centering in generation was made in [Kibble and Power, 2000], but this application was for written text and was not validated through a corpus study.

An important aspect to keep in mind is that most of the previous algorithms for GRE have been applied to written text. Spontaneous dialogs are notorious for being hard to parse. In situated dialogs, the participants exhibit behaviors motivated by many new factors such as: alignment to the partner’s choices [Pickering and Garrod, 2004], politeness [Porayska-Pomsta and Mellish, 2004] or influence of situated factors (such as what objects are visible, what are their actual positions in the world, how is the action performed affecting the reference, etc) [Maass, 1995]. The transcriptions of the spontaneous dialogs contain disfluencies and abandoned turns and are much harder to parse than written text. This makes modeling GRE algorithms from spoken
data difficult, because we do not have access to a parse, and we want the algorithm’s output to be grammatically well formed (no disfluencies or abandoned turns).

In the next subsection, we will concentrate on previous research on integrating the influence from the situated world factors into NL generation algorithms.

2.3 Previous Work on GRE for Situated Environments

Many previous projects [Lauria et al., 2001, Moratz and Tenbrink, 2003, Skubic et al., 2002, inter alia] study interpretation of situated language, most often from the perspective of giving directions to a robot or having a system identify an object in the world. The focus of our work is rather on generating the referring expressions associated with a particular object using features of the context in which the expression was spoken (discourse context, spatial position, the presence of similar items, etc).

Some of the first papers that dealt with the interaction between visual and spatial features and natural language (NL) are related to the VITRA (VIsual TRAnslator) project. [Maass, 1994] presents the system called MOSES, a computational framework for the interaction of visual perception and natural language descriptions. The domain for this model is similar to the domain used in this thesis: incremental route descriptions in a virtual 3D world. The language used in Maass’s study is German, so different patterns for choosing proximity markings and using demonstratives or spatial relations can happen compared to English. The study focuses more on different synchronization aspects of language generation than our work, specifically on timing the speech output with the movement of the hearer when giving directions. The relation between the movement of the user and the form of the description in itself is not studied. In [Maass et al., 1995], the authors use features from the world,
including objects’ color, height, width, and visibility, as well as the users’ direction of travel and distance to objects for generating instructions in a situated task. However, their focus is on selecting landmarks (such as important buildings to be mentioned in a route description, etc.) and making descriptions under time pressure, rather than choosing the particular linguistic form of the expression to be produced. In related work [Maass, 1995], the author proposes a computational model for selecting objects using visual salience determined by an algorithm that relies on Treisman’s feature integration theory [Treisman and Gelade, 1980]. The previous visual model is extended with direction of movement and orientation, and path related intentions are used to constrain the spatial attention area. These models are used for selecting which particular objects should be mentioned during the route description, but how the actual features influence the form of RE is not discussed. In [Blocher and Stopp, 1995], the production of spatial descriptions is addressed in the context of developing an anytime algorithm (an algorithm that can provide a good result under time constraints). The descriptions of spatial locations are judged with respect to the degree of applicability, uniqueness and ease of memorization (the spatial relationships are interpreted using potential fields with a model presented in [Gapp, 1994b, Gapp, 1994a]). The set of possible reference objects is constructed taking into account distance (objects closer to the located objects are considered first), salience (objects more remarkable in color, size or shape are considered first) and linguistic context (objects that have already been mentioned are considered first).

An interesting aspect in this research is the introduction of underspecific definite descriptions (UDDs), which are expressions that are ambiguous but can be understood within the context. Redundant information is discarded using a model of the
visual focus of the listener. Even though the domain and problems addressed in the research related to VITRA are relevant for our own research, there has not been any work on GRE with regard to the integration of extra-linguistic factors in the choice of a particular form of reference (should the referent be expressed through a noun phrase headed by a demonstrative, is a pronoun suitable, what proximity marking should be used, etc.) and the focus is on constructing a spatial description that identifies the target referent.

Interesting results in incorporating visual information in a description task are reported in [Roy, 2002]. The problem here is describing a particular target from a visual scene of computer generated rectangles, using a set of visual attributes including shape, size, location, color and brightness. The statistical modeling of the problem is interesting and relevant to our work, as the author use features derived from the visual scene such as shape, size, location, color and brightness of objects, but the models are designed for static scenes where no dialog is taking place.

Definite description usage in the presence of shared visual information is an important component of GRE for situated environments. The research conducted presented in [Poesio, 1993] analyzes the influence of visual attention on REs using a corpus of human to human conversations (part of TRAINS corpus [Allen and Schubert, 1991]) with a focus on the interpretation rather than generation of noun phrases. Poesio proposes the use of a situation theoretic logic called Episodic Logic to formalize the interpretation of definite descriptions. In this framework, the agents organize events into threads, which are situations, and anaphora resolution takes place in these conversational threads. The author proposes an algorithm that determines shifts in visual
attention for this transportation domain (with the most important principle being *follow the movement* : the terminal location of a movement becomes the new mutual focus of attention) and that is capable of resolving definite NPs that are examples of visible situation use (are visible to both speakers and are unique). Talking about the use of definite descriptions in the TRAINS dialogs, Poesio notes that when an object is in the current focus of attention, it can be felicitously referred to by means of a definite description even though objects of the same type have been introduced in the discourse or are part of the world described by the map.

This is relevant for our research, as we are concerned with generating these phrases in a similar way as humans use them, without creating ambiguity by using a redundant expression.

The most related studies to our own are that of Kelleher and Kruijff. In [Kelleher and van Genabith, 2004] the authors describe an algorithm that integrates visual salience with Reiter and Dale’s Incremental algorithm [Reiter and Dale, 1992]. They propose a function to calculate visual salience based on the objects’ position on the screen, color, and dimension. The termination condition is different: terminate when there are no more distractors, or when all the objects in the set of distractors have a visual salience smaller than the object we need to describe. The algorithm only creates descriptions of the form *the [adj] N*. In later work [Kelleher and Kruijff, 2005] the authors devise models for proximity that enable the use of topological terms like *near, at, and close to*. The algorithms use potential fields to model spatial proximity (based on distances) and can take into account different types of attention (visual salience, discourse salience, distance) and also inhibition effects (how the potential fields of different objects interact with each other). The advantage of using potential
fields is that they model the fuzziness of spatial relations and describe the degree of applicability of the spatial relations. The method has a simple model of discourse history (based on recency) and there is no human validation of the behavior and the interaction between different factors in a potential field (all factors are given equal power). The study concentrates on modeling the applicability of topological terms and there is no discussion about how the findings could be applied to proximity marking of NPs.

[Kruijff, 2005] presents an architecture of a situated dialog system similar to the system we use in a human evaluation in Chapter 7. The platform permits dynamic integration in the logical form of information determined by the visual, spatial or discourse context. The study acknowledges the need for a training corpus, especially for determining the interaction between the visual, spatial and discourse features in choosing a particular noun phrase form.

Another recent improvement on the Incremental algorithm is presented in [Kelleher and Kruijff, 2006] which deals more with the sorting of spatially derived properties of a referent in an order motivated by cognitive load and determining the choice of reference objects to be used in the descriptions. The algorithm lacks validation against human behavior and treatment of deixis, proximity marking or other variations in the noun phrase form besides using *it* vs *the +type of object + list of properties*. However, it is one of the few approaches that gives a cognitively plausible motivation for choosing spatial properties in forming REs, and we will employ this algorithm in one of our human studies, in Chapter 6.

In related work by Kelleher, Costello and van Genabith [Kelleher et al., 2005], the LIVE (Linguistic Interaction with Virtual Environments) system is presented,
in which a dialog system is integrating representations from the visual context (and through the visual field, spatial information) and discourse context for the interpretation and generation of REs. The system receives typed input and produces REs that are generated by incorporating salience scores from visual and linguistic input. An important difference between this system and our final goal system for generating REs is that here the user’s movement in the space is restricted to system execution of commands to turn or walk/run forward. The user is not free to move anytime to any degree he wants, and cannot execute a movement while the system is processing a previous turn. Also, the system completes the actions that the user is requesting, and does not issue commands for the user to change position or orientation. These limitations are reasonable in this particular work, because the problems studied are more related to the integration of visual information in the referring process, both for generation and understanding, and not on the dynamics of the context in a instruction giving task.

### 2.4 Challenges in Producing Situated REs

As our summary of the research in the field of generating referring expressions for situated worlds shows, there are still numerous issues to be addressed.

The influence of situated attributes over the way people talk has been acknowledged by many studies in visual attention [Roy, 2002, Byron et al., 2005]. The ongoing action was also acknowledged as a source of influence over language production, and algorithms that integrate planning and collaboration with RE have been developed [Heeman and Hirst, 1995]. Things that happen in the world, such as a door that opened, or a previous action an object participated in (the door that the speaker just
came through), can be used by the dialog participants in building referring expressions. The goal that we have influences what properties of the object are important for us and we include in the expressions. In work presented in this thesis we show how the obligation to perform a task influences the proximal marking on REs [Byron and Stoia, 2005].

One problem that is still unsolved is determining how these factors (discourse, visual and positional information, or task related measures) influence the GRE process (if they do) and determining what model is suitable for representing their interaction in the referring process. There are theories that define discourse salience or visual salience, that advocate the importance of the action being performed or of the particular spatial configuration with regard to reference, but there is not yet a unified approach that describes which information is used at a particular moment and how these models should be combined. This thesis contributes to research in this area by presenting results from a statistical modeling of the referring process including different contextual factors (discourse, visual, spatial).

Another related problem is the fact that multiple forms for a particular RE could be equally good (or the difference could not be reliably perceived by people). Visual focus or spatial configuration alone might license the use of a particular form in the context, regardless of other parameters. Deixis could be used in a particular context, but a full description will work just as well. This problem is very important in generation, as it is more useful to produce a list of ranked alternatives for a RE rather than a unique solution. For example, in [Nakatsu and White, 2006] the authors chose the language that sounds better as produced by the system’s speech synthesizer, so having ranked options from the GRE stage can help in subsequent stages. This
also makes automatic evaluation of any resulting algorithms very difficult. We will contribute to the research in this area by presenting our results in several human evaluations of the output of automatic algorithms.

Another interesting issue that hasn’t received its due attention in the field is the interaction between reference and actions in the world. In a situated domain, the participants can change their position and the state of the world, thus impacting the process of reference. Actions that objects participate in and that were witnessed by both participants become part of the common ground, and can be used in referring to the objects (e.g. the door that just opened); using language to change the field of view of the partner can influence the ease of reference to a particular object (e.g. directing your partner to a position where the object to be describe is in frontal position); and anticipating the direction of movement of the hearer can influence the form produced to describe a target. The human data that we gathered as a step in this thesis provides examples of such behaviors, and hopefully will prove to be a useful resource for the community.

Another important aspect of situated REs are the synchronization problems that can arise in real-time interaction between a person and a generation system. Incorporating changes in the world and positional/angular information requires the system to be able to react quickly to the user’s actions and to predict its partner’s future position. Chapter 6 will describe the choices made for synchronization in our implemented system.

Finally, an important problem encountered by most of the algorithms proposed in the literature is the lack of a corpus of human data to study/validate the results. This problem arises from the difficulty to track the visual and spatial context of the
dialog participants and synchronize it with the production of speech. To address this problem, we are proposing a virtual world interaction model, in which the two dialog partners share the same visual context (they see the same screen image) and where the world simulator is recording the positions and movements of the objects and avatar in the world. Chapter 3 will describe the collection of this corpus and the annotation of NPs in detail. We are hoping that with the release of this corpus, the community will benefit from having a platform to test and inform new algorithms for situated GRE.

2.5 Challenges for GRE Evaluation

The NLG community has not yet agreed on general techniques for evaluating generation systems in the case of text, and with the introduction of dialog the problem becomes much more complex. The importance of evaluation is great, as it allows us to compare performance to other systems, to compare our own improvement, to discover areas where the system needs work, and to know how far we are from a desirable threshold, thus facilitating progress.

Evaluation can be qualitative (intrinsic), often based on human ratings, or quantitative (extrinsic), where some factor that indicates task performance is measured. In [Belz and Reiter, 2006], the authors present the advantages of automatic vs. human evaluation techniques for NLG systems. Having an automatic method to evaluate has been recognize as extremely useful, as it would allow researchers to quickly and cheaply evaluate changes in algorithms. But a good automatic measure for dialog generation has yet to be found. We consider that the problems encountered in written text, e.g. the need of a high quality corpus for automatic measures to correlate
with human judgments or the fact that the initial data can be given a low ranking by other human subjects are becoming more difficult in the case of dialogues. Methods that reward the similarity between outputs and a corpus will place at disadvantage systems that do not always base their choice on the most frequent case observed in the corpus. Researchers have shown that, in some cases, people prefer variation (see [Foster and Oberlander, 2006] for a study on the generation of facial expressions), but corpus-based metrics give higher ratings to the systems that follow the most frequent observed behavior. Also, in [Foster and White, 2007] it is shown that people prefer a system that avoids repetitions and contains more variation in its output, even though the results are less similar to the corpus examples.

Human evaluations are expensive, and not having to use humans in the evaluation process has advantages, as it ensures that the experiments are repeatable and incremental development is easy. Automatic corpus-based evaluation has the advantage of being fast and cheap, so more research in finding automatic metrics that have high correlation with human judgments is needed.

One problematic aspect of evaluating dialog systems is that we do not always have a complete system to test in direct interaction with people. Even in the case that such a system exists, the contribution of each module of the system in the overall performance is hard to isolate and the actual factor that lead to different scores is hard to pin down. This is why evaluation of dialog systems has concentrated on evaluating either the entire system, or components that are more suitable for quantitative measures of performance, such as speech recognition or understanding.

To overcome some of the challenges in evaluating the generation output in this thesis, we will conduct a corpus comparison, a human evaluation of the output of a
statistical GRE algorithm and a human interaction study that produces quantitative and qualitative measures.

2.6 GRE in Our Work

To treat the problem of RE in situated environments, we propose to collect and study a corpus of human interactions, and model the behavior by tracking both discourse and environment specific features. Our plan is to first develop a corpus of the type of interaction we are interested in, where most of the situated variables that could influence the generation process can be tracked. Using this data, we develop machine learning algorithms and study corpus frequencies of different aspects we consider important for the generation of REs. In the third phase, we implement a system that generates situated REs and test its behavior using human evaluators.

During our study of GRE for situated worlds we will touch on issues that have received little study in the literature before, such as proximity marking assigned to different expressions, synchronization problems when dealing with a real-time system, and the influence that a variety of situational features (such as distance, angles, speed, visual presence/persistence) have over the referring process.

One of the central points of this thesis is looking at GRE from a dynamic perspective, where context can be modified and reference timed such that it is produced in contexts more likely to lead to a successful and natural communication.
CHAPTER 3

THE QUAKE SITUATED INSTRUCTION GIVING CORPUS

To support our research in situated noun phrase generation, examples of human produced language in this particular condition are necessary. The language should contain a wealth of referring expressions, produced in spontaneous dialogs by human partners that share the same environment. Both language and world attributes should be tractable in these dialogs, so that models that mimic human behavior can be developed.

3.1 Why Data Collection was Necessary

Even though a wealth of speech data is available for the dialog systems research community, the particular aspects of language generation we are interested in require a type of resource that was not available at the time we started our investigations. The corpus required to answer our research questions should have a way to relate world information with the human language. As of our knowledge, there currently is no such corpus available for research.

One situated language corpus is available [Byron, 2005], but it does not include information to automatically link world attributes with the dialogs. In the new corpus
we collected, we use the same stimuli and task, however the roles of the partners are modified. Only one partner is placed in the world, and the other partner is given full knowledge of the world so he can plan the tasks. This produces a large number of referring expressions and instructional language.

A virtual reality (VR) world, used to create computer games\(^3\) was used to collect the corpus. It provides information about the user’s position and orientation in the virtual world, about all the object locations and timings of any events that take place in the virtual world. Such a corpus of human dialogs and detailed world knowledge will allow us to research different aspects of noun phrases in a situated environment and train a dialog system that can use environment variables in its processing. In the following sections, we will describe the procedure we used to collect the corpus and the properties of the resulting data set.

The VR world was chosen instead of a real-world setting so that complex spatially-extended tasks could be studied without the expense of specialized equipment to obtain detailed information on context variables such as locations and view angles. The VR materials can be reused by other research groups and could be modified to suit different research questions. Virtual worlds are also suitable for distributed web-based data collections.

Although we used a virtual world, humans have been found to be very robust in treating virtual world spatial representations in the same way as real-world objects, even when the graphical depiction in the virtual world is very impoverished [Peruch et al., 2000]. We take the view that spatial language and references to objects in a virtual world maintain most properties when transferred to the real world domain,

\(^3\)http://www.idsoftware.com/games/quake/quake2/

26
and a corpus collected in a virtual environment will prove useful for studying general language behavior.

3.2 Description of Data Collection Procedure: Pilot Study in Human-Human Direction Giving

In order to study the reference process in situated worlds, we developed a domain in which a human is directed by another human partner through an interior space (a graphically-presented 3D virtual world) to perform a sequence of manipulation tasks. We collected and annotated a corpus of dialogs in this domain.

Our task setup is designed to elicit natural, spontaneous situated language examples from human partners. The experimental platform employs a VR world in which one partner, the Direction Follower (DF), moves about to perform a series of tasks, such as pushing buttons to re-arrange objects in the room or finding treasures. Figure 3.1 shows examples of some objects that populate the VR world. The simulated world was presented from first-person perspective on a desktop computer monitor. The DF had no prior knowledge of the world map or tasks and relied on his partner to guide him on completing the tasks.

His partner, the Direction Giver (DG), had a paper 2D map of the world and a list of tasks to complete. The world is a two level maze, with a total of eighteen rooms, two flights of stairs and a long hallway (the map is provided in Appendix A, Figure A.6). The map contains only a small number of object types: buttons, cabinets, doors, tables, and so on. All objects in a given category share the same attributes such as texture and color and are of identical size. The buttons are placed at the same height on walls and at the same distances. This was done to make the objects harder to
describe and to encourage the subjects to rely on the spatial context when referring to objects in the world, and not on other properties.

In Appendix A we show the materials used during the data collection process. Figures A.1 and A.2 present the instructions received by the two participants, and Figures A.4 and A.6 show the map of the world and information about the five tasks received by the DG.

As the two participants performed the study, the DG had instant feedback about the DF’s location in the VR world, by feeding the DF’s computer screen also on the DG’s computer monitor. The partners communicated through headset microphones and were recruited in pairs. The user view angle was set to 100°, as it provided a natural, not deformed view into the world. Figure 3.2 shows an example view of the world and the accompanying dialog fragment. The REs that identify buttons, doors and cabinets are indicated in bold.

The video output of DF’s computer was captured to a camera, along with the audio from both microphones. A log-file created by the VR software recorded the DF’s coordinates, gaze angle, and the position of objects in the world at a frequency of...
DG: you can currently see three buttons... there’s actually a fourth button that’s kind of hidden
DF: yeah
DG: by this cabinet on the right
DF: I know, yeah
DG: ok, um, so what you wanna do is you want to
go in and you’re gonna press one of the buttons
that’s on the right hand wall, so you wanna go
all the way straight into the room and then face
the wall
DF: mhm
DG: there with the two buttons
DF: yep
DG: um and you wanna push the one that’s on the left

**Figure 3.2:** Sample dialog fragment and accompanying video frame (Session 4, 28 min 5 sec)

10 times per second. These data sources were synchronized using calibration markers. A technical report is available that describes the recording equipment and software used [Byron, 2005].

It is important to note that the knowledge shared by the dialog partners in this domain comes from both the dialog they are engaged in, and also their shared view of
the world. The DF’s actions change the state of the world, and his partner is aware of these changes through the visual input.

<table>
<thead>
<tr>
<th>Question</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) How would you rate your expertise as a computer user? (1=very poor - 5=very good)</td>
<td>4.43</td>
</tr>
<tr>
<td>2) What was the difficulty of the tasks you had to complete? (1=not difficult at all - 5=very difficult)</td>
<td>1.87</td>
</tr>
<tr>
<td>3) How did you feel about the difficulty of the navigation in the Virtual World? (only for the follower) (1=not difficult at all - 5=very difficult)</td>
<td>1.71</td>
</tr>
<tr>
<td>4) Were the descriptions of the objects in the world difficult to follow? (were the objects difficult to identify by their provided descriptions?) (1=not difficult at all - 5=very difficult)</td>
<td>1.45</td>
</tr>
</tbody>
</table>

Table 3.1: General survey questions and average ratings

Figure A.7 shows the questionnaire that the human participants completed at the end of the study. Table 3.1 shows the average rankings over the 30 participants obtained for each of the survey questions. All the participants identified themselves as native speakers of North American English, with an average age of 30. The participants were recruited in pairs and usually they were friends, colleagues or members of the same family. There were 19 male and 11 female participants. All the questions used a 1 to 5 rating scale. The participants rated themselves high on computer expertise, they found that the tasks were not very difficult, the virtual world was not hard to navigate in, and the descriptions they heard were not difficult to follow.

3.3 Data Preparation: Transcriptions and Annotations

Using the above-described setup, we created a corpus consisting of 15 dialogs containing a total of 3 hours and 41 minutes of speech. The corpus was transcribed and word-aligned using Praat [Boersma and Weenink, 2001]. SONIC [Pellom and
Hacioglu, 2001] speech recognition software was used to automatically word align the utterances, which were corrected by two human annotators. The dialogs were further annotated using the Anvil software [Kipp, 2004] to identify a set of target REs in the corpus. Because we are interested in the spatial properties of the referents of these target expressions, the items of interest in this experiment were restricted to objects with a defined spatial position (we only used doors, buttons and cabinets).

Each object in the virtual world was assigned a symbolic identifier (ID), and the ID of each target RE was added to the annotation. REs with plural referents were marked as Set, and were labeled with a list of the members in the set. Expressions were also annotated as either Vague when the referent was not clear at the time of utterance or Abandoned in case the utterance was cut short. Items that did not contain a surface realization of the head of the NP (e.g., on the left), were marked with the tag Empty, but the ID was still included.

The data used in the experiments is a consensus version on which both annotators agreed on the set of target expressions and their properties. Due to the constraints introduced by the task, referent annotation achieved almost perfect inter-annotator agreement (the raw agreement was 99.7).

