View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Publications at Bielefeld University

Incrementally Resolving References in Order to Identify Visually Present Objects in a Situated Dialogue Setting

Casey R. Kennington

Doctor of Philosophy
Department of Linguistics and Literature
Bielefeld University

2016

Copyright ©2016 Casey R. Kennington, Bielefeld, Germany.
Dissertation zur Erlangung des akademischen Grades *Doctor philosophiae* (Dr. phil.) vorgelegt an der Fakultät für Linguistik und Literaturwissenschaft der Universität Bielefeld am 30 September 2015.

Prüfungskommission:
Prof. Dr. David Schlangen (Betreuer und Gutachter)
Prof. Dr. Philipp Cimiano (Gutachter)
Prof. Dr. Marcus Kracht

Datum der mündlichen Prüfung: 16 März 2016.

Typeset mit LaTeX.

⊗ Printed on acid-free, aging-resistant paper (according ISO 9706). Gedruckt auf säure-freiem, alterungsbeständigen Papier (nach ISO 9706).

Abstract

The primary concern of this thesis is to model the resolution of spoken referring expressions made in order to identify objects; in particular, everyday objects that can be perceived visually and distinctly from other objects. The practical goal of such a model is for it to be implemented as a component for use in a live, interactive, autonomous spoken dialogue system. The requirement of interaction imposes an added complication; one that has been ignored in previous models and approaches to automatic reference resolution: the model must attempt to resolve the reference *incrementally* as it unfolds–not wait until the end of the referring expression to begin the resolution process.

Beyond components in dialogue systems, reference has been a major player in the philosophy of meaning for longer than a century. For example, Gottlob Frege (1892) has distinguished between Sinn (sense) and Bedeutung (reference), discussed how they are related and how they relate to the meaning of words and expressions. It has furthermore been argued (e.g., Dahlgren (1976)) that reference to entities in the actual world is not just a fundamental notion of semantic theory, but *the* fundamental notion; for an individual acquiring a language, understanding the meaning of many words and concepts is done via the task of reference, beginning in early childhood. In this thesis, we pursue an account of word meaning that is based on perception of objects; for example, the meaning of the word *red* is based on visual features that are selected as distinguishing red objects from non-red ones.

This thesis proposes two statistical models of incremental reference resolution. Given examples of referring expressions and visual aspects of the objects to which those expressions referred, both model components learn a functional mapping between the words of the referring expressions and the visual aspects. A generative model, the *simple incremental update model*, presented in Chapter 5, uses a mediating variable to learn the mapping, whereas a discriminative model, the *words-as-classifiers* model, presented in Chapter 6, learns the mapping directly and improves over the generative model. Both models have been evaluated in various reference resolution tasks to objects in virtual scenes as well as real, tangible objects. This thesis shows that both models work robustly and are able to resolve referring expressions made in reference to visually present objects despite realistic, noisy conditions of speech and object recognition. A theoretical and practical comparison is also provided.

Special emphasis is given to the discriminative model in this thesis because of its simplicity and ability to represent word meanings. It is in the learning and application of this model that gives credence to the above claim that reference is the fundamental notion for semantic theory and that meanings of (visual) words is done through experiencing referring expressions made to objects that are visually perceivable.

Acknowledgements

Upon arrival at Bielefeld, my knowledge about dialogue systems research was impoverished at best. Despite this, my advisor, David Schlangen, took a chance on me to fill a PhD position in a newly forming research group. Since then, he has been patient through my hastily written source code, incoherent papers, and stumbling through learning about how to do research, but somehow kept an open door, listened to my naive ideas, and made useful suggestions that led me to discover answers to questions on my own. I could not have selected a better advisor; he struck a perfect balance with me of supervision and autonomy. As a result I think we have both learned a thing or two and have made some contributions.

I am grateful to CITEC who not only funded my PhD position which provided enough income so I could be a full-time student and support my growing family, but also funded travel to conferences and granted me a 12-month extension. Thanks also to the CITEC administration, in particular Claudia Muhl. Thanks also to the administration of LiLi for getting me through the paperwork of being a foreigner studying in Germany.

I am grateful to Professor Philipp Cimiano for joining my dissertation review committee, as well as Professor Marcus Kracht for performing the role of defense chair.

Thanks also go to those, whoever you are, who have reviewed my papers, listened to my talks or took the time to visit my posters and demos at conferences and workshops. Questions and comments on my work have allowed me to be a better researcher and take a more critical look at my own work. Along a similar vein, thanks to all my co-authors for working with and putting up with me during long hours of paper writing and editing.

I am grateful to my colleagues, current and former, of the Dialogue Systems Group, namely Spyros Kousidis who got our lab up and running and provided our group with oft needed social relief; Ting Han, my kind and hard working office mate; Sina Zarrieß, Soledad Lopez Gambino, Birte Carlmeyer, Simon Betz, and Ivan de Kok–all who, at one time or another, provided interesting discussion on various areas of dialogue and listened to dry runs of my conference talks. Also thanks to our student assistants who tirelessly annotated and transcribed countless audio files, as well as their help with our many experiments and data collections. Special thanks to our delightful interns Livia Dia who got us started with computer vision and Ramesh Manuvinakurike who got us started with web-based dialogue.

