THE GENERATION
OF NATURAL DESCRIPTIONS
CORPUS-BASED INVESTIGATIONS
OF REFERRING EXPRESSIONS
IN VISUAL DOMAINS

Henriette Anna Elisabeth Viethen

This dissertation is presented for the degree of

Doctor of Philosophy

at

MACQUARIE UNIVERSITY

March 2011
Declaration

The research presented in this thesis is the original work of the author except where otherwise indicated. Some parts of the thesis include revised versions of published papers. This work has not been submitted for a degree or any other qualification to any other university or institution. All verbatim extracts have been distinguished by quotations, and all sources of information have been specifically acknowledged.

The research presented in this thesis was approved by the Macquarie University Ethics Review Committee, reference number: HE27FEB2009-D06283 on 16 January 2009.

Signed: Henriette Anna Elisabeth Viethen

Date:
Referring expression generation (REG) has been studied by computational linguists for nearly three decades. Although other aspects of the task have been examined, most investigations into REG are focussed on the selection of those attributes of an object that best distinguish it from all others in its environment. Historically, much of this work has suffered from two problems: firstly, it does not take account of empirical evidence for how people refer; and secondly, it has not been evaluated against human-produced corpora.

This thesis is based on two related premises which I take to be self-evident if our ultimate goal is to explain how humans refer: first, that naturalness should be the primary goal of computational models of referring expression generation, and second, that the task therefore needs to be approached by using human-produced corpora for the development and testing of algorithms.

Based on these premises, this thesis presents an extensive exploration into how corpora can be used in REG. It makes three main contributions in this area: (1) it presents a study that explores how corpora can be used to evaluate algorithms for the generation of referring expressions, and shows that existing algorithms cannot fully account for the way humans generate referring expressions; (2) it provides a detailed analysis of the different aspects of the human use of referring expressions in two large corpora in order to inform the development of REG algorithms; and (3) it presents experiments in using these corpora to train decision trees for attribute selection for referring expressions. The main conclusion of the analyses and experiments in this thesis is that speaker-specific variation plays a much larger role in the generation of referring expressions than existing algorithms acknowledge.

Chapter 2 begins by surveying existing research in the field of REG. Chapter 3 then provides an in-depth discussion of the methodological choices that have to be made when employing corpora to inform and evaluate REG algorithms. Chapter 4 presents an evaluation of three popular existing REG algorithms using a small corpus of human-produced data. It shows that, while one of the algorithms is capable of generating a large proportion of the referring expressions in the corpus, none of them are even in principle able to generate all of them. The experiment gives rise to a dissection of the issues involved in the evaluation of REG algorithms. Based on the analyses of the previous three chapters, Chapter 5 describes the design, collection and annotation of two large corpora of referring expressions, and analyses how speakers make use of different object properties. These corpora are novel in that they contain spatial relations between objects, allowing a systematic analysis of the circumstances under which people use relations as well as other properties. The second corpus constitutes the largest systematically-designed single-domain collection of referring expressions to date. Finally, Chapter 6 explores the use of the corpora described in Chapter 5 to train algorithms which model the content selection behaviour of the human participants who contributed the data. Modelling this data using decision trees is a natural way to gain insights into the factors that influence a person’s decision to include a particular property in a referring expression and how these factors interact.
Acknowledgements

First of all, I thank my supervisor Robert Dale. I can truly say that without him this thesis would never have happened. I thank him for believing in me and in my project, even when I didn’t, for encouraging me to try things out and for forcing me to finish things off, for sending me to every conference and every summer school, for teaching me how to write research papers, for showing me what it means to treat students with respect, for never putting his own interests before mine, and for forgiving me my inability to let go of German punctuation rules.

I thank Albert Gatt, Alexander Koller, Anja Belz, Ehud Reiter, Emiel Krahmer, Ielka van der Sluis, Intiaz Khan, Kees van Deemter, Mariët Theune, and Ross Turner for interesting discussions and for reminding me that there are other people interested in the same things as me, albeit mostly on the other side of the world. I thank my three thesis reviewers, John Kelleher, Kathy McCoy, and Kees van Deemter for their constructive comments and suggestions. Special thanks go to Meg Mitchell who probably has no idea that in a short online chat she explained to me what my work was all about. Without her I might never have figured out how to pull it all together into one rounded piece of work.

I thank the Markists and the CLT lunch gang for their support and friendship and for making lunch breaks worth the daily trek to uni: Andrew Lampert, Ben Hachey, Ben Phelan, Diego Molla, Elena Akhmatova, François Lareau, Jean-Philippe Prost, Jojo Wong, Luiz Pizzato, Marc Tilbrook, Mark Dras, Mark Johnson, Mary Gardiner, Matt Honnibal, Menno van Zanen, Pawel Mazur, Rolf Schwitter, Stephen Wan, Suzy Howlett, Teresa Lynn, Yasaman Motazedi, various visitors over the years, and, in particular, Simon Zwarts, who patiently helped me with countless trivial technical problems and programming questions without ever being condescending and who just generally is a great friend.