3.4 Referring Expressions: Corpus Distribution

The corpus contains 1736 target expressions, of which 221 were AllVague (vague + abandoned), 45 were Empty, and 228 were Sets. The remaining 1242 form the set of test items in the experiments described below. AllVague items were excluded since we do not wish for the algorithms developed to reproduce this behavior.

I would like to acknowledge Darla Magdalena Shockley for all the help in transcribing and annotating the data.
Table 3.2: Distribution of Determiner and Head values in the corpus

<table>
<thead>
<tr>
<th>Value</th>
<th>Count</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>364</td>
<td>39%</td>
</tr>
<tr>
<td>that/this</td>
<td>264</td>
<td>29%</td>
</tr>
<tr>
<td>none</td>
<td>253</td>
<td>27%</td>
</tr>
<tr>
<td>a</td>
<td>46</td>
<td>5%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Value</th>
<th>Count</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>common noun</td>
<td>558</td>
<td>60%</td>
</tr>
<tr>
<td>one</td>
<td>166</td>
<td>18%</td>
</tr>
<tr>
<td>it</td>
<td>116</td>
<td>13%</td>
</tr>
<tr>
<td>that</td>
<td>57</td>
<td>6%</td>
</tr>
<tr>
<td>none</td>
<td>30</td>
<td>3%</td>
</tr>
</tbody>
</table>

Set items were excluded in order to avoid the more complex calculation of spatial properties associated with plural entities. Table 3.2 presents the distribution of the various determiners and head values in a subset of the annotated expressions (only the Direction-Giver expressions not vague, ambiguous or sets), showing the great variability of our corpus.
CHAPTER 4

MODELING GRE FROM HUMAN DATA

4.1 Planning RE in a Dynamically Changing Context

Previously in the RE literature, the process of referring has been treated statically: given a particular context, what is the appropriate expression to describe an entity. Studying our human-human interaction corpus, we observed that people have a more interesting behavior. The context is not something that is static, and if the partner does not consider it appropriate for reference, it can be changed.

In a classical treatment of referring, algorithms would generate expressions independently of how hard the context is or how complicated the resulting expression would be. But we observed that people take advantage of the mobility of their partner and can modify the context. The Direction Givers in our study told their partners to hold on or stop moving when the speed of the follower was making planning of an expression too complicated, and also directed their partners to the center of the room or into a position where they faced the objects of interest. Utterances that direct attention to a set of objects of interest (e.g.: Two doors here, There are two buttons on the right hand wall, Face the buttons) are frequent in the corpus. An example of the observed behavior is presented in Figure 4.1.
Figure 4.1: An example sequence with adjustment of context (Session 13, 13 min 16 sec)

The Direction Givers in our study produced frequent and linguistically varied messages to reposition their partners. Some examples are: “turn around”, “walk forward”, “turn to your right/left”, “look right/left”, “look at the buttons/cabinets”, “facing the buttons”, “face the left door”, “rotate ninety degrees to the left”, “if your back is to the stairs”. In Figure 4.2, we show another sample dialog fragment where the Direction Giver is driving his partner to retrieve an object from a cabinet ([Cabinet9], in this case, which is in the downstairs level of the map). One possible route involves the following five targets: [Door9] - [Door11] - [Button18] - [Door10] - [Cabinet6]. The DG directs the Follower to each of these objects in turn, and he also produces three additional repositioning commands (in bold), to adjust the context before producing descriptions.

In the following subsection, we will present an attempt to statistically model this behavior in our corpus, by constructing a model that determines when a context was deemed appropriate for referring by the human direction givers, and when they waited or moved their partner towards another context.
DG: ok, yeah, go through *that door* [Door9, first locate]

**turn to your right**

'mkay, and there's *a door* [Door11, vague]

*in there* um, go through *the one*

*straight in front of you* [Door11, first locate]

ok, stop... and then **turn around and look at the buttons** [Button18, Button20, Button21]

ok, you wanna push *the button that’s there on the left by the door* [Button18]

ok, and then go through *the door* [Door10]

**look to your left**

there, in *that cabinet there* [Cabinet6, first locate]

---

**Figure 4.2:** Sample dialog fragment (Session 6, 7 min 29 sec)

**4.1.1 Deciding When a Description Should be Produced: Resulting Algorithm and Implications for GRE**

The problem addressed in [Stoia et al., 2006a] is the process through which the content planner can decide whether the DF’s current location and orientation is appropriate for producing a RE to describe a target object, using some spatial features derived from the DF’s location or to wait for a more favorable context. Compared to the classical architecture of GRE, the behavior we observed is presented in Figure 4.3. We are not committed to placing the computation of the context immediately before the GRE module. It must take place before the GRE stage, and it is possible to be done before the aggregation step, as we encounter examples where a referring expression is produce with the help on multiple turns (for example, in a first turn a group of buttons is brought into focus, and in the next group the particular target is described).
In this study, the focus is to determine the configurations of spatial context variables that allow a target object to be described rather than a movement command to be issued. Therefore, we annotated each referring expression with a boolean feature called First Locate that indicates whether the expression is the first referring expression used for that object that allowed the follower to identify it in the world, in other words, the point at which joint spatial reference is achieved. The $\kappa$ value [Carletta, 1996] obtained on this feature was 0.93. There were 466 referring expressions in the 15-dialog corpus that were annotated TRUE on this dimension.

Figure 4.2 shows several RE that were a First Locate.

The following features were calculated automatically from the VR output: absolute Angle between target and the center of the follower’s view direction, Distance from target, visible distractors (VisDistracts), visible distractors of the same semantic category (VisSemDistracts), whether the target is visible (boolean Visible),
and the target’s semantic category (Cat: button, door or cabinet). Figure 4.4 is an example spatial configuration with these features identified.

![Diagram of spatial context features](image)

$v = \text{Visible area}(100^\circ) \\
\alpha = \text{Angle to target} \\
d = \text{distance to target} \\
\text{In this scene:} \\
\text{VisDistr} = 3 \{B2, B3, C1\} \\
\text{VisSemDistr} = 2 \{B2, B3\} \\
\text{Perceptually accessible:} \\
\{B1, B2, B3, B4, C1, D1\}

**Figure 4.4:** An example configuration with spatial context features identified (the target object of the referring expression is B4)

Training examples from the annotation data are tuples containing the ID of the annotated description, the spatial features of the DF at that moment (from the VR engine log), and a class label: either Positive or Negative. Because we expect some latency between when the DG judges that a felicity condition is met and when he begins to speak, rather than using spatial context features that co-occur with the onset of each description, we averaged the values over a 0.3 second window centered at the onset of the expression.

Negative contexts are difficult to identify since they often do not manifest linguistically: the DG may say nothing and allow the user to continue moving along his current vector, or he may issue a command to move his partner in a different
position or to change orientation. A minimal criterion for producing an expression that can achieve joint spatial reference is that the addressee must have perceptual accessibility to the item (either through vision, or just because it is proximally located). Therefore, negative training examples for this experiment were selected from the time-periods that elapsed between the follower achieving perceptual access to the object (coming into the same room with the target object but not necessarily looking at it), but before the First Locate description was spoken. In these negative examples, we consider the basic felicity conditions for producing a descriptive reference to the object to be met, yet the DG did not produce a description. The dataset of 932 training examples\(^5\) was balanced to contain 50% positive and 50% negative examples.

The Weka\(^6\) toolkit was used to build a decision tree classifier [Witten and Frank, 2000]. Figure 4.5 shows the resulting tree. 20% of the examples were held out as test items, and 80% were used for training with 10-fold cross validation. Marking all cases where the referent was visible as describe-id and all the other examples as delay gives a baseline of 70%, still 16% lower than the result of our tree, difference statistically significant (p=0.0001).\(^7\)

<table>
<thead>
<tr>
<th>Class</th>
<th>TruePositives</th>
<th>FalseNegatives</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>describe-id</td>
<td>74</td>
<td>6</td>
<td>0.822</td>
<td>0.925</td>
<td>0.871</td>
</tr>
<tr>
<td>delay</td>
<td>64</td>
<td>16</td>
<td>0.914</td>
<td>0.8</td>
<td>0.853</td>
</tr>
</tbody>
</table>

**Table 4.1:** Detailed performance of the decision tree

\(^5\)only the DG speech was used and the REs were restricted to the first mention that achieved joint spatial reference, labeled with true on the First Locate feature (466 positive examples)

\(^6\)http://www.cs.waikato.ac.nz/ml/weka/

\(^7\)not all positive examples were visible
VisDistracts <= 3
| Angle <= 33
| | Distance <= 154: describe-id (308/27)
| | Distance > 154: delay (60/20)
| Angle > 33
| | Distance <= 90
| | | Angle <= 83: describe-id (79/20)
| | | Angle > 83: delay (53/9)
| | Distance > 90: delay (158/16)
VisDistracts > 3: delay (114/1)

**Figure 4.5:** The decision tree obtained

Previous findings in spatial cognition [Gapp, 1995] consider angle, distance and shape as the key factors in establishing spatial relationships. The angular deviation from the straight direction is considered the most important feature for projective spatial relations (e.g. *in front of*, *behind*, *right*, *left*, *above*, *below*, etc). The decision tree also selects **Angle** and **Distance** as informative features. **VisDistracts** is selected as the most important feature by the tree, suggesting that having a large number of objects to identify the referent from makes the description harder, which is in agreement with human intuition. In terms of the referring expression generation algorithm described in [Reiter and Dale, 1992], in which the description which eliminates the most distractors is selected, our results suggest that the human subjects chose to reduce the size of the distractor set before producing a description, presumably in order to reduce the computational load required to calculate the optimal description or to facilitate the partner’s understanding of the referent.

The exact values of the features shown in our decision tree (such as distances and angle values) can be specific to our own environment. However, the features themselves are domain-independent and are relevant for any spatial direction-giving task, and their relative influence over the final decision may transfer to a new domain.
To incorporate our findings in a system, we would monitor the user’s context and plan a description only in the cases when our tree predicts it, and try to delay or turn the user to a better context in the other cases. This strategy has been used in the human interaction experiment we will present in Chapter 7. A GRE algorithm incorporated in an interactive system could also consider if it can produce an unambiguous RE before trying to reposition the user, but this can be time consuming (calculating the expression takes more time than just applying the decision tree), and also might produce difficult descriptions (we do not know if the reason why direction givers turn their partner is because it would be hard to produce an unambiguous description, or because they want to be cooperative and minimize the cognitive load on their partners by giving incremental, simpler descriptions).

The process developed in this study for producing dialog moves that contain descriptions only at appropriate times is relevant for spoken dialog agents operating in other navigation domains, where the user is mobile in the environment.

4.1.2 Feature Interaction

The features that were used for training the decision tree model (Distance between user and target, Angle between user’s center of view and target, number of other objects in the field of view (VisibleDistr), number of other objects of the same type in the field of view (VisibleSemDistr), type of object (Cat) and if the target is Visible or not) are not necessarily independent, an assumption that decision trees make during modeling. They are dependent, as it is likely that approaching a target will decrease the number of other objects included in the field of view. Also, during
movement, both angle and distance to target change simultaneously and not independently (for example, if the person keeps the same direction, as he moves forward the distance decreases and the angle increases). The number of visible distractors of the same type is always included in the number of total visible distractors and the visibility of an object is derived directly from the angle to the object.

With this in mind, we trained a different statistical model that allows correlated features for training. We built a Maximum Entropy model using the Stanford Classifier 2.0\(^8\) on the same training data as the decision tree. We used Quasi-Newton optimization with 15 previous iterations, a tolerance of 0.0001 for convergence and a sigma of 3 for smoothing. Changes in these parameters did not significantly improve the performance. The results over the same test data are presented in Figure 4.2, and the weights that were associated with the important features in Table 4.3. Even though the model permits correlations, the results of the prediction have almost the same precision: 86% for decision tree and 83% for max entropy, a difference not significant at a level of 0.2. The same features that were deemed important by the decision tree have been assigned high weight values, with the exception of VisibleSemDistr.

From these results, we conclude that human processing for deciding if a context is appropriate to produce a referring expression might follow a simple model, as the decision tree, where one feature’s presence can lead to a decision without considering the other feature values. More data is necessary to be able to make a claim in this direction, but the performance of these models is interesting in correlation to human processing.

\(^8\)http://nlp.stanford.edu/downloads/classifier.shtml
<table>
<thead>
<tr>
<th>Class</th>
<th>TruePositives</th>
<th>FalseNegatives</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>describe-id</td>
<td>72</td>
<td>8</td>
<td>0.791</td>
<td>0.900</td>
<td>0.842</td>
</tr>
<tr>
<td>delay</td>
<td>61</td>
<td>19</td>
<td>0.884</td>
<td>0.762</td>
<td>0.819</td>
</tr>
</tbody>
</table>

**Table 4.2:** Detailed performance of the maximum entropy model

<table>
<thead>
<tr>
<th>High Weight Features</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>VisibleDistr</td>
<td>0.2270</td>
</tr>
<tr>
<td>VisibleSemDistr</td>
<td>0.0306</td>
</tr>
<tr>
<td>Angle</td>
<td>0.0286</td>
</tr>
<tr>
<td>Distance</td>
<td>0.0160</td>
</tr>
</tbody>
</table>

**Table 4.3:** The maximum entropy model feature’s weights

### 4.1.3 Discussion and Implications for GRE

In the section above we described an experiment in natural language generation (NLG) content planning for spoken dialog agents that provide navigation instructions. Navigation instructions require the system and the user to achieve joint reference to objects in the environment. Human subjects achieve this communication task sometimes by describing the referent, and sometimes by steering their partner into a position from which the object is easier to describe. The algorithm we developed in this study replicates our human subject’s decision to produce a description with 86% accuracy. Our decision procedure selected spatial features that were found to be correlated to the referring process in situated environments in the literature. Thus, although the spatial details will vary for other spoken dialog domains, the process developed in this study for selecting a dialog move that is sensitive to the spatial context should be relevant for spoken dialog agents operating in other navigation domains.
The behavior modeled in our algorithm demonstrates an interesting property of the dynamic context that occurs during interactive dialog. A dialog partner need not be only a passive receiver of context variables against which it should produce coherent linguistic units. To achieve communicative goals, he can work to manipulate that context into a configuration that makes potentially ambiguous dialog behaviors, such as referring expressions, more likely to succeed. In [Clark and Bangerter, 2004], the authors define referring as an interactive process, part of a joint action between two dialog participants. Referring is a collaborative process, and the speaker can use different strategies in an opportunistic fashion: offering options, self-interruptions, waiting or instant revisions of their turns.

4.2 Statistical Approaches to GRE

4.2.1 Modeling GRE as a Statistical Problem

In the following subsection we present our first study on producing REs using statistical methods in a situated domain [Stoia et al., 2006b]. The algorithm will provide input to a surface realization component for NP generation, given the ID of a target referent and a vector of context features. It is desirable for these context features to be automatically derived (from the previous turns or from the state of the dialog system), to limit the reliance on human annotation and to be transferable to a real world system. This study was restricted to features that either were derived automatically, or required minimal human annotation.

One consequence of this decision is that even though the linguistic literature predicts that syntactic features such as grammatical role are important in selecting NP
forms, these features were difficult to obtain. Our corpus contains spontaneous spoken dialog, which has no given sentence boundaries. Automatic parsing of spoken data is still a research question, and currently a reliable parse is hard to obtain. With improved parsing techniques, we may include syntactic information derived from a corpus study in the decision process for statistical NP generation in future, but this was not included in the current study.

<table>
<thead>
<tr>
<th>Dialog history features</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Count and chainCount</td>
<td>the mention counts for the referent over the dialog and inside a reference chain</td>
</tr>
<tr>
<td>2. DeltaTime &amp; DeltaTimeChain</td>
<td>the time elapsed since it was last mentioned in the dialog overall or in a chain</td>
</tr>
<tr>
<td>3. PrevSpeaker</td>
<td>the previous speaker that mentioned the ID</td>
</tr>
<tr>
<td>4. Mod(_i)-1, Det(_i)-1, Head(_i)-1</td>
<td>the values of the slots of the NP frame of the prior mention of the same referent</td>
</tr>
<tr>
<td>5. Mod(_i)-2, Det(_i)-2, Head(_i)-2</td>
<td>the previous-1 values of the slots</td>
</tr>
<tr>
<td>6. WordDistance and chainWordDistance</td>
<td>the number of words spoken by both speakers since overall or in the chain the last mention of the ID</td>
</tr>
<tr>
<td>7. Type(_i)-1</td>
<td>indicates if the previous mention was in a Set, was Vague, or was a test item</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Spatial/Visual features</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>8. Distance</td>
<td>the distance between the referent and the DF’s VR coordinates</td>
</tr>
<tr>
<td>9. Angle</td>
<td>the angle between the center of the DF’s view angle and the center of the referent bounding box</td>
</tr>
<tr>
<td>10. Visible</td>
<td>a boolean value which indicates if the object is visible</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Relation to other objects in the world</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>11. Visible Distractors</td>
<td>the number of other objects besides the target referent in the field of view</td>
</tr>
<tr>
<td>12. SameCatVisDistractors</td>
<td>the number of visible distractors of the same type as the referent</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Object category and its information status</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>13. Cat</td>
<td>the semantic category of the referent</td>
</tr>
<tr>
<td>14. First Locate</td>
<td>indicates if this is the first expression that allowed the DF to identify the object in the world</td>
</tr>
</tbody>
</table>

\(^a\)note that an Angle value smaller than 50° ensures the object is Visible

Table 4.4: The context features used by the algorithm
There is an obvious mismatch between features that can be annotated or obtained automatically from a corpus and the information that a Dialog System has during generation. Also, there is a mismatch between the type of language we want to generate and the corpus we are modeling. The corpus has disfluencies and the grammar rules are much more loose, while the system’s turns are full sentences that are expected to be grammatically correct. People have a very robust way of using language, and corrections, vague expressions and abandoned utterances are not that disturbing, although they would be if produced by a computer system, where the expectations are different.

Following [Poesio et al., 1999], we consider the information conveyed by an NP to be divided into four slots which must be filled to be able to generate the NP form: a determiner/quantifier, a pre or post-modifier and a head noun slot. There were few examples of premodifiers in the corpus, so we collapsed the modifier feature. Therefore, the output from our algorithm is an NP frame specifying values for the three slots for each target expression. Figure 4.6 shows the possible values in each slot and example slot values for two NPs.

To form the training dataset, we processed each target expression with a syntactic chunker.9 The partial parse it produced was further processed with a regular-expression matcher to isolate the values corresponding to the three slots. Parser errors caused some low-count NP frame values, so we retained only items that occurred at least 10 times in the entire corpus. Any parser errors that remained in the data were not hand corrected (because we wanted to keep the human intervention minimal).

9http://www.ltg.ed.ac.uk/software/chunk/index.html
The possible values of each NP frame slot

\[
\begin{array}{c|c|c|c|c|}
\text{det} & \text{head} & \text{mod} \\
\hline
\text{none} & \text{it} & - \\
\text{that} & \text{noun} & + \\
\end{array}
\]

NP frames for it and that button on the right

Figure 4.6: NP frame slot values and examples

4.2.2 Deriving a Decision Tree model

Table 4.4 describes the full set of features derived from the dialog, spatial and visual context (refer to Figure 4.4 for an explanation of the spatial/visual context features). Mention counts are not considered over Vague or Ambiguous tags, or over Sets. A reference chain is the collection of all mentions linked to a particular entity in a dialog fragment. These features were used as input to develop a classifier to determine NP frames for unseen target referents. Five dialogs were held out as unseen data and the remaining 10 were used to train and adjust the parameters of the decision process. The first procedure was to test whether the three slot values are interdependent. Unfortunately, combining the labels exacerbates the data sparsity problem. We trained several decision trees, varying the independence assumptions:

**Independent** - a decision tree was trained for each slot and their outputs combined at the end (consider correct if all three slots are correct).

**Joint** - a decision tree was trained for the combined label for all three slots

**Conditional** - three decision trees were trained in sequence, each having access to
the output of the previous tree. For example, **Mod-Det-Head** means that first the **Mod** tree was trained, then a tree to classify **Det**, using the output from **Mod**, and finally a tree for **Head**, using both the **Det** and **Mod** values.

All possible orderings between **Mod**, **Head** and **Det** were tested. The best result obtained was from the ordering **Mod-Det-Head**, but the differences between the orderings were not significant. The 10 fold cross-validation results are shown in Table 4.5. There were 632 items in the data set\(^{10}\). The **Conditional** trees outperformed the **Independent** trees by 9%, which is significant at the level of \((p < .0002)\).

As our training data suggests, we test the **Mod-Det-Head** trees against our held out data. We decided to use a leave one out method of training/testing due to the sparsity of data (train on everything but one item and test on it, and repeat for all the test items). The results are presented in the following subsection.

<table>
<thead>
<tr>
<th></th>
<th>Independent</th>
<th>Joint</th>
<th>Mod-Det-Head</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>22.0%</td>
<td>28.8%</td>
<td>31.2%</td>
</tr>
</tbody>
</table>

**Table 4.5:** Testing independence of the slot values

Decision tree classifiers offer the opportunity to examine the relevance of particular features in the final decision. We found that there are significant dependencies between the slots in the NP form. Each time one of the slots’ values was available to the decision process, it was selected as the most informative feature by the next tree. The spatial features were selected as informative in all the trees, most prevalently in

\(^{10}\)we held out 5 dialogs for testing
the decision tree for Mod, suggesting that the decision of including extra information is influenced by the spatial configuration. The information status features and discourse history, such as First Locate, Type, and attributes of the prior mention, were selected as good predictors for the Det slot.

4.2.3 Analyzing the Resulting Algorithm: A Human Based Evaluation

We report several methods of evaluating the NP frames produced using the process given by the decision trees. First, we report the results of a strict evaluation in which the system’s output must exactly match expressions produced by the human subjects. We also compare this result with a hand-crafted Centering-style generation algorithm.

Requiring the algorithm to exactly match human performance is an overly-strict criterion, since in many contexts several possible referring expression forms could be equally felicitous in a given context, so we also conducted a human judgment study. We held five dialogs for testing, for a total of 295 target expressions.