Though also a group-member, I left out Julian Hough to isolate him for special thanks for fruitful discussions on reference, incremental processing, semantics, and the philosophy thereof. Thanks also to fellow participants and organizers of YRRSDS; in particular Matt Henderson and Nina Detlefs who, in their own small way and unbeknownst to them, helped me through a somewhat difficult time. Thanks also to Jana Götze for interesting discussions

on speech and language at conferences and via email correspondence. Thanks also to Timo Baumann who patiently answered questions about InproTK. Similar thanks to Pierre Lison for getting me started with OpenDial. I've gained insights, large and small, from all of you.

I am further grateful some dear friends from my LCT masters days who continue to influence me. Annemarie Friedrich continues to provide friendship through emails and fragmented meetings at conferences. Also Farina Freigang who helped me through my first few years at Bielefeld with regular discussions and will ever be my German teacher.

I must also thank the many members of the HLT Gemeinde Bielefeld, in particular Rüdiger and Doris Wagnitz, our dear neighbours who kept an eye on my family while I was off interning and conferencing. They went above and beyond the usual genial German hospitality and filled the role of adopted grandparents for our children. Speaking of internships, thanks to Kotaro Funakoshi, colleagues and supporting staff of Honda Research Institute in Japan for taking the time and effort to set up an internship position for me and seeing me through it. I gained valuable experience that continues to serve me. My love for Japan has greatly increased.

Special thanks to those who, in the final hours, added a pair of eyes to my thesis by looking over at least a couple of chapters or even specific sections: Julian Hough, Farina Freigang, Carrie Kennington, Patty Kennington, and Sina Zarrieß.

I am grateful to the facilities of Bielefeld University, not the least among them the Mensa which, on some days, was the high point; the libraries which provided a haven for reviewing papers, thinking about the philosophy of reference, and finding books and papers written by people much smarter than myself. Minor thanks go to Chopin, Ravel, Debussy, and Paul Cardall (among others) for helping me filter out unwanted noise. Also thanks to Kile, Eclipse, Bitbucket, DokuWiki, Trello, Evernote, Mendeley, Slack, and red crosses everywhere. We must not forget German taxpayers who graciously provided my fellowship funding.

Finally, thanks go to my family. To my in-laws Richard and Becky Jackson who helped us transition to life in Europe, find the beautiful places therein, and are ever examples of what it means to be good people. My own parents, who always saw to it that I did my best and provided the childhood scaffolding and continued support for me to do so. Thanks to my children, Elsa, Leah, Mary, and Isaac for always being happy to see me as I arrived home from work and being patient during my absences. Final thanks go to my wife Katie who has supported me through these past 10 years of married life. We have spent more time abroad than in our native U.S. and with meagre student income and the difficult task of learning foreign cultures and languages, she has made the best of things and provided the support I have needed to become and accomplish what I (we) have.

Relevant Publications

Portions of this thesis are based on previously published material, as specified in more detail in the following publications:

- Kennington and Schlangen (2012, 2014): Provided a basis for the method of representing objects as properties, as presented in Chapter 5.
- Kennington, Kousidis, and Schlangen (2013): Original definition of the SIUM model presented in Chapter 5.
- Kousidis, Kennington, and Schlangen (2013): An explanation of the tools used for gathering and analysing collected multimodal data, some of which is described in Chapter 4, used by the models presented in Chapters 5 and 6.
- Kennington, Kousidis, and Schlangen (2014c): A linguistically motivated extension of the model presented in Chapter 5 that uses a semantic representation.
- Kennington, Kousidis, and Schlangen (2014d): Explanation of an extension of InproTK to pass information between various processes; used for collecting data described in Chapter 4; models are implemented as InproTK modules, explained in Appendix A and B.
- Kennington, Dia, and Schlangen (2015a): Basis and original definition for the WAC model presented in Chapter 6.
- Hough, Kennington, and Ginzburg (2015): Desiderata for incremental semantics, applied the model presented in Chapters 5 for the evaluation.
- Kennington, Iida, Tokunaga, and Schlangen (2015b): Rigorous evaluation of the model presented in Chapter 5 on interactive data.
- Kennington and Schlangen (2015): Extended earlier work on the model presented in Chapter 6, adding more complex referring expression types; established a potential connection to a semantic calculus.
- Han, Kennington, and Schlangen (2015): Application of both models (Chapters 5 and 6) to a novel map-like task.
- Kennington, Lopez Gambino, and Schlangen (2015c): Application of the model presented in Chapter 6 in a realtime interactive dialogue system (demo paper; lightly reviewed).

Declaration

I declare that this thesis was written by myself and that the work contained therein is my own, except in those cases where it is explicitly stated otherwise in the text. This work has only been submitted to Bielefeld University and has not been submitted to another degree or professional qualification.