I thank Ana Castro, Angélica Tomaz, Bonne Eggleston, Duncan Macinnis, Emily Mitchell, Geoff Thilo, Henry ‘Woody’ Woodruff, Jonathan Berant, Matt Roberts, Reut Tsarfaty, Rui Costa, Sascha Jockel, Shelley Bambrook, Susana Correia, Tanja Döring, Telia Curtis and Yael Augarten, as well as Jorge Cham and his phdcomics, for constantly reminding me that I was not the only one fighting a thesis monster and that it is possible to finally slay it. I am especially grateful to Yael Augarten, Telia Curtis and the rest of the crew from G12B at the School for Photovoltaics and Renewable Energy at UNSW, who provided me refuge in their office, where I hid from the world (and the internet) for four months to write up the first draft of this thesis.

I thank my parents, Maria Viethen, Richard Eßer and Volkhart Schönberg, my brother Lasse, and the rest of my family back in Germany and Austria, who despite having lost me to Australia always welcome me back with open arms and have been supporting my PhD adventure from afar.

Most importantly, I thank Smithy for keeping me alive with his love and his amazing cooking skills. Without him, it would all amount to nothing.
Contents

Abstract v

Acknowledgements vii

Notational Conventions xiii

1 Introduction 1
  1.1 Problem Statement .......................... 1
  1.2 Contributions .............................. 4
  1.3 Overview ................................. 5

2 Content Selection for Distinguishing Descriptions 7
  2.1 The Task of REG .............................. 8
  2.2 REG Frameworks .............................. 11
    2.2.1 Greedy Search .......................... 11
    2.2.2 Incremental Reference Generation ........... 13
    2.2.3 Graph-Based REG ........................ 18
    2.2.4 Other Frameworks ......................... 24
  2.3 Relations in REG ............................. 26
    2.3.1 Relational REG Using Constraints ............ 26
    2.3.2 Relational Extensions of the Incremental Algorithm 30
    2.3.3 Relations in Graphs ...................... 33
    2.3.4 Relations in Other Approaches ............... 34
  2.4 The Use of Corpora in REG ..................... 35
    2.4.1 Existing Corpora ........................ 36
    2.4.2 Empirical Approaches to REG ................ 39
    2.4.3 Evaluation against Human-produced Data ........ 42
  2.5 Optimality of Referring Expressions ............... 45
    2.5.1 Aiming for Brevity ....................... 46
    2.5.2 Taking into Account the Needs of the Listener 48
    2.5.3 Aiming for Naturalness .................... 51
  2.6 Discussion ................................ 53

3 Methodological Choices 57
  3.1 Issues in the Collection of Corpora ............... 57
    3.1.1 Collected vs. Found Data .................. 57
### Contents

3.1.2 Reference in Discourse vs. Isolated Reference .......................... 59
3.1.3 Characteristics of Domains ............................................. 60
3.1.4 Web-based vs. Off-line Data Collection ............................... 62

3.2 Issues in the Analysis of Corpora ........................................ 63
3.2.1 Types of Object Attributes .......................................... 64
3.2.2 Minimality and Over-specification .................................... 65

3.3 Comparing System Output to Corpus Data ................................ 73
3.3.1 Common Evaluation Metrics .......................................... 73
3.3.2 Attribute-Level vs. Property-Level Evaluation ....................... 75
3.3.3 Taking Length into Account .......................................... 76
3.3.4 Evaluation against Multiple Gold Standards ......................... 77
3.3.5 Surface-Level Evaluation ............................................. 77

3.4 Summary ........................................................................... 78

4 Corpus-Based Evaluation ................................................................
4.1 The Drawer Data ............................................................... 81

4.2 An Evaluation Experiment ...................................................... 85
4.2.1 Knowledge Representation ............................................. 86
4.2.2 The Algorithms ............................................................ 87
4.2.3 Results ......................................................................... 90
4.2.4 Other Approaches to Relations and Redundancy ................... 97
4.2.5 Discussion ..................................................................... 100

4.3 Issues in the Evaluation of REG Algorithms .............................. 101
4.3.1 Representational Choice .................................................. 101
4.3.2 Non-Determinism of Natural Language Choice .................... 104
4.3.3 Measuring Performance .................................................. 106
4.3.4 Domain Specificity ........................................................ 109
4.3.5 Interim Summary .......................................................... 111