As seen in the previous subsection, the decision tree classifier obtained an accuracy of 31.2%. The most frequent tag gives a 20.0% baseline performance using this strict match criterion. We also compared our algorithm to a baseline provided by a centering algorithm. A manual evaluation of the centering-style generation algorithm described in [Kibble and Power, 2000] against our dialog corpus revealed that both this method and the decision tree performed similarly (64.1% centering, 64.8% our decision tree) and over-generated pronouns. With this baseline, only a subset of the expressions could be compared, since centering can predict a label only on subsequent references.

In the human evaluation, judges compared the NPs generated by our algorithm to the original NPs used by human subjects or NPs created from randomly generated
feature assignments. So that they would understand the context in which the original expression was produced, the judges watched the video, audio, and dialog transcript of the original human subjects experience in the VR using the Anvil annotation tool [Kipp, 2004]. The judges could play or pause the video as they wished. Using the word-alignments established during the data annotation phase, the audio of the test NPs was replaced by silence, and the words were removed in the transcript shown in the time-line viewer. Because the algorithm produces just inclusion/exclusion values for the modifier features, when the algorithm indicated that a modifier should be included, we either included the modifier that was used in the human expressions, or generated a simple spatial description where there was no modifier in the original expression. In Section 6.2 we will present an algorithm that can decide on the content of the modifier, but since this study does not make decisions on the modifiers, we wanted to judge just its modeled behavior, and not the content choice, and this is why we used the human modifiers when it was possible. For each test item, the judges were presented with a selection box showing two possible referring expressions that
they were asked to compare using a qualitative ranking: option 1 is better, option 2 is better, or they are equal. Figures 4.7 and 4.8 show the screen and a close-up of the Anvil tool that the judges used.

Table 4.6 shows the results of human judging. The system’s output was judged as either equal or preferred to the original spontaneous language in 62.6% of cases where these two choices were compared directly. Interestingly, the randomly-generated choice was preferred over the original spontaneous language in 13.0% of trials, and was preferred over the system’s output in 22.5% of trials. Aggregating over all judges, the system’s performance was judged to be much better than random, but not as good as the original human language.
In this section we described a RE generation study for situated dialogs and a novel evaluation setup of the system’s output. The algorithm decides upon the determiner, head and modifier values to be produced in a noun phrase describing an object at a particular moment in the dialog. It achieves 31.2% exact match with human language, but human evaluators judged the output as good as or better than the original human language 62.6% of the time.

These results are the first to show that useful models can be created from spontaneous dialogs for predicting NP form. Previous statistical methods for GRE have used either collections of written text or focus on other aspects of GRE. In the study presented in [Poesio et al., 1999], only a part of the GNOME corpus was used, comprised of museum labels and pharmaceutical leaflets, which are not spontaneous dialogs. [Jordan and Walker, 2005] concentrated on selecting the attributes to be included in the REs, not on the determiner/head variation of REs.

In [Gupta and Stent, 2005], two dialog corpora, MapTask [Thompson et al., 1993] and Coconut [Jordan and Walker, 2005], are used to evaluate generation algorithms and to propose partner-specific adaptations in the RE. [Varges, 2005] also presents a statistical approach to GRE using the MapTask corpus, but his approach is based on a rule-based overgeneration combined with rankings obtained from the corpus. We consider that the use of other statistical methods for this task should be studied. One of the options is Maximum Entropy Models, which would be able to account for the dependencies between the different features utilized in training, and we will present our results obtained with this method in the next section.

The human-computer interaction study described in Chapter 6 showed promising results but its performance could be improved by including some dynamic information
from the history of the user’s movement. With this in mind, we have implemented several new features, but no large improvement was obtained. We consider that this is a consequence of the fact that we have a very small number of data points, the number of labels we want to distinguish between is relatively large (15 different tags), we only have access to automatically obtained features (or features that required a very small human intervention) and many of the features are real-valued, with a large range of variation.

Since we will not do a new human evaluation to assess the improvement of the new features, we ran the decision tree, with a 10 fold cross validation, on the entire data set with all the features mentioned above in this section. The results obtained for the Joint condition (in which we train on the entire label for a NP, containing the determiner, head and modifier slots) is very similar to the one reported on the 5 test dialogs above, classifying correctly 28.2% of the data. Among the most informative features selected by this decision tree are First Locate, SameCatDistractors and chainCount. This is intuitive, as the feature First Locate is determining the type of reference\(^{11}\), the number of distractors are expected to correlate with the type of description needed, and the count inside of the referential chain is a widely acknowledged feature for determining NP form. The Angle and Distance were included in the tree, but not very high up (Angle at the 3rd level in the tree, and Distance at the 6th level).

In trying to improve the models performance, we’ve implemented a set of new features:

- last type of object mentioned in the conversation,

\(^{11}\)it distinguishes between anaphoric references or to a hypothetical entity not yet located and a First Locate, the references where the target could be identified from the world
• if the target is included in a group (grouping is determined as in Section 6.2.5),
• a boolean feature for each basic spatial relation (front, left, right, behind) that
  indicates if the target was in that spatial relation to the follower,
• speed, representing the follower’s speed in the previous second,
• angular speed, representing the DF’s change in angles in the previous second,
• visual recency, representing how recently the object was seen over a 2 s window,
• visual persistence, representing how much time the object was in view in a 2s window.

Adding these new features did not improve the performance, even though our intuition
is that useful information is encoded in them, but probably the size of the data set
is too small. We observed that, for example, building a decision tree using only the
distance feature to predict the inclusion of a modifier or not gives an improvement of
4.5% over the majority baseline, and that using angular speed and visual persistence
we can improve on the determiner decision (4% improvement).

Traditionally in modeling language understanding, noun phrase variation is re-
stricted to a smaller set of values. If we are collapsing our tags in four basic cate-
gories (it, definite, that/that N and one + no surface head, e.g. on the right), then
the best performance obtained is using basic discourse attributes and front-ness in-
formation (46.86%, compared to a baseline of 37.35%, when we use the majority tag).
The spatial information, encoded in the front-ness feature, brings a small relative im-
provement of 2%. These results look promising considering that no human-annotated
features were used (the FirstLocate feature was not included), and no syntactic infor-
mation was available. These results could be usefulness for NLG in creating variation
in the output of the GRE module or assigning demonstrative markings, or using pronouns (for this case, a check of ambiguity is necessary before using an pronoun, since the algorithm overgenerates pronouns and so produces ambiguity).

We conclude that the spatial features we modeled did not bring an improvement in modeling the problem of choosing the right noun phrase form. Some of the factors that can lead to this result are the lack of training data, our assumption that the spatial features should be calculated at the moment that the noun phrase is produced, and the fact that many of the decisions in the NP forms are dictated by the discourse features and spatial features are irrelevant for them (for example, no situated information should influence the choice of an it pronoun). We hope with a larger data collection this problem will be solved, and the importance of different situated features will be further clarified.

4.2.4 Deriving a Maximum Entropy Model

In this section we present a second statistical model trained on the same corpus used in Section 4.2.1. We wanted to test the hypothesis that decision trees, a model that makes decisions at each level on one feature, are capturing the information in our data as well as a more complicated mathematical model, built using Maximum Entropy, a model that uses combinations of features in its decisions. As we observed in Section 4.1.2, for the problem of deciding when a context is appropriate to make a description, the two models did not produce statistically significant different outputs. We wanted to test if the problem of deciding which type of NP to use would benefit from richer modeling.
Compared to the Decision Tree, which classified 28.2% of the data correctly, the Maximum Entropy model produced a similar result, with an accuracy of 25.8% (with a 10 fold cross validation). The difference between the scores is not statistically significant at a 0.05 level (there were 1212 elements in the data set). The Maximum Entropy model assigns high weights to the count of the ID over the whole session, the chainCount and the SameCatDistractors. First Locate was not in the top-weighted features. Angle and Distance were used by the model, but they were not highly ranked.

We conclude that, for the set of features we constructed, we did not obtain better results when we used a model that can take advantage of the interactions between features. Again, similar to the results in Section 4.1.2, we think this might be a consequence of human processing of contextual factors, maybe driven by decisions at the top level.
CHAPTER 5

A STUDY OF DEMONSTRATIVES AND PROXIMITY MARKING

Another objective of our research is to model the influence of situated factors over the choice of a particular proximity marking in demonstratives. These alternations are commonly analyzed as signaling the speaker’s distance from the specified item, where distance is measured in physical space or along some emotional dimension; however, in a study by Byron and Stoia [Byron and Stoia, 2005] other factors are investigated besides spatial configuration of the objects in the world, such as the progress of the task underway and the roles of the partners collaborating on a task. We found that the proximity marking of NPs (e.g. this/this N/these/these Ns vs. that/that N/those/those Ns) and locative adverbs (e.g. here vs. there) is influenced by contextual factors that signal not only the spatial configuration of items in the task world, but also the speaker’s attitude about activities proposed as part of the evolving task. Our analysis indicates that the proximal and the distal expressions may be influenced by different pragmatic dimensions.
5.1 Previous Research on Proximity Marking

Although proximal and distal are the labels commonly employed to discuss the alternation pattern of here/this/this N vs. there/that/that N, there is some discussion that the so-called distal forms are actually unmarked for proximity [Lyons, 1977]. The distal form can be used to signal either the item that is far from me or the item whose position is not specified. We concur with this point, but because the proximal forms are not used for distant items and the other forms are, we will continue to use the label distal. Whether this alternation actually marks distance or not will be a point we return to in the following discussion. The locative adverbs here and there are only used to refer to positions or regions in space, and therefore their distribution is expected to be sensitive to physical spatial arrangements. The demonstrative expressions, however, can be used to specify many kinds of referents, and therefore any proximity relationship that they might signal has a complex relationship to the context in which they are uttered. Demonstrative NPs and demonstrative pronouns are used in English for exophoric reference to entities in the situation of discourse and for anaphoric re-mention of an item already under discussion. The referents of demonstrative exophors and anaphors can either be physical objects or abstract referents such as events or situations. Demonstratives can also be used to perform so-called discourse deixis, a reference to the propositional content of a discourse unit [Webber, 1988].

In their comprehensive treatment of referring behavior in English, Halliday and Hassan [Halliday and Hassan, 1976, page 60], point out several dimensions that account for some instances of nearness/not-nearness markings on demonstrative expressions. A speaker is more likely to use the proximal form to refer back to something he himself said previously, and to use the distal form to refer to something said by
someone else. This tendency was also noted by Glover [Glover, 2000] in a transcript of urban planning discussions. Glover also noticed that the polarity of proximity-marked expressions was affected by the status of referents within the unfolding task. Proximal items were used for open agenda items and distal expressions were used for issues that had been resolved.

In another task-oriented dialog corpus, direction-givers who watched their partner follow instructions for a construction task remotely over a video stream tended to use distal items, but when the experimenters gave the instructor a way to generate deictic gestures into the remote space, the instructors tended to produce proximal expressions to accompany their gestures [Fussell et al., 2004]. Although the actual objects had the same spatial relationship with the speaker in both cases, his ability to gesture to the objects pulled them cognitively closer to him.

The polarity of a proximity-marked expression, therefore, can be used to convey the speaker’s attitude toward the referent, positioning it with respect to himself and the hearer along any number of abstract dimensions. Choosing a proximal expression conveys that the speaker considers the referent to be somehow in my space while a distal expression is chosen when he wishes to convey the fact that the referent is in your space (or just not in mine), where the operative spatial dimension might be physical or abstract. This referent-to-speaker / referent-to-hearer relative comparison may partially explain why the distal forms do not necessarily signal distance from the speaker. They need not signal that a referent is far from the speaker in any absolute sense, but only that the referent is closer to the hearer than it is to the speaker along the operative dimension.
The polarity chosen for a proximity-marked form can also be modulated by temporal relations. Speakers tend to choose the proximal form for a present or future event and the distal form for a past-time referent [Halliday and Hassan, 1976, page 60]. In addition to these dimensions, different polarities of demonstrative forms have been accounted for in terms of the attentional salience of the referent. In the Givenness Hierarchy [Gundel et al., 1993], proximal demonstrative determiners (this $N$/these $Ns$) are expected to be used for referents that have been previously evoked into the short-term memory of the hearer, while the distal demonstrative determiner (that $N$/those $Ns$) are preferred for referents that are not previously evoked but are familiar to the addressee. Piwek et al. [Piwek et al., 1995] also found that cognitive modeling of one’s interlocutor plays a role in proximity marking in a corpus of collaborative dialog. Their subjects were pairs in which a speaker directed another person to construct a small object. Speakers in this setting tended to use distal expressions for referents already known to the hearer, and proximal expressions for new items that the hearer was being instructed to locate, which is counter to the pattern represented in the Givenness Hierarchy. Piwek et al., however, did not report taking the temporal dimension into account, although it may have confounded their findings since it interacts with givenness: objects known to the hearer (+given) tend to be objects that were already used in the task (past-time referent).

In computational models of referring behavior, demonstrative expressions have most often been investigated as a way to formulate exophoric reference, combined with either a pointing gesture or a gaze fixation. Studies such as the ones reported in [Kaur et al., 2003] and [Campana et al., 2002] have looked at the synchronization of spoken and gaze inputs obtained from users referring to items on a multimodal GUI
display. Other groups have worked on integrating demonstrative expressions and pen or mouse gestures to items on a GUI display [Kehler, 2000, Johnston et al., 2002] or using facial expressions [Foster and Oberlander, 2006] while making descriptions. Similarly, Bolt [Bolt, 1980] and Kaiser et al. [Kaiser et al., 2003] created computational models for combining pointing gestures and demonstrative expressions for referring to items in a physical 3D space. However, the meaning of the alternation from proximal to distal polarity has not been systematically studied in terms of computational modeling. In the traditional interaction paradigm for computer applications with conversational interfaces, the human and computer agent partners are not both localized in a shared space, which may hinder the user from adopting the requisite cognitive orientation to characterize referents as either in my space or in your space. Thus, the language that humans produce when interacting with current spoken-dialog systems tends not to contain much proximity-based variation. In contrast, spoken dialog systems of the future will be implemented as localized, embodied agents housed in robots or computer animated characters utilized as partners with humans to collaborate on tasks in spatially-extended workspaces. Search-and-rescue is one such domain in which teams of humans and robots currently collaborate [Kitano et al., 1999], although without the aid of a conversational interface. The study reported in this section was designed to elicit language from human subjects that would be similar to the natural language inputs encountered by automated agents working in such domains. The next section describes the corpus, followed by an analysis of the proximity-marked expressions in this corpus.
5.2 New Factors that Determine Proximity Marking in Instruction Giving Dialogs

The corpus used in this study was different from the one describe in Chapter 3. Both human partners were introduced in the virtual world and none had previous knowledge of the map of the world. Only one of the partners knew the tasks to be completed and collaboration was required for completion. This corpus allowed a more thorough analysis of proximity-marked expressions in collaborative problem-solving dialogs than has been possible with previously existing corpora. A more detailed description of this data can be found in [Byron and Fosler-Lussier, 2006].

![Figure 5.1: Proximity marking compared to absolute distance](image1)

![Figure 5.2: Proximity marking compared to the time of the event](image2)
After examining different forms of proximity, our corpus supports the following conclusions:

- **Proximal-polarity expressions** are used for items close to the speaker in physical space, but distal expressions are used for items either close or far (see Figure 5.1). Also, the type of object and its spatial properties impacts the distribution of polarity: the room the speaker is standing in is almost never described with a distal form, while an object that is close to the speaker, even one the speaker is holding, can be described with a distal form.

- **Temporal relations** are correlated with the polarity of proximity-marked expressions, with proximals being used for referents taking part in current/future events and distals used for past time. The counts and relative proportions are presented in Figure 5.2. References to past time events or items involved in past-time events were predominantly phrased as distal expressions (68%), while present and future time references were more likely to be proximal (59%). The results were significant at a level of 0.01. Again, the proximally marked items’ distribution had a stronger bias towards appearing in events after the speech.

![Figure 5.3: Proximity marking compared to partner completing action](image)
time (82% vs. 18%), while the distribution of distal expressions did not show as strong a preference (41% vs. 59%). This is similar to our observation on the spatial dimension, where the distal shows a less strong marking.

- When a speaker uses a distal-polarity expression in the first pair-part of an adjacency pair, his partner will interpret that as a clue that he should respond with the second pair-part. Distal-polarity expressions are less sensitive to the space and time dimension than proximal-marked items are, but we observed that they might be more sensitive to obligation within the task space. Figure 5.3 shows the distribution of proximal/distal expressions and whether the hearer took up the obligation to respond or not. As the chart shows, the hearer reacts in all cases where a distal item was used. Some examples labeled on this dimension are: *Anything exciting in there?*, *Let’s go ahead and hit this button and see what happens*, *Pushing that button should help us with the other task*, where we marked which one of the partners did the action associated with the turn. Out of the 35 distal items that could be marked on this dimension, there were no examples where a distal expression was used and the hearer did not respond. There were 36 proximal items, and they were allocated in a 2:1 ratio between the cases when the hearer did respond vs. when he did nothing. This predilection for responding to the speaker’s requests is justified by the fact that people participating in this study were collaborative, trying to achieve the tasks together, and it is normal for a partner to take up an obligation introduced by the other partner.

Proximity markings interact with a range of factors in the extra-linguistic context. Understanding these interactions is an important step in developing dialog systems.
that will use these forms in a natural way. In the direction giving corpus we present in Chapter 3 we have noticed a very small number of proximal marked items, which could be a consequence of the fact that in this setting the obligation to complete the task falls always on the follower. In the modeling in Chapter 4, proximal-marked items were not predicted, but only distals, due to the corpus distribution. A machine learning algorithm that generates proximally marked referring expressions in different conditions might benefit from information about who is giving the instructions and who is suppose to complete the tasks.

Our study confirms linguistic theories on the use of proximity marking using corpus data and also adds new evidence that distal polarity expressions may have an effect on intention recognition within collaborative dialogs. To our knowledge, this is the first study in proximity marking that used a corpus of spontaneous human dialogs and recordings of positional information to validate its claims.
CHAPTER 6

HUMAN-COMPUTER INTERACTION: AN IMPLEMENTED INSTRUCTION GIVING SYSTEM

In this chapter we will present the details of our implemented instruction giving system. Having an operational system for situated worlds is a lack in the literature. Very few systems let the user change the environment and navigate through space, and not many of these systems present human evaluation results or statistics obtained from the interaction with the system. We consider that this study is an important contribution to the dialog system/generation community, and fills a gap in the field.

6.1 Human-Computer Interaction in Virtual Environments

The statistical algorithms described in Chapter 4 were obtained from transcripts of human-human interaction. They modeled the human interaction, but were tested off-line, not embedded in a running system. The next step is validating some of the observations we made in a human-computer configuration. For this, we have implemented a natural language instruction giving system that accomplishes the Direction Giver role in the same problem scenario used for our data collection.

The dialog system has full knowledge of the world layout and of the tasks that the user will complete. It also receives information about the position and orientation of
the player (Direction Follower) in the Quake world. An illustration of the system’s physical setting in interacting with a user is presented in Figure 6.1.

For simplicity and due to our focus on the generation aspects, we have devised an interaction model in which the human is not able to give spoken input to the system, and his only input is in the form of actions in the virtual world. In this way, we avoid the problems that arise from speech recognition and understanding, and still have a good framework for testing various aspects of the generated language.

To be able to answer our research questions related to the generation of noun phrases as a dynamic, contextual process, the first stage is building a system that can interact with people with a reasonable task completion rate. We have implemented an Instruction Giving System that produces descriptions for various commands that need to be executed in the virtual world, checks the user’s progress and generates contextual commands. The language depends on the person’s position and orientation in the virtual world, and the particular context (other objects and their placement) that the

Figure 6.1: System’s physical setting
user is currently in. Even though the focus here is not studying how to produce a distinguishing description or what are the optimal strategies for this problem, we still need the system to produce a fluent, distinguishing description if we expect the user to be able to identify the objects of the commands and act on the right referents.

Our system’s strategy is simple: a plan consisting of the steps that need to be completed for each task is part of the system’s knowledge base, and steps are described...
in a sequence. Figure 6.2 presents a diagram of the system’s modules. The Dialog Manager selects the next plan step and calls the Generation module to produce the associated language. We implemented a template lexicalization module that converts the plan steps into commands. The system generates referring expressions in the GRE module, which is aware of the information from the Quake world. After speaking the words to describe the next step in the plan, the system enters a loop in which it checks that the expected effects of the generated turn have been completed. If the effects are not perceived after a small time period (6 seconds), then the system makes a second attempt to produce language associated with the turn (a new GRE will be computed). If after two attempts to describe the step in the plan the conditions for moving on to the next step are not yet met, the system will give up and announce failure on this task. If the conditions are met, the system will produce an acknowledgment and proceed to the next step.

The vocabulary of the system is small, containing a total of 70 words. All possible system turns are presented in Appendix E. Even with such a simple framework, a lot of interesting research questions arise.

The research questions we want to address in the following human-computer study focus more on the interaction between the referring expressions and situated aspects of the environment than on defining the applicability of spatial relations or the selection of properties for constructing a distinguishing expression. Situated GRE in itself poses numerous research questions, as presented in Section 2.4. We developed a basic algorithm following the locative incremental algorithm presented in [Kelleher and Kruijff, 2006]. The previous machine learning algorithm that we developed for choosing a noun phrase form only indicates if we need to include a modifier or not, and not
which modifier, or if the expression obtained is ambiguous. The locative incremental
algorithm from [Kelleher and Kruijff, 2006] builds a GRE by applying spatial proper-
ties in sequence until it either obtains a distinguishing description, or it exhausts all
its strategies, and produces an ambiguous description. The algorithm is adapted for
our particular world which has no distinguishing features for an object other than its
category and position, and tries to produce distinguishing expressions by constructing
expressions that use spatial properties. The algorithm orders possible landmarks (a
reference object or relatum used in the expression involving a locative prepositional
phrase, e.g. the cabinet, in the expression the button closer to the cabinet) and spatial
relations by their cognitive load, to produce a referring expression.