(Casey Redd Kennington)

Kurzfassung: Inkrementelle Referenzresolution für sichtbare Objekte in einem situativen Dialog

Angenommen, zwei Menschen befinden sich in einer Gesprächssituation, in der sie einige interessante Objekte um sich herum sehen. Sprecher (S) will über ein bestimmtes Objekt (I) reden. Bei S entsteht also die Intention, mit Hilfe eines referierenden Ausdrucks (U) die Aufmerksamkeit des Hörers (L) auf dieses Objekt zu lenken. Damit dies geschieht, muss Folgendes zutreffen:

1. S sieht das Objekt I
2. bei S entsteht die Intention mit L über I zu reden
3. S bezeichnet I mit Hilfe eines der folgenden referierenden Ausdrücke U (damit L I aus der Menge der möglichen Objekte identifizieren kann):
   (a) ein beschreibender Ausdruck (z.B. *das rote Kreuz*)
   (b) ein deiktischer Ausdruck (z.B. Zeigen mit dem Finger, *das*)
   (c) eine Kombination eines beschreibenden und eines deiktischen Ausdrucks (z.B. Zeigen mit dem Finger, *das rote Kreuz da*)
4. L sieht die Objekte in S und Ls unmittelbarer Nähe (einschließlich I)
5. L hört U (evt. sieht er die Zeigegeste)
6. L löst die Referenz U auf, indem er im Verlauf der Äußerung von S I aus der Menge der möglichen Referenten identifiziert

Für Menschen ist der Umgang mit sichtbaren, greifbaren Objekten alltäglich. Menschen interagieren jedoch nicht nur mit Objekten, sondern sie verwenden auch *referierende Ausdrücke*, um über sie zu reden, wie der Dialog zwischen S und L im oberen Beispiel zeigt. Das Referieren auf sichtbare Objekte geschieht in sogenannten *situativen Dialogen*. Der Prozess, der es dem Hörer L ermöglicht, den von S intendierten Referenten aus einer Menge von möglichen Objekten zu identifizieren, wird als Referenzresolution bezeichnet. Das Hauptziel dieser Doktorarbeit ist es, die Resolution von referierenden Ausdrücken zu modellieren, um dieses Modell als Komponente in einem Dialogsystem zu implementieren. Menschen bestimmen die Bedeutung von referierenden Ausdrücken oft völlig unbewusst und problemlos.

Für eine automatische, komputationelle Modellierung ist Referenzresolution jedoch ein komplexes Problem. Um den Referenten eines entsprechenden Ausdrucks aufzulösen, muss eine solche Komponente sowohl das Sprachsignal des referierenden Ausdrucks als auch Information über die sichtbaren Objekte repräsentieren, um zu bestimmen, welches der Objekte durch den referierenden Ausdruck bezeichnet wurde. Dabei ist entscheidend, dass ein Gesprächspartner nicht das Ende der Äußerung eines referierenden Ausdrucks abwartet, um die Referenz aufzulösen, sondern die Referenz inkrementell bestimmt, noch während der Sprecher seine Äußerung tätigt. Idealerweise würde eine automatische Komponente ebenfalls auf diese Weise funktionieren.

Die Hauptziele dieser Arbeit sind in den folgenden Stichpunkten zusammengefasst, geordnet nach ihrer Wichtigkeit. Die Modellierung der Referenzresolution soll Folgendes ermöglichen:

- referierende Ausdrücke inkrementell (bzw. Wort für Wort) aufzulösen.
- eine Zuordnung von Wörtern eines referierenden Ausdrucks auf sichtbare Objekte (die entweder direkt oder durch Zwischenrepräsentationen erfasst werden) in einem gegebenen Datensatz zu lernen
- Referenzresolution für ungesehene referierende Ausdrücke in neuen Dialogsituationen zu generalisieren
- definite Kennzeichnungen und deiktische Ausdrücke in einem einheitlichen Modell zu erfassen
- das Modell als Komponente in einem Dialogsystem zu implementieren
- das Modell zu , um zu zeigen, dass referierende Ausdrücke trotz möglicher Probleme in der Repräsentation des Sprachsignals oder der Objekte aufgelöst werden.

Traditionell basieren Modelle für die Resolution von referierenden Ausdrücken auf symbolischen Ansätzen, z.B. der Aussagenlogik, wobei die abstrakten Eigenschaften der Objekte vollständig und exakt bekannt sind, das Wissen über Objektklassen manuell definiert wird (z.B. gehört ein Objekt, das rot ist, der Klasse der roten Dinge an), und der Prozess der Referenzresolution erst beginnt, wenn der referierende Ausdruck vollständig bekannt ist. Diese Doktorarbeit schlägt zwei Modelle zur Referenzresolution vor, die über diese Ansätze hinausgehen. Beide Modelle behandeln referierende Ausdrücke in einer inkrementellen Weise (z.B. Wort für Wort) und beide Modelle lernen aus Daten eine probabilistische Zuordnung von Objekten in bestimmte Klassen. Diese Modelle werden im Folgenden kurz zusammengefasst.