4.4 The Referring Expression Generation Challenges ....................... 112
4.4.1 The Problem with Representational Choice ....................... 112
4.4.2 The Problem with Non-Determinism ................................. 113
4.4.3 The Problem with Measuring Performance ......................... 115
4.4.4 The Problem with Domain Specificity ............................... 116

4.5 Conclusions ....................................................................... 116

5 Collection and Analysis of Two REG Corpora ................................
5.1 Aim of the Corpus Collections .............................................. 120

5.2 Collecting GRE3D3 .............................................................. 122
5.2.1 Stimulus Design ............................................................. 122
5.2.2 Procedure and Participants ............................................. 127
5.2.3 Data Filtering and Annotation ......................................... 128

5.3 Analysis of GRE3D3 .............................................................. 130
5.3.1 General Overview .......................................................... 130
5.3.2 The Use of Spatial Relations in GRE3D3 ............................. 132
5.3.3 Interim Summary .......................................................... 138

5.4 Collecting GRE3D7 .............................................................. 140
5.4.1 Stimulus Design ............................................. 140
5.4.2 Procedure and Participants .............................. 143
5.4.3 Data Filtering and Annotation .......................... 145
5.5 Analysis of GRE3D7 ......................................... 146
  5.5.1 General Overview ...................................... 146
  5.5.2 The Use of Spatial Relations in GRE3D7 .............. 148
  5.5.3 Interim Summary ....................................... 154
5.6 Variation in the Two Corpora ............................... 155
5.7 Conclusions .................................................. 159

6 Corpus-Based Modelling of REG ............................. 163
  6.1 Setting Up the Experimental Framework ................. 164
    6.1.1 The Prediction Classes ................................ 164
    6.1.2 Features to Learn From ................................ 165
    6.1.3 Feature Values ........................................ 170
    6.1.4 Decision Tree Classifiers ............................ 171
  6.2 Modelling the Use of Complete Content Patterns ........ 172
  6.3 Modelling the Use of IndividualAttributes ............. 176
    6.3.1 The Target’s Attributes .............................. 177
    6.3.2 The Landmark’s Attributes ........................... 181
  6.4 Cross-Corpus Testing ...................................... 185
  6.5 Speaker-Dependent Variation .............................. 188
    6.5.1 Speaker as a Prediction Feature ...................... 188
    6.5.2 Training Speaker-Specific Trees ..................... 190
  6.6 Discussion ................................................ 195
    6.6.1 Conclusions .......................................... 195
    6.6.2 Implications for Algorithm Development .......... 197

7 Conclusions ..................................................... 201
  7.1 Summary and Discussion .................................. 201
    7.1.1 Corpus-Based Evaluation ............................. 202
    7.1.2 Corpus Collection and Analysis ..................... 204
    7.1.3 Corpus-Based Modelling ............................. 205
  7.2 Future Research Directions ............................... 206

A Materials for the GRE3D3 Collection Experiment ............ 209

B Materials for the GRE3D7 Collection Experiment ............. 213
  B.1 Screenshots of the Experiment .......................... 213
  B.2 Filler Scenes ............................................. 216

C Tables for Section 6.5.2 .................................... 221

D Publications Related to this Thesis .......................... 223

Bibliography ..................................................... 225
Notational Conventions

<table>
<thead>
<tr>
<th>Example</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>the blue ball</em></td>
<td>Linguistic examples in the body of the text are in italics.</td>
</tr>
<tr>
<td>$E = \sum_{i=0}^{n} x$</td>
<td>Mathematical symbols and variable names are in italics.</td>
</tr>
<tr>
<td>cluster</td>
<td>Boldface is used for technical terms when they are first introduced.</td>
</tr>
<tr>
<td><strong>Question 1</strong></td>
<td>Capitalised boldface is used to highlight hypotheses and research questions in the running text.</td>
</tr>
<tr>
<td>ACL</td>
<td>Small capitals are used for acronyms and names, such as abbreviations for systems and algorithms.</td>
</tr>
<tr>
<td>tg_size</td>
<td>Lowercase sans serif terms are used in knowledge representation contexts.</td>
</tr>
<tr>
<td>TG_Size</td>
<td>Capitalised sans serif terms are used for machine learning features.</td>
</tr>
</tbody>
</table>

**A note on the ‘academic plural’**

This thesis is written in the first person singular. Although many doctoral and masters theses are written in the ‘academic plural’, I feel at odds with this tradition of writing a monograph in the first person plural. This does not mean that I never use the first person plural. I do so when I refer to work that I have published in a co-authored paper, and in cases in which I invite the reader to join me in considering a particular idea or a certain section, table or figure of the thesis.