6.2 The GRE Algorithm Employed in the Basic System

In the process of generating a referring expression, we need to determine the
content we want to include (what are the properties included in the RE), and how
to realize that content. As previously mentioned, we have adapted the algorithm
presented in [Kelleher and Kruijff, 2006] for determining the content of the referring
expressions for objects that the system intends to refer to. We chose the Kelleher and
Kruijff algorithm because it represents the state of the art rule-based algorithm for
GRE in situated worlds, and provides a cognitive motivation for ordering the spatial
relations and landmarks selected for building a RE. The algorithm receives an ID of
an object from the virtual world, the user’s position and orientation and positional
information about all the objects in the world and returns the spatial relation and
the landmark to be used in the RE. In the following paragraphs, we will present the
decisions we made while implementing the GRE algorithm in our dialog system and how it compares to the behavior observed in the human-human corpus.

The previous statistical model presented in Section 4.2 was trained on all mentions of buttons, doors and cabinets in the dialogs, but the Instruction Giving System at this stage has a different configuration, in which the direction giver (the system) always gives one command for which it produces a RE. There are no referential chains\textsuperscript{12}, as the follower will not be able to give spoken input. All mentions in the system are \textbf{First Locate} noun phrases (as introduce in Section 4.1.1), that expect the user to identify the referent and immediately give feedback through his actions. Of course, human-human interaction is very complex, and good algorithms for spoken noun phrase generation and use of pronouns have not yet been developed. Here the focus will be more on the \textbf{First Locate} descriptions, the noun phrases that were associated with the moment when both partners identify an object from the world and achieved joint spatial reference. This type of expression is very important in situated dialogs, as it represents the moment the dialog partner picks the object out of the surrounding world. We consider that studying these cases is very beneficial for situated dialog systems, as they represent the connection between the real world objects and the way people identify them and talk about them.

The statistical algorithm in Section 4.2.1 was trained for determining what type of determiner the resulting noun phrase will have, the head information (will it be a noun for the type of object, \textit{one}, or a pronoun), and if the expression should include a modifier, for all types of references (not only \textbf{First Locate}). The algorithm was not trying to decide which modifier to use in the case one was needed, which is the

\textsuperscript{12}A referential chain is the collection of all mentions linked to a particulate entity in a dialog fragment, such as a first mention and subsequent anaphoric mentions of the same entity.
purpose of the algorithm in [Kelleher and Kruijff, 2006]. In the following we will present the implementation decisions we made in building the GRE algorithm.

In [Kelleher and Kruijff, 2006] the authors propose an extension of the Incremental Algorithm [Dale and Reiter, 1996]. This algorithm cycles through a list of properties for an object (e.g. color, size, etc.), and adds the property to the description if it eliminates some of the distractors, until a distinguishing expression is obtained. By distractors we mean objects that share properties with the target and are competitors in interpreting the expression. Later additions to this algorithm order the objects with respect to their salience [Krahmer and Theune, 2002], and if the target salience is greater than the maximum distractor salience of the remaining distractors, a distinguishing expression is generated. Kelleher and Kruijff’s algorithm first tries to produce a referring expression using this algorithm, and if a distinguishing form is not yet obtained, cycles through a set of candidate landmarks and applies spatial relations in an order motivated by the cognitive load associated with processing the spatial relation. The Locative Incremental Algorithm is presented in Figure 6.3, where the BasicIncrementalAlgorithm(T,D) is Dale’s Incremental Algorithm [Dale and Reiter, 1996] with salience information incorporated. We will elaborate on the construction of the Candidate Landmarks list and the list of Spatial Relations that we have implemented in our system in the next subsections.

Our world design ensures that all objects of the same type have the same properties (e.g. all buttons look the same), so no minimization of the distractor set can be made by including properties like color, size, etc. In this case, calling the Basic Incremental Algorithm will always add only the type property (button, door or cabinet), as no other property distinguishes objects of the same type.
The Locative Incremental Algorithm (T,D) \((T = \text{target}, \ D = \text{Distractors})\)

\[\text{DESC}= \text{BasicIncrementalAlgorithm}(T,D)\]

if DESC = Distinguishing then
  return DESC

else
  create \( CL = \{x : x \neq T, DESC(x) = \text{false}\}\)
  for \( i = 1 \) to length(\( CL \)) \((CL = \text{CandidateLandmarks ordered by salience})\)
    for \( j = 1 \) to length(\( R \)) \((R = \text{SpatialRelations ordered by cognitive load})\)
      if \( R_j(T,CL_i) = \text{true} \)
        TL = \{CL_i\}
        DL = \{z : z \in CL, R_j(D,z) = \text{true}\}
        LANDDESC = BasicIncrementalAlgorithm(TL, DL)
        if LANDDESC = Distinguishing then
          Distinguishing locative generated
          return \{DESC, R_j, LANDDESC\}
      end if
  end for
end if

FAIL (all available landmarks and relations have been tried)

Figure 6.3: The Locative Incremental Algorithm from [Kelleher and Kruijff, 2006]

6.2.1 Salience

 Object salience is considered in [Kelleher and Kruijff, 2006], and the relative salience of an object is calculated as the average between its visual salience \(S_{vis}\) and its discourse salience \(S_{disc}\). Because our system is designed to accept only input in the form of actions in the world, the system initiates turns that always contain a description of a new object, so the \(S_{disc}\), based on recency of mention, is equal for all objects. \(S_{vis}\) is calculated in [Kelleher and Kruijff, 2006] based on placement on the screen using a false coloring technique. This associates each object with a new color that is not present in the scene (thus the name of false coloring), and then re-renders the scene and calculates how many pixels of that color are present in the image, so determining the visible dimension of the object (number of pixels of that color rendered in the scene). More details about the visual salience algorithm can be found in [Kelleher and van Genabith, 2004]. Implementing the false coloring technique for
the Quake engine is not obvious, as it will be necessary to modify the code to re-render each scene. Also, we do not want to use visual salience to discriminate between the target and objects in the distractor set, because this might lead to confusion in our setting. As the user is free to move at any time and can check the entire room configuration, receiving a RE that relies on the visual salience of the object at the moment of calculating the RE to make it distinguishable can pose problems.

The only place where we use salience information is in choosing between the visible objects that can be used as landmarks by the GRE algorithm\textsuperscript{13}, so we ordered our objects by size: cabinets are more salient than doors, which are more salient than buttons.

\subsection*{6.2.2 Frames of Reference}

To generate a locative expression, a frame of reference needs to be selected. A locative expression is \textit{“an expression involving a locative prepositional phrase together with whatever the phrase modifies (noun, clause, etc.)”} (see [Herskovits, 1986, page 7]). The term frame of reference is related to the coordinates in space from where the description is true. In general, three frames of reference are employed in English:

- the absolute frame, based on the environment, the world, where the origin of the space is the same as the origin from where the description is made (e.g. \textit{the town in the North}),

- the intrinsic frame, based on a landmark’s origin and its canonical position (e.g. \textit{the cat in front of the cabinet})

\textsuperscript{13}in the reference algorithm, we consider distractors any entities that match the description, not only the ones with larger visual salience

73
• viewer centered, which uses the origin of the viewer in relations that presuppose a viewpoint to locate an object relative to a landmark (e.g. *the house on the right of the tree*).

We will always use the user’s frame of reference to make a description (egocentric frame), because this is the behavior we observed in the human corpus, when completing tasks similar to the tasks of the system. The Locative Incremental Algorithm does not explicitly mention how the choice of frame of reference is to be made. Other frames of reference could be employed (our world contains objects that have an intrinsic frame of reference, like cabinets, for example, that have a canonical *front*), which can facilitate a description relative to the object’s front or sides. We expect that when two human participants are immersed together in a virtual environment, the intrinsic reference frame of an object that they can both see will be chosen more often, due to ease of interpretation. In our design, there is only one person in the world and the user-centered frame was preferred by people (in the 15 dialogs collected, there was no other frame of reference except user-centered for the NPs that refer to doors, cabinets and buttons that we have annotated).

### 6.2.3 Landmarks

The proposed possible landmarks in [Kelleher and Kruijff, 2006] are *the speaker, the hearer, the scene, another object and a group of objects*. From our human-human corpus, where only the DF, which plays the role of the hearer, is placed into the world, we observed that the speakers in our domain only used three types of landmarks: *the hearer, another object or a group.*
To serve as a landmark in [Kelleher and Kruijff, 2006], an object must complete the following conditions: be distinct from the target and not be in the distractor set. Also, a candidate landmark gets described using the basic incremental algorithm, so only the type will be added in our domain. The Locative Incremental algorithm does not apply for the landmarks so no recursion in describing landmarks can occur. With this in mind, we chose as landmarks objects that are a singleton in the field of view (so that a description such as the cabinet will be sufficient). A condition not treated separately in [Kelleher and Kruijff, 2006] that we also check up front is that the target is the only object of that type in the particular spatial relationship to the landmark. Otherwise the description would not be distinguishable. The Locative Incremental Algorithm does take this case into account by the fact that it will fail to produce a distinguishable expression for the landmark and a distractor list that includes the landmark. We consider it is more intuitive to just do this check upfront instead of calling the basic incremental algorithm with a distractor list that includes the object we want to describe.

The preference order between different types of landmark categories is not rigorously motivated in [Kelleher and Kruijff, 2006], but we have decided to adopt it, as it makes sense intuitively and we do not have other statistical data to support a different ordering. A study that evoked landmarks or rankings from human subjects would elucidate this question, but this is left for future work.

**Groupings**

Details about what is considered a group of objects that can be selected as a landmark or if this feature has been implemented in the system are not given in [Kelleher and Kruijff, 2006]. Determining what a group is and how to make a description of
an object relating it to a group is still an open research question [Funakoshi et al., 2006]. There is no GRE algorithm at this moment that can produce groupings in a general context. Our domain is particularly difficult because there are no distinctive attributes for the objects. Funakoshi et al. [Funakoshi et al., 2006] present a group-based algorithm specific for these situation with no distinctive attributes, but this work is focused on 2D spaces and the authors acknowledge that more investigations are needed to apply their method to three dimensional worlds.

Looking at the corpus, we observed that from the total of 423 First Locate expressions, 14 82 (19.4%) involved groups as landmarks. In all these cases the group was formed by linearly aligned objects of the same type. These mentions relative to a group are important in real interaction, as [Moratz and Tenbrink, 2006], a study of direction giving to a robot in a real space, also shows cases of using counting (5.81% out of total utterances identified the goal object via counting). In work on the same domain, the authors note that “Our findings show that counting and ordering is a rather natural strategy to use for object reference in a row-like configuration” [Tenbrink et al., 2007, page 5]. Also, for lexicalizing these references, the authors notice that “the direction of counting needs to be inferred in as many as 1/3 of the cases”, and that the left direction, or the direction closer to the robot is preferred.

Perceptual groups are determined in the literature based on common attributes (such as color, size, etc.) combined with spatial proximity [Thorisson, 1994, Gatt, 2006, Funakoshi et al., 2004], but none of these computational models includes linear alignments of objects, even though their importance is acknowledged.

14 We used only Direction Giver language, no vague or ambiguous mentions
Examples of human produced expressions for B6:
- the back corner right button
- the button that’s on the wall in front of you to the left
- the button that is farthest from the cabinet on that wall that you are facing
- the second button after the cabinet
- the third one from your left
- that one
- the one right there
- the one on the left

Figure 6.4: Overhead view of a room in our VR world.

We decided to implement a grouping based on the object’s category and spatial alignment due to its frequent presence in our corpus. Other types of grouping would be helpful in other environments, and the frequency of this type is just a consequence of our world design. We modeled a simple algorithm that checks if all objects of a particular type located in a room are aligned with either X or Y axis. If that is the case, the objects are considered a group and references such as the middle button, the left button and the second button from the left will be generated. In our world, objects of interest are placed on the walls, and rooms bounding boxes are rectangles placed without rotation from the world’s axis. So centers of objects that are in a linear arrangement will form a line parallel with either X or Y axis. These linear alignments are equivalent with the objects being on the same wall, and this property of objects being bounded in perceptually significant groups has been called closure in the literature.
Of course, incorporating more complex models for determining groupings will help produce more natural expressions, but algorithms for this problem have not been developed yet. We find a group only in the case where all objects are in a linear alignment, but people can identify more groupings in a variety of contexts. Figure 6.4 presents an example from our virtual world, where different groupings can be created: (B7, B8), (B5, B6), (B4, B5, B6), (B5, B6, B7, B8), etc. Determining all possible groupings is an exponential algorithm in the number of objects of interest. In our world, where there is a small number of objects in the room, this would probably not be a problem. But even then, the groups identified are not necessarily consistent with what humans use for GRE. In our example in Figure 6.4, extending our simplistic algorithm for determining all groupings, (B4, B5, B6) will be a group. To refer to B5 the system will employ the phrase the middle button, which is very unnatural. Proximity has been acknowledged as an important feature in determining groupings, and it needs to be taken into account in conjunction with spatial alignment.

Supposing that we have access to the correct groupings, identifying which one to use in generation and how to construct the final referring expression is still problematic. Let’s say we identified the groups in the setting as (B4), (B5, B6) and (B7, B8). To identify B6, the button that opens the cabinet containing the Quad Damage, we would need to distinguish the grouping of (B5, B6) from the rest of the buttons, and to identify B6 inside of this group. One possible expression, assuming that the user has just entered the room and is facing into the room, would be the left button of the two buttons on the right wall. An absolute description would be even more complicated.
Our world is a dynamic one, and people choose different groupings and description strategies depending on the context. Figure 6.4 shows some of the phrasings people used to refer to B6 in the corpus, showing a great variety of expressions, that use different groupings and strategies to build the REs. Notice that the REs are produced from the viewpoint offered by the current position of the Direction Follower.

We consider that reference using groups of objects opens a wide range of interesting research questions, and our solution for lexicalizing objects in relation to groups is specific for our world design.

6.2.4 Distractors

In a generation algorithm, it is important to define the set of distractor objects, objects that are considered present in the current context of the dialog or are easily accessible for reference. In [Dale and Reiter, 1996] the objects that are included in the context are the entities that the hearer is currently assumed to be attending to (page 236), a definition which is hard to translate into an exact list of conditions for an algorithm. In his thesis that studies the interpretation of spatial language [Kelleher, 2003], Kelleher decided to consider objects introduced previously in the discourse or that have been in the user’s field of view recently as part of the context. This might be appropriate for interpretation, but not in our domain. People can understand reference to entities that they have not yet seen, but are located in their proximity. Figure 6.5 illustrates an example of this behavior, where the expression the cabinet on the right is felicitous, even though the user hasn’t yet seen the cabinet. This accentuates one of the interesting factors of our configuration, the possibility for movement that the users have.
Our virtual world is a hierarchical domain as described in [Paraboni et al., 2007]. As Paraboni et al. show, referring expressions need to be easy to identify, and might include redundant information in cases where this would accelerate resolution. In a domain that is hierarchically ordered (structured like a tree), interpretation is shown to follow an algorithm called Ancestral Search. Ancestral Search means that, in a hierarchical domain, the hearer will search for a referent exhaustively in the current subtree before he will move up on the tree. Building interiors are usually hierarchical, and our domain is hierarchical, because it is organized in two levels, each level containing a set of rooms, and each room containing several objects (doors, buttons, cabinets).

It is then intuitive to consider as distractors for the generation algorithm only the objects present in the room that the user is located in, restricting it to a subset of the objects in the world. This also corresponds to the type of mutual knowledge defined as physical co-presence in [Clark and Marshall, 1981]. Physical co-presence can be current, when both partners are simultaneously attending to an object, prior, when an object has been seen before, and it is memorable, and potential, for items that are not seen yet, but are locatable. The dialog system is not located in the virtual
world (only its human partner has a position and orientation in the world), but it presumably has full knowledge of the user’s information and it indexes the description to the user’s current location. In the human data, direction givers frequently assume the follower’s origin, using expressions like *on our left* and the follower presumes that all the spatial information he has is shared with the direction giver, as they share the same view of the world.

We consider that this hierarchy matches what we have observed in the corpus, and also motivates us to restrict the distractor set to the objects from the mutual knowledge that are introduced through physical co-presence. This speeds up the generation algorithm, and also produces intuitive expressions, as none of our human subjects used any lexical items to distinguish the target from an object not located in close proximity if it did not have other special discourse properties to make this distinction necessary.

This is not to say that humans in our corpus have been that strict about using as distractors only the objects in the same room. There are some special cases when objects are not in the room, but they are still accessible. Sometimes the objects from the next room are taken into consideration by people, when the rooms are connected through an open door. Also, the way our map is built, doors have their own bounding box that are not overlapping with a room, and there are cases in the corpus when people are stopped in a doorway and talk about the objects in the room they are facing. In the case when the person’s coordinates are in a doorway, in our dialog system we consider as distractors the objects from both rooms the door connects. A more complex model, that takes into account the cases where people are looking through an open door into another room, might be necessary to match
human behavior. In our dialog system, the turn taking behavior we implemented assures that the follower is in the room (or in the doorway that leads to the room) containing the target of the RE before producing the expression.

### 6.2.5 Spatial Relations

Unlike static properties, spatial relations must be calculated, and this takes time. We want to find a spatial relation that applies, and that is similar with what people would chose in that context. To order the spatial relations that are calculated by the system we will use the same preference ordering proposed in [Kelleher and Kruijff, 2006], developed based on the psychological research in [Logan, 1994, Logan, 1995], and displayed in Table 6.1. Between contrastive or relative uses of the spatial features, the contrastive ones are considered computationally easier to interpret. A contrastive use of a spatial relation is when the object’s location is contrasted to that of the distractors, for example, when we use *the button on the left* and there is only one button on the left, the other objects are in other spatial relations. A relative use of the same phrase would be when there are other objects in the acceptable area for *left*, but the target matches the prototypical *left* direction much better than the competitors.

<table>
<thead>
<tr>
<th>Low Cognitive Load</th>
<th>TYPE</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ABSOLUTE ADJECTIVE</td>
<td>e.g.: <em>button, door, cabinet</em></td>
</tr>
<tr>
<td></td>
<td>RELATIVE ADJECTIVE</td>
<td>not used in our system</td>
</tr>
<tr>
<td></td>
<td>TOPOLOGICAL LOCATION</td>
<td>e.g.: <em>near, far</em></td>
</tr>
<tr>
<td></td>
<td>PROJECTIVE LOCATION</td>
<td>e.g.: <em>right, left, front</em></td>
</tr>
</tbody>
</table>

Table 6.1: The cognitive load associated with object properties [Kelleher and Kruijff, 2006]
In our algorithm, we implemented the projective relations *right, left, front* and *behind* (e.g. *the button on the right*), both relative and contrastive uses, and the topological relations *close* and *far*, in their relative use (e.g. *the button closer to the cabinet*). Also, due to observing similar referring expressions in the corpus, we will also check for two simple relations of an object relative to a grouping that it belongs to, generating expressions such as *the middle button, the left/right button* (from a group of two buttons, even if both are in front) or *the second/third button from the left* (in an alignment of more than 3 objects, we always used left as the direction of counting). The combination of spatial relations/landmarks implemented in our system and the order they are considered is presented in Table 6.2.

For determining if the relation *front, behind, left* or *right* holds in its contrastive use, we check the inclusion of the angle between the person’s view center and the object’s bounding box center in the enlarged acceptance areas defined in [Moratz and Tenbrink, 2006] for the intrinsic frame of reference and presented in Figure 6.6.

<table>
<thead>
<tr>
<th>Landmark</th>
<th>Spatial Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 user</td>
<td>front (contrastive use)</td>
</tr>
<tr>
<td>2 user</td>
<td>behind (contrastive use)</td>
</tr>
<tr>
<td>3 user</td>
<td>left (contrastive use)</td>
</tr>
<tr>
<td>4 user</td>
<td>right (contrastive use)</td>
</tr>
<tr>
<td>5 user</td>
<td>front (relative use)</td>
</tr>
<tr>
<td>6 user</td>
<td>behind (relative use)</td>
</tr>
<tr>
<td>7 user</td>
<td>left (relative use)</td>
</tr>
<tr>
<td>8 user</td>
<td>right (relative use)</td>
</tr>
<tr>
<td>9 another object</td>
<td>close (relative use)</td>
</tr>
<tr>
<td>10 another object</td>
<td>far (relative use)</td>
</tr>
<tr>
<td>11 group</td>
<td>middle</td>
</tr>
<tr>
<td>12 group</td>
<td>left (contrastive use)</td>
</tr>
<tr>
<td>13 group</td>
<td>right (contrastive use)</td>
</tr>
<tr>
<td>14 group</td>
<td>2nd from the left</td>
</tr>
<tr>
<td>15 group</td>
<td>3rd from the left</td>
</tr>
</tbody>
</table>

**Table 6.2:** The spatial relationships/landmarks employed in our system
For a relative use of the same spatial relations, we use the closeness to the prototypical direction, as in [Moratz and Tenbrink, 2006] and presented in Figure 6.7. The domain in [Moratz and Tenbrink, 2006] is similar to our own: humans are giving instructions to move to a target object to a robot situated in a small real world setting. In this research, the robot is trying to interpret human language and not generate it. That is why relations are given weight and only the best matching combination is used to resolve the referent. We want to accomplish the opposite task - we want to create a referring expression that allows a person to identify the referent. That is why we decided to produce a relative use of a spatial relation only in the case when the target is closer to the prototypical direction with a difference of at least 30° than any of the distractors.
When the spatial relations left and right are used in relation to a group, we check that the object is on the margin of the group and which relation we should use given the current position/orientation of the user.

For determining if the topological relations hold, we employ a threshold that we chose experimentally by examining our corpus. An object is considered the closest to a landmark when \( \text{distance(landmark, anyDistractor)} > \text{distance(landmark, target)} + \text{threshold} \) and the farthest when \( \text{distance(landmark, anyDistractor)} < \text{distance(landmark, target)} - \text{threshold} \).

Two important facts in implementing the spatial relations for our dialog system are that 1) we want to generate easy to interpret expressions and 2) the checking of conditions has to be very fast, because the system is used in real-time interaction with the user. More complicated treatment of spatial relations could produce more appropriate expressions. One possibility would be using potential fields such as in [Gapp, 1995, Gapp, 1994b, Gapp, 1994a] and order all applicable spatial relations by their probability, but we decided that this simpler and faster solution is sufficient for our needs. Kelleher and Kruijff suggest treating spatial relations and frame of reference ambiguity using potential fields as in [Kelleher and van Genabith, 2006]. Still, the framework is focusing on topological relations near, close to, far, etc and only on the in front, back projective relations, and from an interpretation perspective.