Das erste Modell wird als das *Simple Incremental Update Model* (SIUM) bezeichnet. SIUM ist ein generatives probabilistisches Modell. Es modelliert die unabhängige Wahrscheinlichkeit

der Kandidatenobjekte I, des referierenden Ausdrucks U und die Eigenschaften R der Kandidatenobjekte. R hat die Funktion einer Zwischenrepräsentation zwischen I und U und ist eine Menge von Klassen, denen die Objekte zugeordnet werden können. Das Modell lernt diese Abbildung zwischen U und R mit Hilfe von Daten. Ein gutes gelerntes Modell würde sich dadurch auszeichnen, dass zum Beispiel das Wort *rot* der Eigenschaft in R zugeordnet wurde, welche die rote Farbe von Objekten ausmachen. Das Modell wurde eingehend in mehreren Referenzresolutionanwendungen evaluiert, wobei verschiedene Aspekte des Modells sowie die Sprache (Englisch, Deutsch, und Japanisch) variiert wurde. Das Modell erzielte gute Ergebnisse und ist robust, es kann mit einer verrauschten Repräsentation des referierende Ausdrucks, bedingt durch automatische Spracherkennung, umgehen. Des Weiteren kann das Modell auch in gewissem Maße mit Ungenauigkeiten in der visuellen Repräsentation der Objekte umgehen. Das Modell und die entsprechenden Evaluationen werden in Kapitel 5 ausführlich erläutert.

Das zweite Modell wird als das *Words-as-Classifiers* (WAC) bezeichnet. WAC ist ein diskriminatives Wahrscheinlichkeitsmodell und modelliert die bedingte Wahrscheinlichkeit eines Kandidatenobjekts I bezeichnet durch den referenzierende Ausdruck U, gegeben eine Menge von tieferen (nicht symbolisch repräsentierten) Merkmalen des Objekts. Das Modell benötigt keine Zwischenrepräsentation für I und U, sondern lernt die Abbildung direkt. Die Menge der Klassen ist nicht vorgegeben, sie ergibt sich durch der Menge der Wörter und damit aus den Daten, auf denen das Modell gelernt wurde. Die Basisaufgabe des Modells besteht darin, zu lernen, inwieweit jedes Wort in einem referierenden Ausdruck zu den Merkmalen eine Kandidatenobjekts *passt*. Das Modell funktioniert robust, auch mit einer verrauschten Repräsentation des referierende Ausdrucks und einer tieferen Repräsentation des Objekts. Das Modell und die Evaluationen werden in Kapitel 6 ausführlich erläutert.

Ein weiteres Ziel dieser Arbeit ist, die Bedeutung von Wörtern zu modellieren. Diese Fragestellung liegt nahe: um referierende Ausdrücke zu verstehen, welche Wörter verwenden, um Referenten zu bezeichnen, muss die Bedeutung der Wörter beschrieben werden. Das Problem der Referenz hat eine lange Geschichte in der Philosophie zur Bedeutung. Zum Beispiel hat Gottlob Frege (1892) zwischen Sinn und Bedeutung differenziert, wobei sich letztere auf das referierte Objekt bezieht. In der formalen Terminologie wird zwischen *Intension* (in Bezug auf Sinn) und *Extension* (in Bezug auf Bedeutung) unterschieden. In erster Linie interessiert sich diese Arbeit für die Extension, bzw. das Objekt, auf das referiert wird und dafür, wie der Prozess der Referenzresolution abläuft. Um jedoch einen referierenden Ausdruck aufzulösen, muss auch die Intension der entsprechenden Wörter gelernt und repräsentiert werden. In Bezug auf Dahlgren (1976) verfolgen wir die Annahme, dass das Intension durch Extension gelernt wird. Die Experimente in Kapitel 5 und 6 liefern für diese Behauptung glaubwürdige Evidenz, weil Wortbedeutungen angenähert werden können, wenn man lediglich Beispiele von

referierenden Ausdrücken und Informationen zu den entsprechenden Objekten zur Verfügung hat. Beispielsweise kann aus genügend Beispielen roter Objekte, die Bedeutung von *rot* gelernt und für bestimmte Anwendungen operationalisiert werden.

Diese Dissertation beginnt mit einem einleitenden Kapitel, die das Problem Referenzresolution und das Szenario des situativen Dialogs vorstellt. Der Fokus auf dieses Problem wird im gleichen Kapitel mit mehreren Beispielen aus dem Sprachgebrauch und der Sprachentwicklung motiviert.

Kapitel 2 erläutert den interdisziplinären Hintergrund der Arbeit. Es beginnt mit Dialogsystemen und schließt situativen und inkrementelle Dialogsystemen ein, bei denen zeitliche und räumliche Präsenz der Benutzer gegeben ist. Es folgt ein erster Versuch der Modellierung von Referenzresolution mit Hilfe der Aussagenlogik. Der Abschnitt endet mit einer Liste von Anforderungen, die der logische Ansatz nicht lösen kann, wenn er in einer praktischen Anwendung umgesetzt wird. Der folgende Abschnitt legt dar, wie einige dieser Probleme in einer sog. *grounded semantics* behandelt werden können. Der folgende Abschnitt bezieht sich auf die philosophische Referenztheorie, deren Grundideen eine Rolle dabei spielen, wie die Modellierung der Komponenten erfolgen soll.

Kapitel 3 bietet einen Überblick zur einschlägigen Literatur über Referenzresolution einschließlich der traditionellen, *"grounded"*, und inkrementellen Varianten. Außerdem wird Literatur aus verwandten Bereichen des Sprachverstehens, der Generierung natürlicher Sprache, und einigen anderen Forschungsbereichen, die sich Referenz beschäftigen, sowie andere Arbeiten zum Problem der Bedeutungssepräsentation diskutiert. Kapitel 4 erläutert die Daten, die für das Lernen und die Evaluation der Modelle in Kapitel 5 und 6 verwendet werden.