Some problems arise from treating objects as bounding boxes and using their center in calculations as has been acknowledged in the field [Kelleher and van Genabith, 2006], even though all the objects we consider for GRE are of a rectangular shape and the bounding box is the actual shape of the object. Figure 6.8 shows an example where door1 is very close to the user, who might be approaching it on his way to
Figure 6.8: The expression generated using the angle to the center of the bounding box will be *the door on your left* which does not correspond to our intuition.

Figure 6.9: The expression generated using the angle to the center of the bounding box will be *the door in front of you* which does not correspond to our intuition.

the next room. The system might generate an expression such as *the door on your left*, which would be confusing for the person, who perceives it as *in front*. This shows an example where just the angle information is not enough, and distance and visual salience interact in the choice of an optimal expression. The same conclusion was found in [Gapp, 1995], that notes that in choosing spatial relations *“there was no major effect due to the distance. However, distance interacted with the angular deviation if the located object was very close to the reference object”*. In our case the object is very close to the user.

Another example, shown in Figure 6.9, emphasizes the problems that can arise from the object’s canonical orientation. The door has a front and a side, and even if the door’s angle to the person falls into the acceptable area for *front*, the fact that the door is sideways makes the expression *the door in front of you* not intuitive. There might also be a bias in the use of spatial prepositions related to the actions that the
objects participate in. Spatial attributes interact with the object’s affordances: for example, on the path a user takes to go through a door, the door will implicitly be in front. These cases show examples of the problems that can arise from a simplified treatment of objects, but their frequency in the real human/system interaction is not significant, and human partners can accommodate such expressions, even if they are not optimal.

6.2.6 Realization

The output obtained from the Incremental Locative Algorithm is underspecified for obtaining a lexicalization. For example, the result of the algorithm might look like 
(type:button, SpatialRelation:left, landmark:user), (type:door, SpatialRelation:closer, landmark:cabinet), (type:button, SpatialRelation: middle, landmark: group of B1, B2, B3). The resulting noun phrase might not contain all the information, as we have noticed in the corpus there are many things left out of the surface form. For example, the frame of reference is sometimes explicitly stated, as in the button on your right, but there are also a large number of noun phrases with no explicit frame information, such as the button on the right, which can be used if it is to the right of the user, or to the right in a group of two buttons, even if the group is situated directly in front of the user. In our corpus, for the First Locate items described with a spatial relation, 38% contained an explicit mention of the user-centered view. Also, the landmark can be sometimes not mentioned, as it usually happens in the case of groups. The corpus contains expressions such as the middle button or the left button, and very rarely something like the middle button of those three buttons in front of you or the left button of those two buttons on your right. People leave this information out, and
producing it in the generated system utterance would be unnatural. With this in
mind, and after studying the corpus behavior, the algorithm for surface realization
implements the following design choices:

- Every generated expression uses the determiner *the*. We plan a new study for our
  future work to determine the influence of different determiners (do people like
  variety, is there any perceived difference, does the proximal marking behavior
  we observed in the off-line study transfer to human-computer interaction, etc.),
  but for now we decided to not introduce more variables in the basic system.

- When a group is chosen as the landmark, it will never be lexicalized (e.g. the
  system will say *the middle button* or *the left button*).

- We will always include information associated with the user as a landmark,
  to avoid ambiguity (this could be included only in the cases where there ex-
  ists possibility for ambiguity, but to determine that we need a more complex
  model of groupings, so we decided it was more beneficial to be redundant than
  ambiguous).

- When the landmark is a single object, it is always included in the noun phrase
  as simple RE formed from the type (e.g. *the cabinet*)

- Since the only frame of reference we are using is the egocentric one, we will
  never mark it specially in the resulting noun phrases.
6.3 Comparison between Corpus Distributions and the GRE Algorithm Output

In this section we present our observations on comparing the First Locate expressions from the human corpus with the output of the algorithm described in Section 6.2. Evaluation of GRE algorithms has been recognized as a difficult task in the NLP field, and we presented some of the challenges in Section 2.5. Comparing between the GRE output and the human output is interesting, because even if we made some choices in the algorithm looking at the corpus (like making decisions about what is a landmark or when to compute spatial properties and what spatial relations to use), we still used a methodological way to construct the noun phrases. The algorithm follows Kelleher and Kruijff’s Locative Incremental algorithm [Kelleher and Kruijff, 2006], and basic spatial relations are calculated using results from [Moratz and Tenbrink, 2006]. A rule-based algorithm can tell us when its rules did not produce a distinguishing expression and can guarantee the correctness of the spatial expressions used in the referring expressions, with regard to the rules and information it has access to. Showing how generation algorithms compare to corpus data has been the focus of recent research [Viethen and Dale, 2006, Gupta and Stent, 2005].

In the following subsections we will look at the properties of the First Locate expressions in the corpus and how the rule-based algorithm output compares to the human output, from the point of view of the content selected and the distribution of spatial relations, landmarks and frames of reference used.
<table>
<thead>
<tr>
<th>Determiner</th>
<th>Value</th>
<th>Count</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>225</td>
<td>53.2%</td>
<td></td>
</tr>
<tr>
<td>that</td>
<td>121</td>
<td>28.6%</td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>31</td>
<td>7.3%</td>
<td></td>
</tr>
<tr>
<td>none</td>
<td>31</td>
<td>7.3%</td>
<td></td>
</tr>
<tr>
<td>this</td>
<td>15</td>
<td>3.5%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Head</th>
<th>Value</th>
<th>Count</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>common noun</td>
<td>324</td>
<td>76.6%</td>
</tr>
<tr>
<td></td>
<td>one</td>
<td>80</td>
<td>18.9%</td>
</tr>
<tr>
<td></td>
<td>none</td>
<td>11</td>
<td>2.6%</td>
</tr>
<tr>
<td></td>
<td>that</td>
<td>7</td>
<td>1.65%</td>
</tr>
<tr>
<td></td>
<td>it</td>
<td>1</td>
<td>0.24%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Modifier</th>
<th>Value</th>
<th>Count</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>+</td>
<td>278</td>
<td>65.7%</td>
</tr>
<tr>
<td></td>
<td>−</td>
<td>145</td>
<td>34.3%</td>
</tr>
</tbody>
</table>

Table 6.3: Distribution of Determiner, Head and Modifier values in the First Locate expressions

<table>
<thead>
<tr>
<th>Type of information used in the surface form</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>use spatial info</td>
<td>258</td>
<td>61%</td>
</tr>
<tr>
<td>do not use spatial info</td>
<td></td>
<td></td>
</tr>
<tr>
<td>no modifier</td>
<td>145</td>
<td>39%</td>
</tr>
<tr>
<td>non spatial modifier</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td></td>
<td>165</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.4: The distribution of spatial/non spatial descriptions

6.3.1 Corpus Properties of First Locate Expressions

A total of 466 expressions were marked First Locate in the corpus, a boolean feature that indicates whether the expression is the first one that allows the follower to identify the object in the world, in other words, the point in the dialog where joint spatial reference is achieved, and the referent can be picked out from the world. 439 expressions are Direction Giver turns, from which 7 are marked vague, 3 were abandoned and 6 delayed, leaving us with 423 expressions we consider in our statistics. Figure 6.3 shows the distribution of determiner, head and modifier values over these expressions, indicating a wide variety of NP forms.

\[\text{see Chapter 3 for annotation procedure and inter-annotator agreement}\]
**Table 6.5:** The modifiers for the non-spatial relations

<table>
<thead>
<tr>
<th>Modifier</th>
<th>Count</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>no modifier</td>
<td>145</td>
<td><em>that door, that one, the cabinet</em></td>
</tr>
<tr>
<td>an action</td>
<td>14</td>
<td><em>the open door, the door that you came in, the cabinet that did open</em></td>
</tr>
<tr>
<td>seen</td>
<td>2</td>
<td><em>the door that you just saw, the first door you see</em></td>
</tr>
<tr>
<td>other</td>
<td>2</td>
<td><em>the other cabinet, the other one</em></td>
</tr>
<tr>
<td>adjective</td>
<td>1</td>
<td><em>the pretty door</em></td>
</tr>
<tr>
<td>more complex</td>
<td>1</td>
<td><em>the door to the room with the chair</em></td>
</tr>
</tbody>
</table>

**Spatial Relations Distribution**

From these expressions, only 61% used a spatial relation in the surface form to describe the target (see Table 6.4). Other ways of locating a referent are forms with no modifier, a reference to an action that the target participated in, a connection to the follower’s visual flow, etc. The distribution and an example for each of these types is presented in Table 6.5.

From the expressions that did include a spatial relation, we observed a predominant use of projective relations, a smaller number of topological relations and only a few that are related to the follower’s path (see Table 6.6). Note that a single expression can use more than one relation, as in *the cabinet on the right closest to you*, but these cases were not very frequent in the corpus (10 mentions from the total expressions contained more than one spatial relation). A more detailed distribution of the spatial relations in the corpus is presented in Table 6.7. In a related study [Moratz and Tenbrink, 2006] of a WOZ\textsuperscript{16} experiment to direct a robot in real space through natural language conducted in German, similar distributions with this corpus data were observed. The German dialogs were typed instructions sent to a robot located in

\textsuperscript{16}a Wizard of Oz experiment is an experiment where people are lead to believe they interact with a system, but a human is producing the system’s response.
Table 6.6: The types of spatial relations observed in the corpus

<table>
<thead>
<tr>
<th>Type</th>
<th>Count</th>
<th>Percentage</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>projective</td>
<td>209</td>
<td>77.41%</td>
<td>front/behind/left/right, second/third</td>
</tr>
<tr>
<td>topological</td>
<td>52</td>
<td>19.26%</td>
<td>middle, close/far, next-to</td>
</tr>
<tr>
<td>path</td>
<td>9</td>
<td>3.33%</td>
<td>across, next</td>
</tr>
</tbody>
</table>

Table 6.7: Detailed counts of spatial relations observed in the corpus

<table>
<thead>
<tr>
<th>Spatial Relation</th>
<th>Count</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>left</td>
<td>89</td>
<td>that button on the left</td>
</tr>
<tr>
<td>right</td>
<td>52</td>
<td>the door on your right</td>
</tr>
<tr>
<td>front</td>
<td>42</td>
<td>the door straight ahead of you</td>
</tr>
<tr>
<td>second/third</td>
<td>19</td>
<td>the second one from the left</td>
</tr>
<tr>
<td>far/farther/farthest</td>
<td>14</td>
<td>the button that's farthest from the door</td>
</tr>
<tr>
<td>middle</td>
<td>13</td>
<td>the middle one</td>
</tr>
<tr>
<td>close/closer/closest</td>
<td>9</td>
<td>the one that is closest to you</td>
</tr>
<tr>
<td>next</td>
<td>6</td>
<td>the next door</td>
</tr>
<tr>
<td>in</td>
<td>5</td>
<td>the doorway that's in this room</td>
</tr>
<tr>
<td>next-to</td>
<td>5</td>
<td>the button next to that one</td>
</tr>
<tr>
<td>other relations (4 items)</td>
<td>&lt; 5</td>
<td></td>
</tr>
</tbody>
</table>

the real world. The authors noticed a small number of referring expressions that used more than one projective term, and a small number of modifications of the projective terms (e.g slight right).

Landmarks Distribution

Looking at the choice of landmarks, the corpus exhibited the distribution shown in Table 6.8. A relatively large number of expressions used a group as a landmark. Projective expressions that are relative only to the user’s position can be consider as having the user as landmark (or no landmark) as in [Kelleher and Kruijff, 2006].
Frames of Reference Distribution

Another aspect that we observed in the corpus and later used in our realization choices is the number of times the frame of reference is explicitly mentioned. Out of all the spatial relations, frames of reference other than the user’s were very rarely used\(^\text{17}\), even though some objects of interest are suitable for generating expressions using an intrinsic frame, for example cabinets have an intrinsic front. The frame of reference was explicitly mentioned in 98 cases, which represent 38% of the total spatial relations.

Of course, all these distributions are an artifact of the world design, and the tasks that were assigned, but it is interesting to see how much variation we have and to later compare these statistics with the results of the GRE algorithm.

6.3.2 Comparison

In this section we present the output obtained applying the GRE algorithm implemented in the system to generate REs in the same contexts where the First Locate expressions were produced in the corpus. These First Locate expressions identify the targets from the world (not anaphoric mentions), so the algorithm from Kelleher and Kruijff can be applied to obtain a referring expression. For the purposes of this

\(^{17}\)two case in the First Locate expressions, in the expressions the door on the...at the back of the room and the back corner right button, which use the room’s intrinsic back
comparison, we have to chose a moment to calculate the spatial context. We de-
cided to use the moment in time before the start of the RE to calculate the context.
Optimizations for this choice can be attempted by moving this moment in time or
averaging over time windows, but this was not attempted in the scope of this study.

From the 423 cases of First Locate, 5 cases happened to contain a reference to
an object that was not in the same room as the follower or in the doorway connected
to that room. We looked at these examples, and the follower was walking towards the
room containing the target, and was seeing through a straight line into that room.
The implemented system only describes objects in the current room, so it does not
have a mechanism to calculate the spatial context of objects in another room, so we
count them as failed.

Comparison Technique 1: Exact String Matching

The exact match on the entire string between the system output and the human
output is very low (6.8%), but we have to take into consideration that this is an overly
strict criteria. Even if the system is not producing the same string, it might still
produce a good referring expression, possibly expressing the same content. Examples
of REs from our generated output that are considered wrong by an exact matching
criteria are presented in Table 6.9. All of these examples are unambiguous from the
point of view of the rules implemented in the system. The GRE algorithm generates a
total of 82.8% distinguishing REs, and only 17.2% ambiguous. Ambiguous expressions
are generated when all the strategies that the system has access to for describing an
object produce ambiguous descriptions. In this case, the system will generate an
expression using a simple projective relation (front, back, left, right). Note that we
use the unambiguous and ambiguous labels as they are calculated with the system’s
implemented rules and perception of the environment, and not necessarily the same from a person’s point of view.

<table>
<thead>
<tr>
<th>Human Output</th>
<th>System’s Output</th>
<th>Exact Match</th>
<th>Content match (Dice score)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a cabinet on your right</td>
<td>the cabinet on your right</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>that cabinet</td>
<td>the cabinet in front of you</td>
<td>0</td>
<td>0.67</td>
</tr>
<tr>
<td>that doorway</td>
<td>the door</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>the door directly in front of you right now</td>
<td>the door in front of you</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>the one to the right</td>
<td>the door on your right</td>
<td>0</td>
<td>0.67</td>
</tr>
<tr>
<td>the one to the left of if</td>
<td>the second button from the left</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>the button on the right</td>
<td>the right button</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>the button in the middle</td>
<td>the middle button</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 6.9: Examples of system output judged wrong by exact matching

If we look separately at each of the **Determiner, Head** and **Modifier** values, the algorithm always uses *the* determiners and *common noun* heads, and only makes decisions about modifiers. Table 6.10 shows the precision obtained for each of these values, compared to a majority baseline. In this table, we show the precision for including a modifier (+/− values for modifier), and not on the content of the modifier, which is what the rule-based algorithm predicts. No statistical difference is obtained from the majority baseline, which will be an expression of the form *the + common noun + modifier*.\(^{18}\)

**Comparison Technique 2: Content Matching**

A comparison between the content returned by the automatic algorithm and the information used by people is still a strict criteria for evaluating the GRE algorithm’s

\(^{18}\)Notice that these statistics are only for the **First Locate** expressions, and for the entire corpus using the majority baseline would be much lower (20.0%), as in Section 4.2.3
Table 6.10: The correctness values for different features obtained by the GRE algorithm

<table>
<thead>
<tr>
<th></th>
<th>Determiner</th>
<th>Head</th>
<th>Modifier</th>
<th>All three values correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRE algorithm</td>
<td>53.2%</td>
<td>76.6%</td>
<td>69.03%</td>
<td>35.19%</td>
</tr>
<tr>
<td>Majority baseline</td>
<td>53.2%</td>
<td>76.6%</td>
<td>65.72%</td>
<td>34.04%</td>
</tr>
</tbody>
</table>

performance. The spatial relations used by the system are always correct (from the point of the rules implemented), and only in a small number of cases they are ambiguous. Direction givers sometimes use a different spatial relation because the follower is constantly moving, and his orientation changes fast, or simply because they decide on a different strategy for building the RE. Also, there might be a discrepancy between the moment in time we decide to calculate the RE (at the onset of the NP in the corpus data), and what moment humans use. We do not incorporate information from the past (such as direction of movement or speed), which might play a role in how people choose REs in this domain.

In the corpus data, we only have access to the surface form, and for some expressions it is not obvious to determine the content expressed. Content may interact with focus marking realized through *this/that* determiners or pronouns. The use of *that* can signal deixis, and the content would be equivalent with *the button in front of you*, or it can be that a different property made the object salient (such as *the door that just came into view now* or *the cabinet that is open*).

In the corpus, only 61% of the REs used a spatial relation, 34% did not have a modifier, and almost 5% used modifiers such as an action or a seeing event.

For the **First Locate** expressions, we annotated the content using a small number of categories: basic spatial relations (*front, back, left, right*), regardless if they are...
expressed relative to a group or the user, *middle, counting, close, far, and other.*

We also added a label *type* to indicate if the category of the objects was included in the surface form. This makes a total of 10 labels. A few expressions contained a modification of the spatial relation (such as *all the way on the left*, or *right/straight in front of you*), and they were assigned the basic spatial relation.

The Dice score has recently been used in the NLP field for assessing generation results, in the Attribute Selection for GRE Challenge. This score evaluates the content selection of GRE algorithms compared to human expressions that have been hand annotated with predications. Given two sets *a* and *b* of attributes included in the REs (one generated by the GRE algorithm and one extracted from human expressions), the Dice score is calculated using the formula:

$$dice(a, b) = \frac{2 \cdot |a \cap b|}{|a| + |b|},$$

where $|a|$ represents the cardinality (number of elements) in the set *a*.

The Dice score gives partial credit for outputs that contain some of the attributes included in the gold standard. For example, the phrase *the door just to your left there that’s closed* will be labeled with the content *type + left + other* (because of *door + to your left + closed*), and a system output of the form *the door on your left* (with the content *type + left*) would receive a score of $4/5 = 0.8$.

The ASGRE Challenge held in the summer of 2007 has proposed several criteria for evaluation, such as unique identification, minimality and humanlikeness. To be able to assess the unique identification and minimality of an expression, further human annotation is needed. From the point of view of the rules implemented in our algorithm, the REs produced are uniquely identifying the referent 82.8% of the

---

19 see reporting of systems performance at http://www.itri.brighton.ac.uk/ucnlg/
times and are always minimal. In the following paragraphs we will report on the humanlikeness of our GRE output, which measures the similarity between the content of the system’s output and of spontaneous human productions.

Using a majority baseline for content selection, we obtain a score of 60.87% by assigning no content to the expression (except the type information). The algorithm always uses the type of object in lexicalizing the expression, and Table 6.3 shows how many of the corpus REs had lexicalized this information. The Dice score for the GRE algorithm is not much better: 61.20%. It is clear that always including no extra information would not have helped our human participants to complete the tasks, so we needed to refine these measurements to elucidate the algorithm’s contribution.

In our domain, the GRE algorithm overgenerates front relations. Humans have a way to lexically mark focus, for example by using this/that markings or prosody, but how to include salience information in the lexical choice of the GRE algorithm has not yet been formalized. Salience information has been included in the algorithm proposed in [Kelleher and Kruijff, 2006] for ordering the set of distractors and to stop the generation process when a description is not yet distinguishable, but the target’s salience is greater than the salience scores assigned to all the distractors. Because our domain is different and we did not want to rely on visual salience for making a description distinguishing (and maybe introduce ambiguity), we did not use this in our algorithm. Still, the lexical form produced by The Locative Incremental Algorithm [Kelleher and Kruijff, 2006] would be of the form the+type, with no focus

\[20\] we observed the algorithm is over-generating front relations when people use a focus marking, but this is not a violation of minimality, as the algorithm doesn’t have this strategy for marking REs, and using its available strategies chooses the front relation.
indication. From the 150 expressions that were labeled with no extra content\textsuperscript{21}, only 16 did not have a focus marking (\textit{this/that} determiners or pronouns). It is left as a point of future work how to indicate where is appropriate to use this focus, because the property selected and the focus assigned are interdependent.

Generating a different spatial preposition (e.g. the system says \textit{right} at a point where the human speaker used \textit{left}) is clearly a greater difference than adding a \textit{front} relation on a focused item. If we equate the \textit{front} relation with no extra content except type information, the Dice score obtained is 69.31\%. This figure is comparable with the numbers reported in the GRE Selection Challenge. The scores there ranged between 52.7\% and 77.1\% on the humanlikeness scale with an average of 68.9\% for all 22 systems. We consider that our domain is more difficult, since the datasets used for the challenge were one time descriptions of different furniture items or people pictures presented on a computer screen, and not in a dialog conducted in a simulated world where participants move about.

The GRE algorithm we implemented is not using any non-spatial attributes (except the type of object), and the content expressed by focused expressions is not clear. To make a more detailed assessment of our algorithm’s behavior, we only looked at the subset of expressions that do include a spatial modifier (and possibly other type of modifiers). There were 258 elements in this set. The majority baseline in this case would be assigning the content \textit{left + type}, with a Dice score of 56.25\%, compared to 62.8\% for the GRE algorithm. This difference is significant only at 0.1 level, but if we exclude the \textit{type} information from our calculations and focus only on the spatial

\textsuperscript{21}Notice that some of the modifiers do not bring any content in the labeling scheme we devised, such as \textit{the button right in the wall}, since all buttons are on walls, or \textit{the door right there}, since we do not know the location the speaker was intending
Dice Scores for Content Selection

<table>
<thead>
<tr>
<th>All First Locate expressions (423 items)</th>
<th>Dice Scores for Content Selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority baseline - type - adding no extra content</td>
<td>60.87%</td>
</tr>
<tr>
<td>GRE algorithm</td>
<td>61.20%</td>
</tr>
<tr>
<td>GRE algorithm considering no extra content and front relations equivalent</td>
<td>69.31%</td>
</tr>
<tr>
<td>Only First Locate expressions that include spatial content (258 items)</td>
<td></td>
</tr>
<tr>
<td>Majority baseline - left + type</td>
<td>56.25%</td>
</tr>
<tr>
<td>GRE algorithm</td>
<td>62.80%</td>
</tr>
</tbody>
</table>

Table 6.11: The content selection performance of the GRE algorithm

properties added, the difference between the baseline Dice score (33.45%) and the system’s score (44.96%) becomes statistically significant (p=0.005).

Table 6.11 summarizes the results of comparing the content included in the human expressions with the content selected by the rule-based GRE algorithm. Even though at a first glance the algorithm is not performing much better than a majority baseline, we consider that it makes a lot of similar choices to the human content selection procedure, especially in the cases of selecting spatial relations. This study also raises an interesting research question about judging the content of focused REs, present with a higher frequency in spoken data than in written descriptions.

Comparison Technique 3: Distributional Similarity

Spatial Relations

Looking at the spatial relations used by the system, we saw that it generated only 63 expressions that do not use any spatial relation (e.g. the door, in cases where there is only one door in the current room), compared to 145 expressions in the corpus that contained no modifier. It failed in 5 cases, and used spatial relations on the type is always added by both the baseline and the GRE algorithm.
remaining 355. The distribution of spatial/non-spatial descriptions in the system’s output is presented in Table 6.12.

Table 6.13 presents the assignment of different spatial relations made by the generation algorithm. The distribution in the human corpus was repeated here for ease of reference. Overall, we observe that the system uses left/right, middle, second/third from the left and closer/farther relations in a proportion very similar to humans, but front relations are more frequent. We explained our hypothesis between the focus assignment in the human data and front relations used by the GRE algorithm in the Content Matching section.

<table>
<thead>
<tr>
<th>Spatial Relation</th>
<th>GRE Count</th>
<th>Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>front</td>
<td>162</td>
<td>42</td>
</tr>
<tr>
<td>left</td>
<td>75</td>
<td>89</td>
</tr>
<tr>
<td>right</td>
<td>50</td>
<td>52</td>
</tr>
<tr>
<td>farther/closer</td>
<td>34</td>
<td>23</td>
</tr>
<tr>
<td>second/third</td>
<td>17</td>
<td>19</td>
</tr>
<tr>
<td>middle</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>behind</td>
<td>4</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 6.13: The types of spatial relations observed in the corpus

**Landmarks**

Compared to the human output, the system output contains 65 references that used a group as a landmark vs. 82 in the human output (15.3% compared to 19.39%
from total expressions, difference significant only at 0.1 level), and 33 compared to
32 mentions that were related to another landmark in the scene (8.0%, respectively
7.56% out of total 423 expressions). Table 6.14 presents the distribution of landmarks
in the algorithm output (the distribution in the human data presented in Table 6.8
was repeated for ease of reference). These are very small counts, but the system’s
distribution in choosing landmarks is similar to the human data. The smaller number
of group landmarks might be a consequence of our system’s ability to identify just
a subset of the groupings selected by humans, and its choice of possible landmarks.
Collecting more data and refining the model for determining groupings/landmarks
might also influence our ordering of landmark selection.

The performance of the GRE algorithm will be further elucidated by looking
at the task completion rates and times to identify a target in the human-computer
interaction presented in Chapter 7.

6.4 Grounding

In this section we present the choices we made for implementing the grounding
behavior of the system. Grounding represents the collective process by which the
dialog participants try to reach a mutual belief about what has been understood [Clark
and Brennan, 1991]. During interaction the user does not give any spoken feedback,
so the system needs to observe his actions to assess if the commands have been

<table>
<thead>
<tr>
<th>Landmark</th>
<th>GRE Count</th>
<th>Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>group</td>
<td>65</td>
<td>82</td>
</tr>
<tr>
<td>other object</td>
<td>33</td>
<td>32</td>
</tr>
<tr>
<td>all other spatial relations</td>
<td>257</td>
<td>144</td>
</tr>
</tbody>
</table>

Table 6.14: The landmarks used by the GRE algorithm
understood. As the user moves freely in the world and completes each command, the system will produce an acknowledgment that the current step has been completed. To complete this task we need to track the user’s position in the world with a high frequency so the system can react fast to his behavior, and we need to know what we are looking for to consider each plan step completed.

6.4.1 Tracking User Position and Interpreting Actions

The Quake game engine keeps track of the user’s position and orientation, and can be queried at any time. It will keep a record of user’s information at every 10th of a second. To check the completion of a turn, our system keeps track of the effects in the world that it expects after each of the issued commands. For example, in the case of a command to push a button, we can check the events generated in the quake world and wait for that button to get pushed. For a command to get through a door, we need to wait for the coordinates of the user to become included in the area for the next room that the door was leading to. For picking up or dropping an object, we need to check that the user’s inventory of objects has changed accordingly. The system enters a loop that checks the world until the conditions are met, or a set amount of time has passed. In our case, we wait for 6 seconds after a command was issued for its completion, and check at every 10th of a second. Reacting fast to the user completion of the step is important, but too frequent checks of world conditions make the system slow.

6.4.2 Timing Issues in a Situated Natural Language System

An ideal Instruction Giving System should have a way to opportunistically find good moments to describe objects, by either waiting until the person is in a good
context, predicting future trajectory and coordinating the time when the description will be ready to be spoken with the actual context, or trying strategies to get the user in better contexts (see our human study in Chapter 7). But there are many problems in timing the speech. For example, we observed that there should be a way to interrupt the TTS, and maybe a way to synthesize parts of the turn and wait until the very last moment to calculate the spatial relation.

For example, the system waits 6 seconds after giving the user a command, and then re-enters the loop to generate a second attempt at 6.01 seconds. If the user in fact completed the command after 6.01 seconds, he will hear the second command way after he already finished the action, and confusion may arise (was the system, referring to the door I just went through, or is it maybe referring to the next door that is now straight ahead?). The way the system works now is to check grounding only after it fully generated the command, so it will produce the acknowledgment only then. Another strategy to stop and acknowledge the completion of the plan step sooner might work better.

Also, tracking intention and observing that the user is moving in the right way could influence the waiting time between attempts, so that the system is not planning to regenerate the turn when the person seems to be really close to finishing the turn, but rather extends the waiting time a couple of seconds.

Another interesting example of people timing their productions observed in the corpus is that the Direction Giver does not always know what he will say until the moment of the RE. That is, he will start the sentence without knowing what will follow, and will fill in the right spatial relation for the target taking in account the follower’s position at that exact time, and not when he started to speak. That is
sometimes seen in the data because of the pause before the RE, or through some lexical evidence, such as in Figure 6.10. That kind of behavior in a dialog system is hard to replicate, as the speech synthesizer output sounds best when it is allowed to plan prosodic contours over an entire sentence. Generating chunks and sticking them together will not produce the right prosody and will most likely introduce perceivable pauses before the REs (to have time to calculate the spatial position and to synthesize the result). Another solution would be to have all the possible turns precomputed and produced with the right prosody and later sliced-up at the start of the REs. The system would have to chose the right continuation of the turn at the latest moment it can, but synchronizing the play back of the two wav files and the GRE module decision has not been attempted within the scope of this thesis.
CHAPTER 7

AN EXPERIMENTAL STUDY IN HUMAN-COMPUTER INTERACTION

In Chapter 6 we described the details of a direction giving system that can guide a person in completing several simple manipulation tasks in an interior virtual reality setting. The system is aware of the user’s spatial orientation and of the world attributes. In this chapter, we will present the results obtained using this system in a human evaluation study.

7.1 Testing the System’s Referring Behavior in Relation to the Context: Study Design

The research question that we are addressing in this human-computer study is the relation between context complexity and users’ ratings/automatic measurements for two systems implementing different referring strategies. The first system creates a description for the target independent of the difficulty of the context (henceforth the Immediate Description condition), while the second system adjusts the context before creating a description to get the user into an easier context (the Adjusted Context condition). Human subjects interacted with the two systems, and qualitative and quantitative measures were collected from the user surveys and the logs generated by the systems.
The system in the Immediate Description condition produces a description the moment the plan step arrives to the generation module, regardless of the difficulty of the context (number of distractors, angle or distance information). If the instruction is not completed in a particular amount of time, a new description will be generated. The system tries two times to describe a target ID, then exits with failure (see Figure 7.1 for a diagram of the system’s strategy). The Refer state can be exited on the OK branch when the action was completed and grounding has returned success, or on the Fail branch when the effects of the turn have not been completed and the allocated time has elapsed.

In the Adjusted Context condition we will be using the decision tree learned in our previous study (see Figure 4.5) to decide if a context is appropriate for a description. If the context is not good, the system will try (for a maximum of two tries) to move the direction follower to a better location by issuing a simple redirect command (Turn left, Turn right, Go forward), and only afterwords will produce the description. The redirect command is chosen from the condition that failed in the decision tree. The dialog strategy of this system is presented in Figure 7.2. Notice that if the decision tree considers this is a good moment to describe, then no extra turn is produced, and we advance on the OK branch of the Adjust state, just as in the case where a redirect command was generated and the appropriate reaction from the user was grounded and acknowledged.

Both systems have a very small vocabulary and a small set of possible system turns (67 vs. 70), presented in Appendix E.

We blocked out the tasks and took into account the order effects that will appear while users interact with the two systems. Table 7.1 shows the order in which the two
systems were presented to the users, and what tasks each system was assigned. Tasks were of different difficulty (computed as the number of plan steps that include referents from the world). Half of the subjects interacted with the Immediate Description system first, and all the possible orders between the given tasks were included (with the constraint of keeping the one easy task always first). We always started with one of the easier tasks (either moving the picture - P, or moving the boxes - B, which both include just 3 descriptions of IDs) and used all possible combinations from the more difficult tasks: moving the Quake Logo from cabinet C7 to C14 (Q), moving the Rebreather from C5 to C4 (B), and moving the Silencer from C12 to C15 or C9 (S, S2). These harder tasks had a maximum of 23 steps per task and a minimum of 14. The subjects received information about their assigned tasks to be completed with the systems through the pictures presented in Appendix C, Figure C.2
<table>
<thead>
<tr>
<th>Subj. ID</th>
<th>System Order</th>
<th>Task Order</th>
<th>Subj. ID</th>
<th>System Order</th>
<th>Task Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Immediate-Adjusted</td>
<td>PQSBRS2</td>
<td>17</td>
<td>Adjusted-Immediate</td>
<td>PSQBR52</td>
</tr>
<tr>
<td>2</td>
<td>Immediate-Adjusted</td>
<td>BQSPRS2</td>
<td>18</td>
<td>Adjusted-Immediate</td>
<td>BSQPR52</td>
</tr>
<tr>
<td>3</td>
<td>Immediate-Adjusted</td>
<td>BRSPQS2</td>
<td>19</td>
<td>Adjusted-Immediate</td>
<td>BS2RPQS</td>
</tr>
<tr>
<td>4</td>
<td>Immediate-Adjusted</td>
<td>PRS2BQS2</td>
<td>20</td>
<td>Adjusted-Immediate</td>
<td>PS2RBQS</td>
</tr>
<tr>
<td>5</td>
<td>Immediate-Adjusted</td>
<td>PQS2BRS2</td>
<td>21</td>
<td>Adjusted-Immediate</td>
<td>PS2QBR5</td>
</tr>
<tr>
<td>6</td>
<td>Immediate-Adjusted</td>
<td>BQS2PRS2</td>
<td>22</td>
<td>Adjusted-Immediate</td>
<td>BS2QPR5</td>
</tr>
<tr>
<td>7</td>
<td>Immediate-Adjusted</td>
<td>BRSPQS2</td>
<td>23</td>
<td>Adjusted-Immediate</td>
<td>BSRPQS2</td>
</tr>
<tr>
<td>8</td>
<td>Immediate-Adjusted</td>
<td>PRSBQS2</td>
<td>24</td>
<td>Adjusted-Immediate</td>
<td>PSRBQS2</td>
</tr>
<tr>
<td>9</td>
<td>Immediate-Adjusted</td>
<td>PSQBS2R</td>
<td>25</td>
<td>Adjusted-Immediate</td>
<td>PSQBS2R</td>
</tr>
<tr>
<td>10</td>
<td>Immediate-Adjusted</td>
<td>BSQPS2R</td>
<td>26</td>
<td>Adjusted-Immediate</td>
<td>BSQPS2R</td>
</tr>
<tr>
<td>11</td>
<td>Immediate-Adjusted</td>
<td>BS2RPSQ</td>
<td>27</td>
<td>Adjusted-Immediate</td>
<td>BR2PSQ</td>
</tr>
<tr>
<td>12</td>
<td>Immediate-Adjusted</td>
<td>PS2RBSQ</td>
<td>28</td>
<td>Adjusted-Immediate</td>
<td>PS2RBSQ</td>
</tr>
<tr>
<td>13</td>
<td>Immediate-Adjusted</td>
<td>PS2QBSR</td>
<td>29</td>
<td>Adjusted-Immediate</td>
<td>PQS2BSR</td>
</tr>
<tr>
<td>14</td>
<td>Immediate-Adjusted</td>
<td>BS2QPSR</td>
<td>30</td>
<td>Adjusted-Immediate</td>
<td>BS2QPSR</td>
</tr>
<tr>
<td>15</td>
<td>Immediate-Adjusted</td>
<td>BSRPS2Q</td>
<td>31</td>
<td>Adjusted-Immediate</td>
<td>BRSPS2Q</td>
</tr>
<tr>
<td>16</td>
<td>Immediate-Adjusted</td>
<td>PSRSB2Q</td>
<td>32</td>
<td>Adjusted-Immediate</td>
<td>PRSBS2Q</td>
</tr>
</tbody>
</table>

P=move the picture, B=move the boxes, Q=hide the Quake Logo, R=hide the Rebreather, S/S2=hide the Silencer

| Table 7.1: | The order that the tasks were presented to the human subjects |

7.2 Study Materials

The materials used in this study are presented in Appendix C. The participants received a set of instructions (Figure C.1), information about the possible tasks (Figure C.2) and the controls for the video game (Figure A.3). The experiments were run on an iMac G5, with 1.8 GHz processor and 2GB of RAM, running MacOS X.

We recruited a total of 32 human subjects, from which 24 were male and 8 female. The participants were all native speakers of American English, with an average age of 26.

At the end of the study, several surveys were given to the participants, presented in the Appendix C, Figures C.3, C.4 and C.5. The surveys related to system performance were shown at the end of each interaction (consisting of 3 tasks) with a system.
Some of the questions for determining the differences between how people perceived the interaction with the two systems were informed by the human evaluation of the COMIC system presented in [White and Foster, 2004].

<table>
<thead>
<tr>
<th>Question</th>
<th>Average</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) How would you rate your expertise as a computer user? (1=very poor - 5=very good)</td>
<td>4.25</td>
<td>0.73</td>
</tr>
<tr>
<td>2) How many hour/week do you play computer games?</td>
<td>2.52</td>
<td>3.26</td>
</tr>
<tr>
<td>3) How did you feel about navigating in the Virtual World? (1=very difficult - 5=not difficult at all)</td>
<td>4.28</td>
<td>0.92</td>
</tr>
</tbody>
</table>

Table 7.2: General survey questions

The questions in the general survey and the average rankings are presented in Table 7.2. Participants ranked high in computer expertise, and felt the navigation in the virtual world was easy. Participants had various levels of previous computer game experience, from playing no computer games, to playing up to 10 hours per week.

7.3 Results Obtained from Human Ratings and System Logs

7.3.1 Qualitative Metrics: User Survey Results

We have collected statistics from the human subjects and computed a pair-wise t-test to determine the significance of the results obtained from the questions related to the evaluation of the two systems.\(^{23}\) The human participants were encouraged to look at the scorings they gave to the first system while completing the evaluation.

\(^{23}\)We made a correction in the system by removing the fill pauses between system turns (\textit{then, and then, now}), because this was slowing down the task progress. Six participants saw this system (this change affected both conditions, and was balanced in between which condition was seen first). This resulted in very small changes in the average scores obtained excluding this data, and none of the dimensions tested changed their significance.
of the second system, to allow them to calibrate their responses based on previous ratings.

The questions related to the performance comparison between the two systems are presented in Table 7.3, together with the average value given to the system in each condition, and the $p$ value obtained from the pair-wise t-test. Values that are statistically significant are indicated in bold.

<table>
<thead>
<tr>
<th>Question</th>
<th>Immediate</th>
<th>Adjusted</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) How would you rate the average ease or difficulty of the tasks you completed? (1=very difficult 5=very easy)</td>
<td>3.75</td>
<td>3.94</td>
<td>0.129</td>
</tr>
<tr>
<td>2) How would you rate the intelligibility of the system’s voice? (1=very poor 5=very good)</td>
<td>3.94</td>
<td>4.16</td>
<td>0.015</td>
</tr>
<tr>
<td>3) How would you rate the naturalness of the language used by the system? (1=very poor 5=very good)</td>
<td>3.16</td>
<td>3.25</td>
<td>0.076</td>
</tr>
<tr>
<td>4) How easy was it to follow the directions given by the system? (1=very difficult 5=very easy)</td>
<td>2.94</td>
<td>3.63</td>
<td>0.00018</td>
</tr>
<tr>
<td>5) Did you feel the directions were - too informative (number of “yes” answers)</td>
<td>0</td>
<td>5</td>
<td>0.011</td>
</tr>
<tr>
<td>- not informative enough (number of “yes” answers)</td>
<td>24</td>
<td>10</td>
<td>0.00003</td>
</tr>
<tr>
<td>6) How easy was it understand the system’s descriptions of different objects in the world? (1=very difficult 5=very easy)</td>
<td>3.88</td>
<td>3.84</td>
<td>0.093</td>
</tr>
<tr>
<td>7) How would you rate the friendliness of this system? (1=very low 5=very high)</td>
<td>3.38</td>
<td>3.63</td>
<td>0.011</td>
</tr>
<tr>
<td>8) I found the system to be cooperative during the conversation. (1=not agree 5=agree)</td>
<td>3.5</td>
<td>3.63</td>
<td>0.041</td>
</tr>
<tr>
<td>9) I found the conversation engaging. (1=not agree 5=agree)</td>
<td>2.66</td>
<td>2.78</td>
<td>0.08</td>
</tr>
<tr>
<td>10) I found it exciting to interact with the system. (1=not agree 5=agree)</td>
<td>2.84</td>
<td>3.03</td>
<td>0.015</td>
</tr>
<tr>
<td>11) I was satisfied with the system’s ability to help me complete the task. (1=not agree 5=agree)</td>
<td>3.03</td>
<td>3.66</td>
<td>0.00095</td>
</tr>
</tbody>
</table>

Table 7.3: The System related survey questions and the average ratings obtained in the two conditions (significant differences are indicated in bold)
The ratings showed that the system in the Adjusted condition was perceived as *too informative* by only 5 participants, while the one in the Immediate condition was perceived as *not informative enough* by 24.

Even though the overall number of tasks that were successfully completed is not very different (see next subsection, Table 7.6), the system in the Adjusted condition ranked higher on the perceived *ability to help complete the task*.

The directions given by the system in the Adjusted condition were considered *more easy to follow* than those of the baseline system. The Adjusted condition ranked better than the Immediate condition also on the *excitement* level, where people gave it an average score of 3.03 vs. 2.84. On the question of the system’s *friendliness* and *cooperativeness*, the Adjusted condition system again ranked higher.

There were no significant differences between the two systems on the *engagement* dimension or on the *naturalness of the language*. Also, the *descriptions* in themselves were considered equally easy to understand in both systems.

Appropriately, the tasks to be completed were considered not difficult at all, and the difference between systems was not statistically significant. A dimension that we did not expect to find statistical differences was the intelligibility of the system’s voice, but people considered that the voice in the Adjusted condition was a little better. We used the same voice to synthesize the turns in both systems, but it is possible that some of the language produced in the online interaction by the system in the Adjusted Context condition sounded more natural, and that produced this preference. The average scores assigned to the two systems are relatively high: 3.94 and 4.16, showing that understanding the instructions was not impeding the subjects’ interaction with the systems.
On the final survey (Figure C.5), people were asked if they saw any difference between the systems and if they had any preference. Out of the total 32 participants, 21 chose the system in the Adjusted condition (see Table 7.4).

During interaction, the users are expected to become better at understanding the synthetic voice\textsuperscript{24}, at navigating through the virtual environment, and at figuring out the strategy used by the system. The data shows a preference for the system seen second, as the subjects preferred it 17 times over the one they saw first. This difference is though smaller than the preference difference we noticed between the two conditions we tested. The System in the Adjusted Context condition was chosen 9 times when it was presented first, while the system in the Immediate Description condition was chosen only 2 times.

<table>
<thead>
<tr>
<th></th>
<th>When Adjusted Context Was Seen First</th>
<th>When Immediate Description Was Seen First</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjusted Context</td>
<td>9</td>
<td>12</td>
<td>21</td>
</tr>
<tr>
<td>Immediate Description</td>
<td>5</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>No Preference</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>16</td>
<td>16</td>
<td>32</td>
</tr>
</tbody>
</table>

Table 7.4: The participant’s choice of system

7.3.2 Quantitative Metrics: Obtained Automatically from the System’s Log

The following quantitative metrics are automatically calculated from the logs that the systems produce at the end of each round of interaction. The system knows what type of referent it produced, how much time it spent on each step in the interaction,

\textsuperscript{24}We used AT&T text to speech engine, available at http://www.naturalvoices.att.com/, to generate the system prompts, using the Crystal voice.
what was the configuration of the world, where the user was located, which tasks were presented and if the tasks were successful or failed.

Both systems will produce sometimes ambiguous descriptions, as they only have a limited range of strategies to form the descriptions and the users are moving unrestricted in the virtual environment space. Still, people were able to successfully finish 65.6% of the tasks, and were able to deal with the ambiguity and adapt to the system’s capabilities.

<table>
<thead>
<tr>
<th></th>
<th>Immediate condition</th>
<th>Adjusted condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Time per task</td>
<td>57.1 s</td>
<td>71.6 s</td>
</tr>
<tr>
<td>(96 tasks with each system)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average Time for successfully</td>
<td>89.2 s</td>
<td>106.3 s</td>
</tr>
<tr>
<td>completed tasks</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 7.5:** Average time spent per task

Task performance was better for the system in the Adjusted Context condition, but the overall difference is significant only at 0.1 level (58 tasks completed in the Immediate condition vs. 68 tasks completed in the Adjusted Context condition). From the total of $3 \times 32 = 96$ tasks with each system, the completion rates for both systems are similar: 60.4% for the Immediate Description and 70.8% for the Adjusted Context one. Table 7.6 shows the distribution of completed tasks for the two systems across all the 6 possible tasks.