Im letzten Kapitel werden SIUM und WAC qualitativ verglichen. Wir haben gezeigt, dass beide Modelle robust funktionieren, auch wenn die Repräsentationdn des referierenden Ausdrucks und des Objekts verrauscht sind. Beide Modelle lernen eine "grounded" Bedeutung und funktionieren inkrementell, und erfüllen damit die Ziele dieser Arbeit. Das Kapitel schließt mit einer Zusammenfassung der Ergebnisse dieser Arbeit und zeigt der Bereiche für die weitere Untersuchungen auf. Zum Schluss werden abschießende Gedanken über Referenz und Bedeutung geäußert.

Contents

1 Introduction **17**

1.1 The Problem: Resolving References made to Identify Objects 17

1.2 The Setting: Situated Dialogue 19

1.2.1 Incrementality 21

1.3 Motivation 22

1.3.1 Referring Expressions in Noun Phrases 22

1.3.2 Reference in Semantics and Pragmatics 23

1.3.3 Situated Dialogue and Grounding 24

1.3.4 Reference in Child Development 24

1.4 Proposed Solution: Two Models for Resolving References 26

1.4.1 A Generative Model 27

1.4.2 A Discriminative Model 27

1.5 Scope and Aims of this Thesis 28

1.6 Outline of the Chapters 29

2 Background **32**

2.1 Incremental, Situated Spoken Dialogue Systems 34

2.1.1 Spoken Dialogue Systems 34

2.1.2 The Reference Resolution Component 35

2.1.3 Situated Spoken Dialogue Systems 36

2.1.4 Incremental Spoken Dialogue Systems 38

2.1.5 Incremental Computation 40

2.1.6 The IU Approach to Incremental Dialogue Processing 41

2.2 Types of Referring Expressions 46

2.2.1 Some Types we don't Consider 46

2.2.2 The Two Types we do Consider 48

2.3 Modelling Reference with First Order Logic 50

2.3.1 Syntactic Assumption 50
2.3.2 Definite Descriptions 50
2.3.3 Demonstratives (and Pronouns) 52
2.3.4 Functional Application of Objects 53
2.3.5 Limitations of Intersection 54
2.3.6 A Brief Survey of Other Semantic Approaches 56
2.4 Reference and Grounding 58
2.4.1 The Symbol Grounding Problem 59
2.4.2 Grounding, Semantics, and Probabilities 62
2.5 Reference and Meaning 64
2.5.1 Philosophical Background 64
2.5.2 Intension and Extension 67
2.5.3 Intension via Extension 68
2.6 Additional Assumptions 70
2.7 Chapter Summary 72

3 A Review of the Reference Resolution Literature 74
3.1 Previous Work in Reference Resolution 75
3.1.1 Traditional Approaches 75
3.1.2 Approaches with Grounding 83
3.1.3 Incremental Approaches 91
3.2 Natural Language Understanding and Reference 94
3.3 Generation of Referring Expressions 99
3.4 Reference in other Research 101
3.5 Related Work on Meaning Representation 103
3.6 Chapter Summary 104

4 Data 105
4.1 Pentomino Puzzle Tiles 105
4.1.1 ACTION 106
4.1.2 TAKE 107
4.1.3 TAKE-CV 110
4.2 The REX Corpora of Tangram Puzzle Dialogues 113
4.3 Airline Travel Information System 115
4.4 A Closer Look at the Data 115
4.4.1 Objects, Properties, and Features 116
4.4.2 Comparison of ACTION and REX 117

4.4.3 Gaze and Pointing Gestures in TAKE . . . 117
4.4.4 Spatial Language . . . 120
4.4.5 Reference Domains . . . 121
4.5 Chapter Summary . . . 122

5 The Simple Incremental Update Model: A Generative Model of Incremental Reference Resolution 123
5.1 Discovering SIUM in the IU-network . . . 124
5.2 Model Definition . . . 127
5.2.1 General Derivation . . . 127
5.2.2 Deriving an Incremental Model . . . 128
5.2.3 Examples . . . 132
5.2.4 Open Questions about the Model . . . 134
5.3 Experiment 1: Varying Representation and Grounding of Referring Expressions 136
5.3.1 Data . . . 136
5.3.2 Task & Procedure . . . 137
5.3.3 Abstracting over U: Robust Minimal Recursion Semantics . . . 137
5.3.4 Metrics . . . 139
5.3.5 Results . . . 139
5.3.6 Remarks . . . 141
5.4 Experiment 2: Fusion with Gaze & Deixis . . . 141
5.4.1 Data . . . 141
5.4.2 Fusing SIUM with Gaze & Deixis . . . 142
5.4.3 Task & Procedure . . . 143
5.4.4 Metrics . . . 143
5.4.5 Results . . . 144
5.5 Experiment 3: Fusing Deixis and Gaze as Properties under High Interactivity . 144
5.5.1 Metrics . . . 148
5.5.2 Results . . . 149
5.5.3 Analysis . . . 150
5.5.4 Remarks . . . 151
5.6 Experiment 4: Uncertainty in the Perception of the World . . . 151
5.6.1 Data . . . 151
5.6.2 Scene Processing . . . 151
5.6.3 Task & Procedure . . . 153
5.6.4 Metrics . . . 153