The total time that was spent interacting with the system in the Adjusted Context is obviously longer, as the system produces more turns and waits for those turns to be grounded, but the difference is not large at all. The average time spent per task with it is around 15 seconds longer than with the Immediate Description system (see Table 7.5), and just a little longer for the successfully completed tasks. This is a
relative increase in time of almost 25%, but it is not noticeable, as the average time per task (completed or not) was only 1 minute in the Immediate Description system.

<table>
<thead>
<tr>
<th>Task</th>
<th>Plan Steps</th>
<th>Immediate condition</th>
<th>Adjusted condition</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>P=picture</td>
<td>3</td>
<td>16</td>
<td>16</td>
<td>32 (100%)</td>
</tr>
<tr>
<td>B=move the boxes</td>
<td>3</td>
<td>14</td>
<td>16</td>
<td>30 (94%)</td>
</tr>
<tr>
<td>S2=hide the Silencer (2)</td>
<td>19</td>
<td>9</td>
<td>11</td>
<td>20 (62.5%)</td>
</tr>
<tr>
<td>S=hide the Silencer (1)</td>
<td>18</td>
<td>8</td>
<td>11</td>
<td>19 (59%)</td>
</tr>
<tr>
<td>R=hide the Rebreather</td>
<td>23</td>
<td>8</td>
<td>8</td>
<td>16 (50%)</td>
</tr>
<tr>
<td>Q=hide the Quake Logo</td>
<td>14</td>
<td>3</td>
<td>6</td>
<td>9 (28%)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>58 (60.4%)</strong></td>
<td><strong>68 (70.8%)</strong></td>
<td><strong>126 (65.6%)</strong></td>
<td></td>
</tr>
</tbody>
</table>

**Table 7.6:** Task completion per task (each task was presented 16 times)

The systems produced a similar number of ambiguous descriptions\(^{25}\) in both cases (143 in Immediate Description condition vs. 126 in the Adjusted Context) and a little more unambiguous descriptions in the Adjusted Context condition (734 vs. only 623 in the Immediate Context).\(^{26}\) We calculated the time it took to ground different types of references: unambiguous, ambiguous, first mentions or repetitions. By time to ground, we consider the elapsed time that the system spent waiting for the person to complete the condition associated with the spoken turn. For example, in the case of a command to push a button, the time it took from the system finishing uttering the command until the system got a message from the virtual world that the button was pushed. The obtained times are very similar, but we observed that it takes 2.23 seconds on average in the Immediate Description condition to resolve a first

\(^{25}\)ambiguous from the point of view of the rules implemented in the NP generation module and at the moment of planning the speech

\(^{26}\)these are only the descriptions produced by the algorithm. The system produces more Unambiguous descriptions that are grounded that do not refer to buttons, doors or cabinets, such as the stairs, the treasures, the tunnel, etc.
mention unambiguous reference, vs. 1.81 seconds with the system in the Adjusted condition (see Table 7.8). The standard error for all the Successful Unambiguous 1st Try references for the Immediate Description system is 0.076 and 0.072 for the Adjusted Context, making the difference between grounding this references in the two systems significant. This might be a signal of a clearer description in this system, but it can also be a consequence of the system’s strategy (the person is brought in some cases closer to the referent). When resolving second mentions, for both ambiguous and unambiguous, grounding was also faster in the Adjusted Context condition. Note that the time to ground an Adjust command is not taken in account into the reference times, and is presented separately. As we shown in Figure 7.5, completing tasks with the system in the Adjusted condition took more time.

<table>
<thead>
<tr>
<th></th>
<th>Immediate Description</th>
<th>Adjusted Context</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>not understood</td>
<td>successful</td>
</tr>
<tr>
<td>Unambiguous</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st Try</td>
<td>75</td>
<td>709</td>
</tr>
<tr>
<td>Retry</td>
<td>12</td>
<td>102</td>
</tr>
<tr>
<td>Total</td>
<td>87</td>
<td>811</td>
</tr>
<tr>
<td>Ambiguous</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st Try</td>
<td>44</td>
<td>46</td>
</tr>
<tr>
<td>Retry</td>
<td>23</td>
<td>25</td>
</tr>
<tr>
<td>Total</td>
<td>67</td>
<td>71</td>
</tr>
<tr>
<td>Total REs</td>
<td>154</td>
<td>882</td>
</tr>
<tr>
<td>Adjust</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total Grounded</td>
<td>154</td>
<td>882</td>
</tr>
</tbody>
</table>

**Table 7.7:** The different types of referents and their counts in the two systems

With a small difference (0.922 vs. 0.817 seconds average time), it took longer to ground first mentions that were ambiguous in the Adjusted condition. This is a rather unexpected result, as we would have thought that it would take less time to
<table>
<thead>
<tr>
<th>Average time spent grounding a:</th>
<th>Immediate Description (sec)</th>
<th>Adjusted Context (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Successful Unambiguous 1st Try</td>
<td>2.23</td>
<td>1.81</td>
</tr>
<tr>
<td>Successful Unambiguous Retry</td>
<td>0.68</td>
<td>0.65</td>
</tr>
<tr>
<td>Successful Ambiguous 1st Try</td>
<td>0.817</td>
<td>0.92</td>
</tr>
<tr>
<td>Successful Ambiguous Retry</td>
<td>0.26</td>
<td>0.18</td>
</tr>
<tr>
<td>Successful Adjust Commands</td>
<td>0</td>
<td>0.56</td>
</tr>
</tbody>
</table>

Table 7.8: The average time it took to ground different types of referents in the two systems

ground ambiguous descriptions too with this system, because they are produced in a presumably better context. The difference is very small, and one possible explanation could be that with the system in the Adjusted Context, possibly after a redirect command, people might have spent a little more time trying to guess which referent the system intended to describe.

The average time it took to ground a redirect turn is comparable with the average time it took to resolve an ambiguous description in the first condition.

Table 7.7 presents the number of different types of referents produced by the two systems. During the entire interaction, the decision tree that evaluates if the context should be adjusted or not has resulted in an Adjust command 590 of the times, and has evaluated the context as appropriate for description 602 of the times. Table 7.8 presents the average time it took for the system to ground different types of referents. The time that was spent grounding a missed reference is 6s, a threshold we design the system to wait for an action to be completed before speaking the next command.

Our subjects had different expertise levels in playing video games or navigating virtual worlds. We informally noticed that the more expert users were sometimes frustrated with the speed of interaction of the Adjusted context system. A user
model that considers the user preferred strategy and level of expertise is appropriate in our domain.

7.4 Discussion

The qualitative metrics show a clear preference for the Adjusted Context condition. The quantitative metrics show that even though the overall average time was longer and the improvement in the task completion was not very large, the system had the opportunity to produce less ambiguous language (with the same basic rule-based GRE algorithm), grounding times of referents were shorter. The system in the Adjusted condition also had better success rates for harder tasks.

This study supports the idea that natural language generation in situated contexts should be approached as a whole, taking into account the dynamics and possibility of movement, and not as a static procedure. It would be ideal if a system could generate an unambiguous description in any context, but often this expression is complicated and not natural. People have been observed to manipulate the context and break down the goal of referring to a target in multiple steps.

The exciting result from this chapter is that although GRE is consistently modeled as a static process, our study found that dialog partners consider the dynamics of the situation in producing REs and also human users interacting with a system prefer a system that is considering the context before producing an referring expression.
The problem of situated language and in particular situated noun phrase generation gives rise to a wealth of interesting research problems, and this dissertation provides answers only to a small subset. The interaction model we developed and the data collection we have completed left us with a lot of interesting observations about human referring process that deserve further investigation, and we will mention a few in the following paragraphs.

We observed an interesting behavior in the case of objects that are part of a group. Even though the group is not lexicalized in the noun phrase referring to the object, it is sometimes previously lexicalized in a sentence that draws attention to the group. In most of those case, the head of the noun phrase changes from the type of object to one. Examples from the corpus are presented in Figure 8.1.

These kind of references are similar to bridging references, because they refer to a part of an entity (the group) that has been mentioned earlier, but they can sometimes lexically overlap with their antecedent. This type of reference has not been extensively studied in the field of generation. For interpretation, [Poesio et al., 2004] presents a machine learning technique, but this paper is looking at resolving the reference
of subsequent, non lexically overlapping bridging references, and not for generating them.

It would be interesting to determine a special strategy for generation that introduces the objects that are part of a group, by first generating a turn to bring the grouping into focus, and then identifying the target inside of the group, possibly through a one NP (see Figure 8.1).

In the situated noun phrases produced by humans in the corpus we observed the need of temporal indexing on the property of the NP for a correct resolution. For example, in Figure 8.2, the expression a door on your right is produced when the door is straight in front of the Follower, but it would have been on the right immediately as the person was getting down the stairs.
Session 10, 3 min 3 sec

DG: after you come down the stairs there should be a door on your right
[ D12 - firstLocate ]

DF: mhm

DG: uh go through that

Session 10, 0 min 25 sec

DG: hit that button ... that’s not the right button ... go to your right the back corner right right button ... that should open the cabinet

DF: let me see maybe i didn’t pop it right

DG: here it goes that’s open a cabinet on your back
[ C7 -firstLocate ]

Figure 8.2: Examples from the corpus where the property needs temporal indexing

In Figure 8.2, in the second dialog fragment there is another example, where by the time the Direction Giver starts the expression a cabinet on your back the cabinet is actually right in front/right area of the Follower. But the follower knows he just turned, and that was the cabinet in his back, and no problem in communication occurs, so Direction Followers are robot to their own movements in interacting with a human partner.

These noun phrases contain spatial information, but to properly resolve them we need to project the future position of the person in the world or to remember past configurations. So Direction Givers are referring to spatial relations as they were or will be at another time during the dialog. Similar effects have been signaled in research about the temporal aspect of noun phrases conducted by Prof. Judith Tonhauser, see [Tonhauser, 2005], where interpreting a property on a noun phrase has connection with the time that property is assumed to hold. The influence of the time of speech and the time that the property holds also transfers for NP generation, and it is interesting to observe how much people accommodate for this discrepancy in a
computer system. As we informally noticed in our human-computer interaction study, people were pretty good at figuring out that the system planned the expression as the spatial configuration was at the beginning of the turn and picked the right referent, even though it was not always matching the spatial property at the end of the spoken turn.

So far, Systems A and B have the same algorithm for describing objects, which only uses the determiner always. Previously in a study of a similar domain [Byron and Stoia, 2005], we noticed the small presence of this pronouns and determiners in our corpus. We concluded that this choice is not so much governed by proximity factors (use this for close objects, that for far objects), but that there are a lot of other influencing factors such as politeness, temporal aspects, and whose obligation is to complete the task. Learning what are the acceptable options vs. the bad choices in particular contexts would be helpful for GRE. Our current online system is appropriate for validating this observation and getting new insights on the use of proximity markings, as the obligation to complete the task is always on the user in our setting. For our future work we want to implement a system so that this determiners are used all the time when the target is located closer than a particular threshold, and observe the human ratings for this system compared to the preferred system from our first study. We predict that it will be ranked lower on fluency, and it will also feel less polite. The human survey can validate this proposal, and would be an important contribution for the field.

Other aspects of GRE that would be interesting to study in more detail are the configurations that are perceived as a group by humans and their translation in the
generation module for a dialog system, and if a special strategy is needed for bringing to attention the group and describing the targets that are part of a group.

Also, incorporating events (e.g. the opening of a door, or the door that the user just came through) in the generation process is one of the aspects we are interested for our future work. Events can appear as modifiers in the referring expressions, but they also are important for calculating the saliency of objects. We observed a small number of cases where the direction followers interpreted the expression *the button closer to the door* not as the button closer to the only visible door located in the front of the user, but as the button closer to the door that the person just came through. This is an example of how actions in the world influence the salience of objects.
CHAPTER 9

CONCLUSIONS

We feel that the problem of generating RE for situated environments is very important because it investigates the connection between linguistic behavior and the surrounding world. While there have been many studies that examine the appropriate way to refer, they don’t take into account the location of the speaker in an incremental way, or as a factor that can change the generation strategy. This study advances the state of the art by examining methods to incorporate real-time positional information of the user in the referring process, which will enable us to construct dialog systems that can react to the environment variables.

Most of the studies in linguistics are based on constructed examples and do not take into account the state of the world in which the speech was produced. But people’s communication very often takes place in a particular space and actions are performed during dialog. The development of a corpus of spontaneous human speech data in a situated world is a valuable resource for the NLP community. In NLG, a free corpus annotated with reference and having spatial and orientation information synchronized at word level will help develop algorithms and direct comparison of results over a common set of naturally occurring data. The evaluation of GRE algorithms has been a problematic issue in the field of generation for many years.
To overcome the problem of evaluation in this thesis, we completed two human studies. The first one takes the results of an automatic algorithm and substitutes it to the human produced language and asks for human ratings, and the second is a full evaluation on different qualitative dimensions (naturalness, friendliness, informativeness, etc) and quantitative metrics (time to completion, success rate, time to ground a description, etc.) in a human-computer interaction.

Our research in the way people use REs in situated worlds proposes a dynamic approach on GRE and informs future models for integrating various stimuli in the production of language. A direction giving task where the partner has mobility opens the door to new research in coordination between GRE and the manipulation of current positions of the participants.

Even though we separated the problem of generating referring expressions from the larger problem of generation to be able to study the effects of situated factors, our research shows it is hard to make such a separation. Different stages of generation interact among themselves: for example, the goal of identifying a referent can take several turns and be composed of utterances that bring into focus a grouping, directions to turn, or descriptions of objects, which affects the aggregation of information. Using a theory based on goals, actions and communicative intentions to plan entire sentences, such as [Stone et al., 2003], has been shown to be beneficial for GRE. The speed concerns and reliance on a rich modeling of the domain made it not practical for our research, but it is a possibility for the future, as faster solutions are proposed (see [Koller and Stone, 2007]). Our research shows that the level of sentence might not be enough, and that in some cases, planning an entire dialog system contribution, that may take several turns, might be necessary (for example, in referring to an object
in a situated environment). This will be able to account for the behavior we observed in our corpus, and is in line with the observations about reference as a collaborative, interactive process presented in [Clark and Krych, 2004, Clark and Bangerter, 2004].

The natural language generation community has started to give more attention to situated language, as mobile devices are becoming ubiquitous in our everyday life, and language interfaces that are aware of the user’s spatial information and visual field can be a large market in the future. Studying language in virtual environments has been received well by the community, and the research presented in this thesis has inspired a proposal for organizing a challenge in instruction-giving in virtual environments as an evaluation testbed for natural language generation [Byron et al., 2007].
APPENDIX A

DATA COLLECTION MATERIALS

Instructions for the Follower

**Phase 1.** Learning the controls

First you will be put into a small map with no partner, to get accustomed to the quake controls. The keyboard controls are posted on the wall in front of you.

You can ask to stop the experiment at any time if you feel queasy or dizzy. Please let the experimenter know as soon as you experience any discomfort.

Practice moving around using the arrow keys.

Practice these actions:
1. Pick up the Rebreather or the Quad Damage
2. Push the blue button to open the cabinet
3. Drop the Quad Damage or the Rebreather inside the cabinet and close the door by pushing the button again.

Adjust the audio level to a comfortable level.

**Phase 2.** Completing the tasks

In this phase you will be put in a new location. Your partner will direct you in completing 5 tasks. He will see the same view that you are seeing, but you are the only one that can move and act in the world. You will not be timed, but penalty points will be taken for pushing the wrong buttons or placing things in the wrong cabinets. At the end, if your team did not make more than five mistakes, you get an extra small prize. When you are done, let the experimenter know.

**Figure A.1:** The instructions received by the Direction Follower
Instructions for the Direction Giver

Phase 1. - Planning the task.

Your packet contains a map of the quake world with 5 objectives that you have to direct your partner to perform. Read the instructions and take your time to plan the directions you want to give to your partner. Walk around the maze to get familiarized with the environment if you wish. You do not have to complete the tasks. You can make notes on your map and write down anything you want. Take your time to get accustomed to the interface and the tasks and let the experimenter know when you are ready to proceed to phase 2.

You can ask to stop the experiment at any time if you feel queasy or dizzy. Please let the experimenter know as soon as you experience any discomfort.

Phase 2. - Directing the follower.

In this phase your partner will be placed into the world in the start position. Your monitor will show his/her view of the world as he/she moves around. He/she has no knowledge of the tasks, and has not received a map. You have to direct him/her through speech in order to complete the tasks. You can keep the map to help you. The objective is to complete all the task, but the order does not matter. You will not be timed, but penalty points will be taken for pushing the wrong buttons or placing things in the wrong cabinets. At the end, if your team did not make more than five mistakes, you get an extra small prize. When you are done, let the experimenter know.

Figure A.2: The instructions received by the Direction Giver
Quake Controls

Use the arrow keys for Movement:

Walk forward: ↑
Walk Backward: ↓
Turn Right: →
Turn Left: ←

To jump: use Spacebar.

To press a button: Walk over to the button. You will see it depress.

To pick up an object: Step onto the item then press Ctrl. (Control key)

To drop an object: Hit TAB to see the list of items that you are currently carrying. Press the letter beside the item you wish to drop. Press TAB again to make the menu go away.

Objects in the world:

<table>
<thead>
<tr>
<th>Cabinet</th>
<th>Button</th>
</tr>
</thead>
</table>

There are three objects that you can pick up and move:

| Quad damage | Rebreather | Silencer |

Figure A.3: The Quake controls received by all participants (in data collection and in the human-computer interaction study)
Your task:

You should direct your partner in completing the following tasks in whatever order you choose:

A) move the picture to the other wall

B) move the boxes on the long table so that the final configuration matches this picture:

C) hide the Rebreather in Cabinet9
   (To "hide" an item, you have to find it, pick it up, drop it in the cabinet and close the door. Make sure that your partner steps close enough to the cabinet when he is dropping the item, otherwise it will not be dropped inside, but outside of it.)

D) hide the Silencer in Cabinet4

E) hide the Quad Damage in Cabinet14

At the end, return to the starting point ★ and tell the experimenter that you are done.

Figure A.4: The task descriptions received by the Direction Giver
Buttons functions:

B1- opens/closes C11 and C14
B2- opens/closes C10
B3- opens/closes C9
B4- opens/closes C8
B5- nothing
B6- opens/closes C7
B7- nothing
B8- nothing
B9- moves boxes on long table
B10- slides pipe
B11- opens/closes C12
B12- moves picture
B13- opens/closes C1
B14- opens/closes C2
B15- opens/closes C4
B16- opens/closes C3
B17- opens/closes C5
B18- opens/closes C6
B19- moves box on table
B20- nothing
B21- opens/closes C15
B22- nothing
B23- nothing
B24- nothing
B25- nothing
B26- nothing
B27- nothing

Cabinet contents:

C6- contains Rebreather
C7- contains Quad Damage
C12- contains Silencer
The rest of the cabinets are empty.

Note: Buttons and cabinets ids are for the clarification of the map and should not be used in the instructions.
Figure A.6: The map of the world received by the Direction Giver
Survey

Background data

Sex: [ ] female  [ ] male

Year of birth: 19 ....

Native speaker of American English [ ] Yes [ ] No

What is the State where you grew up in and go to school

Answer the following questions, using a 1(low)-5 (high) scale.

- How would you rate your expertise as a computer user?
  
  1 (very poor)  2 (poor)  3 (average)  4 (good)  5 (very good)

- What was the difficulty of the task you had to complete?

  1 (not difficult at all)  2 (somewhat difficult)  3 (average)  4 (difficult)  5 (very difficult)

- How did you feel about the difficulty of the navigation in the Virtual World? (only for the follower)

  1 (not difficult at all)  2 (somewhat difficult)  3 (average)  4 (difficult)  5 (very difficult)

- Were the descriptions of the objects in the world difficult to follow? (were the objects difficult to identify by their provided descriptions?)

  1 (not difficult at all)  2 (somewhat difficult)  3 (average)  4 (difficult)  5 (very difficult)

Figure A.7: The survey at the end of the data collection
APPENDIX B

MATERIALS USED IN THE HUMAN EVALUATION

Survey

Background data:

Sex: [ ] female [ ] male

Year of birth: 19....

Native speaker of American English [ ] Yes [ ] No

Task:
You will need evaluate the items marked in red in the transcript as expressions to refer to the object id indicated. Watch the movie and listen to the sound carefully. You can stop and replay any section you want. It’s permitted to go back in the file and change your decisions. Your decision should be between the choices presented to you: option1, option 2 or equal. Please read the options carefully and take your time to evaluate them. The decision is not about the quality of the strings, but which one is better, so if you can tell a difference, please pick one of the choices. If the options seem equally bad/ or equally good, just pick “equal”.

You can stop any time you feel the need for a break and go back the the study after a small pause.

Thank you very much :)

Figure B.1: The instructions received by the human evaluators
APPENDIX C

HUMAN-COMPUTER INTERACTION STUDY
MATERIALS

This appendix contains a more detailed description of the materials used in the human evaluation of the implemented systems from Chapter 7. The Quake Controls, in Figure A.3 were also handed out to the participants.
Instructions

**Phase 1.** - Learning the virtual world controls.

First you will be put into a small practice map, to get accustomed to the controls. The keyboard controls are posted on the wall in front of you. In this toy world, there is one door, one button, and one cabinet.

You can ask to stop the experiment at any time if you feel queasy or dizzy. Please let the experimenter know as soon as you experience any discomfort.

Practice moving around using the arrow keys. Practice going in and out through the door.

Practice these specific actions:
1. Pick up an object (either the Rebreather or the Quad Damage)
2. Push the button to open the cabinet
3. Drop the Quad Damage or the Rebreather inside of the cabinet and close it by pushing the button again.

Adjust the audio volume to a comfortable level.

**Phase 2.** – Getting acquainted with the system.

In this phase you will be put in a new environment, with multiple rooms. Your partner is a computer program that will direct you in completing a task. The system will tell you step by step instructions and can monitor some of your actions.

Please follow the instructions as well as you can.

Each session begins with the system describing the particular task and starting its instructions. When the task is completed, the system will inform you by saying: “We successfully completed our task! Good bye!”, or, in cases when the system has problems in giving the instructions, it will produce a message to inform you of this: “I am sorry. We are unable to finish this task. Good bye!”. At this time, quit the Quake environment by pressing Esc, then choosing Quit, and answering Yes.

In this phase you will complete a task for training purposes.

**Phase 3.** - Completing the tasks.

In this phase you will complete a series of tasks while directed by two different systems. Please follow the instructions as well as you can. After interacting with each system, you will complete a survey to evaluate its behavior.

*Figure C.1:* The instructions received by the human participants
During your session, the software will instruct you to complete one or more of the following tasks:

- **Finding the cabinet that contains a specific object**, picking it up and hiding it in a different cabinet.
  - The objects that you can pick up are:
    - Quake logo
    - Rebreather
    - Silencer

- **Moving a picture from one wall to another.**
  - The picture in the initial configuration
  - The picture after

- **Moving some boxes around.**
  - The boxes in the initial configuration
  - The boxes after

- **Repairing a pipe.**
  - The pipe before repair
  - The pipe after

---

**Figure C.2:** Information received by the human participants: possible tasks
Evaluation of System ____

1) How would you rate the average ease or difficulty of the tasks you completed?

   □  □  □  □  □
   (very difficult) (difficult) (average) (easy) (very easy)

2) How would you rate the intelligibility of the system’s voice?

   □  □  □  □  □
   (very poor) (poor) (average) (good) (very good)

3) How would you rate the naturalness of the language used by the system?

   □  □  □  □  □
   (very poor) (poor) (average) (good) (very good)

4) How easy was it to follow the directions given by the system?

   □  □  □  □  □
   (very difficult) (difficult) (average) (easy) (very easy)

5) Did you feel the directions were:
   a. too informative
   Yes: ___  No: ___
   b. not informative enough
   Yes: ___  No: ___

6) How easy was it to understand the system’s descriptions of different objects in the world?

   □  □  □  □  □
   (very difficult) (difficult) (average) (easy) (very easy)

7) How would you rate the friendliness of this system?

   □  □  □  □  □
   (very low) (low) (average) (high) (very high)

**Figure C.3:** System related survey questions received by the human participants (page 1)
Rate the following statements on a 1 to 5 scale (1=not agree, 5=agree).

8) I found the system to be cooperative during the conversation.

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>(not agree)</td>
<td></td>
<td></td>
<td></td>
<td>(agree)</td>
</tr>
</tbody>
</table>

9) I found the conversation engaging.

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>(not agree)</td>
<td></td>
<td></td>
<td></td>
<td>(agree)</td>
</tr>
</tbody>
</table>

10) I found it exciting to interact with the system

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>(not agree)</td>
<td></td>
<td></td>
<td></td>
<td>(agree)</td>
</tr>
</tbody>
</table>

11) I was satisfied with the system’s ability to help me complete the task.

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>(not agree)</td>
<td></td>
<td></td>
<td></td>
<td>(agree)</td>
</tr>
</tbody>
</table>

12) Do you have any overall comments on the system's performance?

<p>| |</p>
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

**Figure C.4:** System related survey questions received by the human participants (page 2)
Survey

Background data

Sex: [ ] female [ ] male
Age: ...
Native speaker of American English [ ] Yes [ ] No

Answer the following questions, marking the statement you agree with the most.

1) How would you rate your expertise as a computer user?

   □ □ □ □ □
   (very poor) (poor) (average) (good) (very good)

2) How many hours/week do you play computer games?

3) How did you feel about navigating in the Virtual World?

   □ □ □ □ □
   (very difficult) (difficult) (average) (easy) (very easy)

4) Did you perceive any difference between the two systems, and if yes, which one did you prefer?

   Any Difference? Yes: ... No: ...
   Prefer: System 1: ... System 2: ...

• Please write down any comments or suggestions:

Figure C.5: General survey questions received by the human participants (received at the end)
APPENDIX D

EXAMPLE TRANSCRIPT OF A HUMAN-COMPUTER INTERACTION

Session 22, with the system in the Adjusted Context condition

1:09:42 PM EDT 69 millisec; hello
1:09:43 PM EDT 296 millisec; our task is to find the cabinet that has the quake logo
1:09:47 PM EDT 384 millisec; pick it up and hide it in a different cabinet
1:09:50 PM EDT 199 millisec; ok so
1:09:52 PM EDT 778 millisec; turn to your right
1:09:54 PM EDT 554 millisec; ok
1:09:55 PM EDT 691 millisec; go through the door in front of you
1:10:00 PM EDT 668 millisec; ok
1:10:01 PM EDT 867 millisec; go through the door in front of you
1:10:05 PM EDT 799 millisec; ok
1:10:06 PM EDT 850 millisec; go forward
1:10:08 PM EDT 716 millisec; ok
1:10:09 PM EDT 798 millisec; push the button on your right
1:10:19 PM EDT 285 millisec; push the button in front of you
1:10:23 PM EDT 95 millisec; ok
1:10:24 PM EDT 138 millisec; turn to your left
1:10:25 PM EDT 879 millisec; ok
1:10:26 PM EDT 942 millisec; go to the cabinet in front of you that is now open
1:10:30 PM EDT 449 millisec; ok
1:10:31 PM EDT 447 millisec; pick up the quakelogo
1:10:35 PM EDT 129 millisec; ok
1:10:36 PM EDT 171 millisec; turn to your left
1:10:44 PM EDT 605 millisec; turn to your left
1:10:49 PM EDT 480 millisec; step out of the cabinet
1:10:51 PM EDT 855 millisec; ok
1:10:53 PM EDT 355 millisec; turn to your right

141
1:10:55 PM EDT 979 millisec; ok
1:10:57 PM EDT 107 millisec; go through the door
1:11:00 PM EDT 744 millisec; ok
1:11:02 PM EDT 663 millisec; go through the door in front of you
1:11:04 PM EDT 763 millisec; ok
1:11:05 PM EDT 833 millisec; go through the door in front of you
1:11:08 PM EDT 587 millisec; ok
1:11:09 PM EDT 712 millisec; turn to your left
1:11:11 PM EDT 860 millisec; ok
1:11:12 PM EDT 872 millisec; push the second button from the left
1:11:17 PM EDT 259 millisec; ok
1:11:18 PM EDT 230 millisec; turn to your right
1:11:21 PM EDT 880 millisec; ok
1:11:22 PM EDT 864 millisec; go to the cabinet in front of you that is now open
1:11:32 PM EDT 373 millisec; turn to your left
1:11:35 PM EDT 129 millisec; ok
1:11:36 PM EDT 116 millisec; go to the cabinet closer to the door that is now open
1:11:45 PM EDT 744 millisec; ok
1:11:46 PM EDT 739 millisec; drop the quakelogo
1:11:51 PM EDT 732 millisec; ok
1:11:52 PM EDT 753 millisec; turn to your right
1:11:54 PM EDT 730 millisec; ok
1:11:55 PM EDT 681 millisec; step out of the cabinet
1:11:57 PM EDT 595 millisec; ok
1:11:58 PM EDT 570 millisec; go forward
1:12:02 PM EDT 745 millisec; ok
1:12:03 PM EDT 699 millisec; push the second button from the left
1:12:11 PM EDT 485 millisec; ok
1:12:12 PM EDT 431 millisec; the quake logo is now hidden
1:12:14 PM EDT 525 millisec; we successfully completed our task
1:12:17 PM EDT 639 millisec; good bye
APPENDIX E

ALL POSSIBLE TURNS OF THE INTERACTION SYSTEM

Drop the Quake Logo.
Drop the Rebreather.
Drop the Silencer.
Go all the way through the tunnel.
Go around the bend.
Go down the stairs.
Go forward.\textsuperscript{27}
Go through the door.
Go through the door behind you.
Go through the door closer to the button.
Go through the door closer to the cabinet.
Go through the door farther from the button.
Go through the door farther from the cabinet.
Go through the door in front of you.
Go through the door on your left.
Go through the door on your right.
Go through the left door.
Go through the right door.
Go to the cabinet behind you that is now open.
Go to the cabinet closer to the button that is now open.
Go to the cabinet closer to the door that is now open.
Go to the cabinet farther from the button that is now open.
Go to the cabinet farther from the door that is now open.
Go to the cabinet in front of you that is now open.
Go to the cabinet on your left that is now open.
Go to the cabinet on your right that is now open.
Go to the cabinet that is now open.

\textsuperscript{27}only the system in the Adjusted Context condition

143
Go to the front cabinet that is now open.
Go to the left cabinet that is now open.
Go to the right cabinet that is now open.
Go up the stairs.
Good bye!
Hello!
I am going to give you step by step instructions to complete a task.
I am sorry we are unable to finish this task.
Ok!
Ok, so
Our task is to find the cabinet that has the Quake Logo.
Our task is to find the cabinet that has the Rebreather.
Our task is to find the cabinet that has the Silencer.
Our task is to make a box disappear.
Our task is to move a picture from one wall to another.
Our task is to move some boxes around.
Pick it up and hide it in a different cabinet.
Pick up the Quake Logo.
Pick up the Rebreather.
Pick up the Silencer.
Push the button.
Push the button behind you.
Push the button closer to the cabinet.
Push the button closer to the door.
Push the button farther from the cabinet.
Push the button farther from the door.
Push the button in front of you.
Push the button on your left.
Push the button on your right.
Push the left button.
Push the middle button.
Push the right button.
Push the second button from the left.
Step out of the cabinet.
The box disappeared.
The boxes have been moved.
The picture has been moved.
The Quake Logo is now hidden.
The Rebreather is now hidden.
The Silencer is now hidden.
Turn to your left. ²⁸
Turn to your right. ²⁹
We successfully completed our task.

²⁸ only the system in the Adjusted Context condition
²⁹ only the system in the Adjusted Context condition
APPENDIX F

EXAMPLE TRANSCRIPT FROM THE HUMAN CORPUS

Session 15 - two male speakers, total time: 8 min 54 secs

DF: WHAT ARE YOU ON A PC BACK THERE
DG: YEAH YEAH IT'S JUST THE JUST THE SCREEN IS CONNECTED I
GOT NOTHING ELSE
DF: AH
DG: HERE
DF: OH I GOT YOU OK
DG: UM WOW OK YEAH LET'S DO A SPIN AROUND OK NOW TURN
AROUND COM-PLETELY TURN BACKWARDS THERE YOU GO THAT
DOOR RIGHT AHEAD OF YOU GO THROUGH THAT ONE OK
DF: YOU'RE DIZZY
DG: AND
DF: YET [laugh]
DG: YEAH [laugh] WE MIGHT NEED TO STOP
DF: [laugh]
DG: UH TURN TO TH- TO THE LEFT THERE GO THROUGH THE DOOR
STRAIGHT AHEAD OF YOU
DF: MHM
DG: ALRIGHT NOW WE'RE GONNA DO WE'RE GONNA MOVE THE
PICTURE
DF: MHM
DG: SO THERE'S A PICTURE ALL THE WAY TO THE LEFT AND THEN
THERE ARE THREE BUTTONS
DF: YES
DG: YOU NEED TO PRESS THE CENTER BUTTON
DF: OK
DG: THERE YOU GO YOU YOU MOVED THE PICTURE THAT'S IT
DF: OH THAT'S IT OK

146
DG: YEAH
DG: IT'S LIKE A PICTURE IT'S LIKE A RIVER
DF: RIGHT
DG: SWEET UM OK NOW WE NEED TO UH DO THE NEXT TASK WHICH IS MOVE THE BOXES ON THE LONG TABLE SO
DF: MHM
DG: ALRIGHT HANG ON SO I DON'T LOSE YOU HERE UM OK YOU'RE GONNA GO THROUGH
DF: MHM
DG: THE DOOR NEXT TO THE CHAIR SO THAT DOOR RIGHT THERE
DF: RIGHT
DG: OK AND THERE'RE THE BOXES ON THE TABLE
DF: GOT YOU
DG: THERE ARE TWO BUTTONS ON THE WALL
DF: MHM
DG: AND YOU NEED TO PRESS THE LEFT ONE OK THAT SHOULD DO IT YEAH THERE YOU GO
DF: THAT'S IT
DG: SWEET MAN
DF: NICE IT JUST
DG: UM
DF: MOVED BACK
DG: YEAH THEY MOVE WELL I DON'T KNOW MAYBE THEY ARE SUPPOSED TO MOVE BACK HAVE NO IDEA
DF: OK
DG: DID YOU MOVE AGAIN OK THAT WORKS
DF: OK
DG: UM OK SO NOW WE'RE GONNA DO LIKE THE NEXT THREE TASKS ALL AT ONCE
DF: OK
DG: AND UM SO THE FIRST THING WE NEED TO DO IT'S GET THE SILENCER
DF: MHM AND THAT WOULD BE WHERE
DG: AND I NEED TO GUIDE YOU THERE
DF: OK
DG: SO GO BACK OUT THE DOOR THAT YOU CAME IN
DF: MHM
DG: AND NOW HM THAT'S INTERESTING OK NOW GO UM STRAIGHT YEP
DF: MHM
DG: AND GO STRAIGHT AGAIN OK AND NOW GO DOWN THE HALLWAY
HERE SWEET
DF: [laugh]
DG: UH WHERE IS THE SILENCER OH OK GO THROUGH THE DOOR THERE
DF: OK
DG: AND THERE IS A BUTTON THAT OPENS THE CABINET
DF: YEP NOTHING IN IT
DG: UH THERE'S ANOTHER CABINET
DF: OH RIGHT
DG: YEP
DF: THAT JUST
DG: THAT'S
DF: CLOSED
DG: THE SILENCER NO
DF: OH
DG: THERE
DF: GOT
DG: YOU
DF: IT
DG: GO IT'S
DF: OK
DG: OPEN
DF: GOT IT
DG: NICE OK WE GOT THE SILENCER
DF: MHM
DG: UM NOW WE NEED TO GO DOWNSTAIRS THERE'RE STAIRS IN THIS ROOM
DF: YEP GOT IT THROUGH THIS DOOR
DG: UH OH I NEED TO SWITCH TO MY OTHER MAP [laugh]
DF: I THINK THAT'S WHERE I WAS BEFORE THAT TOP ROOM THERE
DG: YEAH AND THE BLOCKS MOVED BACK THAT'S
DF: YEAH
DG: A BUG WE NEED TO SEND
DF: [laugh]
DG: A BUG REPORT
DF: HERE WE GO BUGZILA
DG: UM OK LET ME SEE HOW TO GET TO THE REBREATHER CABINET
DF: OK GO STRAIGHT THROUGH THIS DOOR
DG: AND TAKE A LEFT
DF: MHM
DG: UH WAIT A MINUTE YEAH OK
DF: OK
DG: YEAH YOU’RE IN THE RIGHT PLACE
DF: MHM
DG: HOW DO YOU OPEN THE CABINET
DF: THERE IS A BUTTON THERE
DG: OH THE BUTTON WAS IN THE OTHER ROOM
DF: OK THAT ONE’S
DG: YEAH IT’S NOT THAT YOU UH NEED
DF: OH
DG: TO PUT IT IN THE IN THE RIGHT CABINET
DF: OK SO MAYBE IT’S ONE OF THESE
DG: YEAH IT’S ONE OF THOSE HANG ON JUST SO I CAN TELL YOU WHICH ONE
DF: ALRIGHT
DG: UM OK IT’S THE ONE THAT’S NEXT TO THE DOOR RIGHT
DIRECTLY IN FRONT OF YOU THERE YOU GO NOW GET THE
REBREATHER
DF: GOT IT
DG: OK NOW WE NEED TO HIDE THE SILENCER
DF: IN THERE WHICH WERE WO-
DG: IT’S GONNA BE IN CABINET UH OH or YOU DON’T HAVE THE
NUMBERS ON THE CABINETS
DF: NO
DG: UM OK SO WE’RE GONNA NEED TO LEAVE THIS ROOM
DF: MHM
DG: AND GO BACK INTO THE ROOM YOU WERE IN
DF: OK
DG: AND GO TO THE DOOR THAT’S TO THE LEFT THERE OK
DF: MHM
DG: NOW GO AROUND THE BEND HERE AND THROUGH THE DOOR
DF: OK
DG: THE CABINET OH OK
DF: THERE’S TWO THERE OTHER RIGHT
DG: [noise] ALRIGHT YEP PRESS THE BUTTON THE FARTHEST BUTTON [noise]
DF: ALRIGHT THE OPPOSITE
DG: YEP
DF: ONE THAT OPENED RIGHT
DG: AND PUT THE SILENCER IN THERE
DF: [noise]
DG: AWESOME I ASSUME YOU PROBABLY HAVE TO WE HAVE TO HIDE IT
SO CLOSE THE CABINET

DF: OK GOT
DG: HIDDEN
DF: IT
DG: NO ONE WILL FIND THE SILENCER
DF: [laugh] YEAH RIGHT
DG: IT IS GONE
DF: [laugh]
DG: UM OK NOW WE NEED TO UM MMM NOW WE NEED TO HIDE THE REBREATH IN CABINET NINE WHICH IS UPSTAIRS
DF: OK
DG: SO
DF: THE STAIRS RIGHT HERE
DG: YEAH GET US UPSTAIRS
DF: UPSTAIRS [laugh] GO THROUGH THAT DOOR
DG: UH YEAH
DF: DO YOU NEED TO STOP ARE YOU FEELING QUEASY OR DIZZY OR
DG: WE MIGHT NEED A SOFTER
DF: [laugh]
DG: RIDE
Both: [ SO ]
DF: OK
DG: CAN CAN YOU GET US BACK TO THE ROOM WHERE WE STARTED OR DO YOU WANT ME TO GUIDE YOU TH- OH YOU’RE ALMOST THERE SWEET
DF: IT LOOKS FAMILIAR YEAH I
DG: YEAH
DF: THINK I
DG: NOW
DF: WENT STRAIGHT
DG: I THINK YOU TAKE A UM
DF: NO
DG: OH
DF: OH THAT
DG: NO
DF: WASN’T
DF: THERE NO IT WAS ONE OF THESE
DG: NOT THAT ONE THE
DF: OK
DG: OTHER
DG: ONE
DF: THE OTHER ONE
DG: TO YOUR TO YOUR RIGHT
DF: ALRI- OH I I I DIDN'T START HERE DID I
DG: YEAH YOU DID
DF: OK
DG: YEAH UM GO GO BACK OUT OF THIS ROOM AND IT'S THE DOOR ON
YOUR LEFT THERE OK
DG: AND NOW WE'VE GOTTA HIDE THE UH REBREATHER
DF: MHM
DG: IN ONE OF THE CABINETS IN THIS ROOM
DF: OK
DG: AND THE WAY YOU'RE FACING RIGHT NOW IT'S GOING TO BE THE
CABINET ON YOUR LEFT SO PUSH THE LEFT BUTTON AND HIDE
THE REBREATHER IN THERE NO TEA FOR YOU
DF: [laugh]
DG: AND CLOSE THE CABINET DOORS WHAT IS THAT OUTSIDE THERE
LOOKS LIKE IT'S THE STADIUM ON FIRE
DF: LOOKS LIKE MARS UNTIL
DG: [laugh]
DF: A WEEK All [laugh] RIGHT
DG: YEAH I DON'T KNOW IT'S LIKE IT'S LIKE OHIO STATE IN MID
APOCALYPSE
DF: ALRIGHT [laugh]
DG: [laugh]
DF: [breath]
DG: UM OK SO UM LAST TASK IS TO FIND THE QUAD DAMAGE AND
HIDE IT
DF: MHM
DG: AND THAT'S ALL GONNA BE ON THE FLOOR THAT WE'RE ON
RIGHT NOW
DF: OK
DG: SO WITHOUT LEAVING THIS ROOM YEP TURN TO YOUR RIGHT
DF: MHM
DG: OP BACK OUT TURN TO YOUR RIGHT MORE THERE IS A DOOR
RIGHT THERE
DF: GOT IT [breath]
DG: GO THROUGH THERE
DF: MHM
DG: GO STRAIGHT INTO THE NEXT ROOM
DF: MHM
DG: AND UM THE QUAD DAMAGE IS IN THIS ROOM
DF: OK
DG: AND WHICH NEED TO FIGURE OUT WHICH BUTTON TO PRESS OH
   IT'S GOING TO BE THE BUTTON TURN TO YOUR RIGHT IT'S GONNA
   BE THE BUTTON D- ON THE LEFT ON THAT WALL
DF: THIS ONE
DG: YEP THAT ONE AND
DF: [mumble]
DG: YEP THERE
DF: OK
DG: IT IS
DF: GOT IT
DG: QUAD DAMAGE
DF: YEP OK
DG: UM NOW GO BACK THROUGH THAT DOOR YOU CAME IN AND
   STRAIGHT AGAIN AND GO STRAIGHT AND
DG: OOPS
DF: OH JUST
DG: PRESS
DF: PRESS A BUTTON
DF: HUH
DG: YEAH THAT'S OK PRESS THE THE UH PRESS THE BUTTON NEXT TO
   THAT ONE
DF: OK
DG: OH THEY BOTH OPEN
DF: WOW
DG: OH WHICH ONE IS IT IT'S THE CABINET ON THE LEFT HIDE IT IN
   THAT ONE
DF: OK AND CLOSED
DG: YEAH AND
DF: TH-
DG: CLOSE THEM UP AND NOW YOU GOTTA GO BACK TO WHERE
   WE STARTED WHICH
DF: UH
DG: SHOULD BE TAKE A RIGHT
DF: [noise]
DG: UM AND IT'S THE DOOR ON THE RIGHT
DF: OK
DG: AND THAT'S IT
DG: THIS
DF: OH
DG: IS THIS IS WHERE WE STOP
DF: RIGHT
DG: WE’RE DONE
DF: OK GOOD
DG: AWESOME
DF: NICE
DG: [noise]
DF: OVER AND OUT
DG: YEAH OVER AND OUT
DF: [laugh]
DG: WELL DONE
DF: THANKS
DG: [noise]
DF: YOU TOO [noise] [laugh] MISSION ACCOMPLISHED [noise]