5.6.5 Results . . . 153
5.6.6 Systematic Insertion of Uncertainty . . . 154
5.7 Experiment 5: Using SIUM for Natural Language Understanding . . . 155
5.7.1 Data . . . 155
5.7.2 Task & Procedure . . . 156
5.7.3 Metrics . . . 156
5.7.4 Results . . . 157
5.8 Discussion . . . 157
5.9 Intension, Extension, and the Simple Incremental Update Model . . . 158
5.10 Chapter Summary . . . 159

6 The Words-as-Classifiers Model: A Discriminative Model of Incremental Reference Resolution 160
6.1 Model Definition . . . 161
6.1.1 Word Meanings . . . 161
6.1.2 Application and Composition . . . 163
6.1.3 Open Questions about WAC . . . 166
6.2 Experiment 1: Resolving References made to Virtual Objects . . . 167
6.2.1 Data . . . 167
6.2.2 Evidence from Gaze and Deixis . . . 168
6.2.3 Task & Procedure . . . 168
6.2.4 Metrics . . . 169
6.2.5 Results . . . 169
6.2.6 Further Analysis . . . 170
6.3 Experiment 2: Resolving References made to Tangible, Real-World Objects . . . 172
6.3.1 Data . . . 172
6.3.2 Task & Procedure . . . 172
6.3.3 Metrics . . . 174
6.3.4 Results . . . 174
6.3.5 Further Analysis . . . 176
6.4 Discussion . . . 179
6.5 Intension, Extension, and the Words-as-Classifiers Model . . . 180
6.6 Chapter Summary . . . 181

7 Closing Remarks: Comparisons and Outlook 183
7.1 Comparing SIUM and WAC . . . 183
7.1.1 General Comparison . . . 183

7.1.2 Comparison on Reference 184
7.1.3 Learning Curves 186
7.1.4 Semantic Comparison 187
7.2 Conclusion 189
7.3 Further Work 191
7.4 Parting Thoughts on Meaning and Reference 193

Appendices

A Implementation of SIUM 195
A.1 Overview 195
A.2 Java Implementation 196
A.3 InproTK Module 197

B Implementation of WAC 198
B.1 Overview 198
B.2 Java Implementation 199
B.3 InproTK Module 200

References 201

1
Introduction

All men by nature desire to know. An indication of this is the delight we take in our senses; for even apart from their usefulness they are loved for themselves; and above all others the sense of sight. For not only with a view to action, but even when we are not going to do anything, we prefer seeing (one might say) to everything else. The reason is that this, most of all the senses, makes us know and brings to light many differences between things.

- Aristotle, *Metaphysics*

1.1 The Problem: Resolving References made to Identify Objects

If you were to look around in your immediate vicinity, it is very likely that you will see various objects such as books, papers, pens, maybe a mug, a chair, or some kind of computer or laptop. If you were to find a window and look through it, you might see a tree, a building, or a car driving by. If you recall when you woke up this morning, there was likely some kind of bed, pillow, or alarm nearby. Recall the last meal you had; you ate something and probably used some kind of utensil to do so. These are all examples of things we call *objects*.

These are just a few examples of the seemingly uncountable objects that we as humans interact with daily. Objects are simply a part of our spatial existence. We use objects as tools to achieve some kind of goal (e.g., a pen for writing or a fork for eating), for visual aesthetics (e.g., a painting or a plant), for going from place to place (e.g., a bicycle), for recreation (e.g., a ball or a book), for practical purposes (e.g., the clothes we wear; our phones for communication), or any other host of reasons. We take it for granted that interacting with objects is a very common part of our human experience.

We don't just interact with objects, however; we also *talk about* objects with each other. Looking around you again, choose any object and you can probably effortlessly describe that object's particular shape, size, colour, and relative position to other objects such that you can easily direct another person's attention to that object–that is, you can easily verbally *refer* to that object. There are many settings where referring to objects is necessary; for example, in a meeting when a participant describes a figure portrayed in a presentation, during a meal when someone says *please pass the salt*, or on a walk through a forest when someone directs your attention to a specific type of plant by describing its colour and where it rests relative to a particularly large tree. *Referring* to visually present objects is a very common and very necessary daily occurrence for most of us.

For a more concrete example of referring to objects, consider the red circle (or dot, if you like) below.[1]

(1)

In the previous sentence, the words *the red circle below* were probably sufficient for you to draw your attention to the red circle in (1). The words in that sentence which make up *the red circle below* form a ***referring expression*** (RE) and the object to which they refer is *visually present* on this page.

Referring to objects by describing them, such as using the RE *the red circle below* is not the only way that visually present objects can be referred. Consider the RE in (2), also refers to the red circle represented in (1)

(2) a. **that** circle

This shows that we don't only use descriptive utterances to refer to objects, we also use pointing gestures accompanied by certain words known as *demonstrative pronouns* (e.g., *that*; the entire phrase is a *demonstrative description*). In all cases, there is a *first-mention* of an object where

[1] Throughout this thesis, we use red circles and green squares as examples. The circles are always red and the squares are always green.

the REs are uttered with the purpose of directing a listener's attention to that object.[2]

Referring expressions, the visually present objects that they refer to (in our example, the red circle itself portrayed in (1)), and how such a RE is resolved to its referred object, are the focus of this thesis. Specifically, the goal is to model this process sufficiently that a computational component can automatically resolve (i.e., identify) novel REs made to to visually present objects.

Even though resolving REs is a seemingly common and simple operation for humans, it is no trivial task for an automatic component to accomplish. How can the objects be perceived, recognised, and represented? How can the RE be "heard" and represented in the component? How can a component be modelled such that it can learn the relation between what is seen and what is heard? What general assumptions can be made about such a model in terms of what information is used or ignored? These questions are discussed and addressed throughout the course of this thesis. Another focus, though to a lesser extent, is to take a closer look at the *meaning* of the words that make up the REs. Indeed, if resolving REs is to take place, then some notion of word meaning needs to be represented. We will see that when it comes to learning and representing meanings of words, resolving reference to visually present objects is a good place to start.

This chapter describes the problem, the setting, and sets forth motivation for approaching the problem within that setting. A brief sketch of the proposed solution–two models that resolve REs automatically–is then given. At the end of the chapter, a brief synopsis of each of the remaining chapters is given, with notes to the reader. In the next section, we will see an explanation of the setting where the REs we are interested in will occur.

1.2 The Setting: Situated Dialogue

As put by Charles Fillmore (1981), p.152,

> ...the language of face-to-face conversation is the basic and primary use of language, all others being best described in terms of their manner of derivation from that base.

Certainly, things have changed since the publication of Fillmore's article in the way we as humans communicate with each other. The Internet and mobile phones connect us as never before with many choices in how we can communicate (e.g., via speech or text), and we don't

[2]Also of interest are pronouns (e.g., *it*) to refer to objects that have been previously identified, but continue to be under discussion.

even need to discuss how social media has changed the shape of easy, widespread communication. Yet even if every person had continual access to this vast, interconnected communications network, there are still numerous examples of face-to-face communication; for example, professors lecturing to students, a store clerk and a customer, a witness in a courtroom, two friends having a meal together, etc. The fact remains that, as Fillmore stated, face-to-face conversation is the *basic* and *primary* use of language.

This is taken a step further by Herbert Clark (1996) (from where some of the above examples of face-to-face communication were taken); focusing on face-to-face conversation between two people (p. 8-9). He notes that it is estimated that about one-sixth of the world's population are illiterate, and most languages evolved before the spread of literacy. Moreover, face-to-face conversation is the principle setting which doesn't require any social skills. Reading and writing take years of schooling, and many people never get good at either. Most people find non-standard conversational settings more difficult, such as telling jokes, giving lectures, or narrating stories. Face-to-face dialogue between two people has only minimal requirements, and is hence very basic, making it a good setting when looking at how REs resolve to visually present objects.

Continuing with Clark (1996), he notes several characteristics of face-to-face dialogue (p. 9):

- Copresence: the participants share the same physical space
- Visibility: the participants can see each other
- Audibility: the participants can hear each other
- Instantaneity: the participants perceive each other's actions at no perceptible delay
- Simultaneity: the participants can produce and receive at once and simultaneously
- Extemporaneity: the participants formulate and execute their actions extemporaneously, in real time

(There are other characteristics, but these suffice for our purposes here.) The above list of characteristics can be grouped into two categories: copresense, visibility, and audibility denote that a *space* is shared, whereas instantaneity, simultaneity, and extemporaneity denote that *time* is shared.

Sharing space such that participants can see and hear each other directly, and sharing time such that participants have no delay in perceiving each other is the setting of dialogue we use for the purposes of this thesis. Indeed, in order to refer to visually present objects,

the participant performing the RE, and the participant resolving that reference both need to be able to visually perceive that object directly. Importantly (as is shown below) non-linguistic cues from the participant performing the RE, such as pointing to it or looking directly at it, give important information to help the other participant resolve the reference. This is done most commonly in face-to-face dialogue settings.

In the previous section, we saw two types of REs that we are interested in. It is in this context of face-to-face dialogue where space is shared that we encounter both. For example:

(3) a. N: please hand me **the book on the left**
b. K: **this** one?
c. [pointing to object]

(3) is an example of a typical interaction between two participants in which we see (each in bold typeface) a descriptive reference in (3-a) and a demonstrative reference in (3-b), both of which refer to a particular book that is visually present to both participants; moreover they can hear and see each other, which is necessary for success in the demonstrative RE in (3-b) accompanied by the pointing gesture in (3-c).

In this thesis, we refer to this kind of co-present, face-to-face dialogue as (following the literature) *situated* dialogue. Certainly, other types of dialogue that are not face-to-face can be considered "situated"; e.g., two participants speaking to each other via telephone with some distance between them still have a situated context of being on the same planet, but for this thesis, situated dialogue is of the face-to-face, co-present variety where time and space are shared.

1.2.1 Incrementality

Focusing now on shared *time*: during dialogue, a dyad of humans participate in a give-and-take of speech, where one of the participants plays the role of the *speaker* and the other plays the role of the *listener* and these roles switch fluidly throughout the dialogue. When a speaker is performing some kind of utterance like a RE, the listener is *incrementally* comprehending that utterance as it unfolds; i.e., the listener does not wait until the end of an utterance before she begins processing it (Tanenhaus and Spivey-Knowlton, 1995; Spivey et al., 2002). For Example, in (3-a), as N speaks it is not the case that K sits idly, waiting for some kind of signal that N has finished speaking. Rather, K resolves the reference as it unfolds, taking in each sound and word, possibly walking towards the referred book during N's utterance or reaching out her hand to begin the pointing gesture before she even utters the expression in (3-b).

This has implications in modelling the resolution of REs: such a model should be able

to process input incrementally (e.g., word by word), updating its internal representation of the ongoing comprehension process. Previous approaches to modelling the resolution of REs has not taken this important constraint into account. Incrementality and its implications in a component that resolves RE to visually present objects is explained in greater detail in Chapter 2 as part of an explanation of incremental spoken dialogue systems.

1.3 Motivation

We now turn to motivating *why* one should put focus on resolving REs to visually present objects and why situated dialogue is an ideal setting for this kind of task. In this section, we will see that REs are commonly used in our everyday interactions, that the study of reference has very deep roots in semantics, that situated dialogue is a natural setting to model meaning of words as they are used in context, and, finally, that this task of resolving references and the setting of situated dialogue are also essential in child language development.

1.3.1 Referring Expressions in Noun Phrases

As noted above, objects are ubiquitous and we as humans talk about these objects in every day speech. The types of words we use to describe these objects which make up REs fall under the category of definite descriptions. Poesio and Vieira (1997) explained that definite descriptions, which are a specific type of noun phrase (NP) that begin with the article *the* (e.g., *the red circle*), are one of the most common constructs in English. We put focus on a subset of these kinds of NP constructions. Consider the examples in (4), taken from Poesio and Vieira (1997):

(4) a. please, pass me **the salt**
b. don't break **the vase**
c. beware of **the dog**
d. mind **the step**

The examples in (4) make up only a specific type of definite description (*immediate use*), where visually present objects are described. In this thesis we consider the meaning and usages of such NPs. Though we are also interested in demonstratives and, to a lesser degree, pronouns (of a specific type), of greater interest to us are these kinds of definite descriptions, where the words that make up the RE are chosen specifically to draw an interlocutor's attention to that object. This is a commonly-occurring phenomenon in language use, making it a good candidate for our attention in dialogue systems research.

1.3.2 Reference in Semantics and Pragmatics

In her recent book, Barbara Abbott (2010) explains the main lines of thought regarding reference: "...confronting what is arguably the most basic question of all for semantic and pragmatic theory: what is the link between words and the world?" (see Preface). She begins with some of the ideas of John Stuart Mill, Gottlob Frege, Bertrand Russell, and later considers the ideas of Keith Donnellan (among others), how their ideas overlapped and where they were at odds. She shows where some of these ideas fit into Richard Montague's well-known semantic formalism (Hobbs, 1983).

Though the above-mentioned individuals disagreed in varying degrees to the meaning of proper names (e.g., *Aristotle*, *New York*, etc.), there is a general uniformity of thought in that something to which an expression refers is different from the fundamental meaning of the words that make up that expression. Consider the well-known example in (5):

(5) a. The morning star is the evening star.

There are two REs, *the morning star* and *the evening star*. Both actually refer to the same object (the planet Venus), but the words that make up the individual REs have different individual meanings, as do the individual REs. This difference between meaning and reference is discussed in greater detail in Chapter 2.

The focus of Abbott (2010) is put on the intersection of semantics and philosophy; examples are generally complete sentences that don't have any particular discourse context, though how this fits into a more dialogue-oriented framework (e.g., the ideas of Paul Grice and John Searle) is also given some attention. However, in this thesis, we will see many examples of non-sentential, indeed non-grammatical REs which are a product of spontaneous speech. Some examples:

(6) a. red ... red cross top left
b. green circle here (with potential pointing gesture)
c. left of blue book that one

Though ofttimes ungrammatical, such REs can still be semantically composed (Schlangen, 2004) (composition (Frege, 1892) is discussed in Chapter 2), and since it is the goal of this thesis to construct a model that works as a practical component, this kind of spontaneous speech needs to be taken into account.

The good news is, this is a well-studied and debated area, and we have solid ideas with ample foundational and recent work upon which we can build. Yet, as we will see in later chapters in this thesis, applying these theories to a practical, automated component that can

resolve utterances to objects in a specific context is a challenging task.

1.3.3 Situated Dialogue and Grounding

In this thesis, we are interested in the notion of meaning as it pertains to the words in the REs that we encounter. Referring to visually present objects is a good place to look at meaning: it limits the scope of processing natural language to the specific phenomenon of definite descriptions (as well as demonstratives). Such descriptions can be fairly simple (e.g., *the dog*) or quite complex (e.g., *the red one next to the green one on the bottom left*), giving us a range of linguistic phenomena with which we can work, without leaving our area of interest. An individual (or, for the purposes of this thesis, an automatic component) resolving the RE *the red circle* to the red circle represented in (1) amounts to recognising each individual word, i.e., *the*, *red*, and *circle*, and how they work together to form a complete RE such that a person or component can resolve the intended red circle in (1).

This leads us to the notion of *grounding*, where the symbols of a language (e.g., the words) are linked somehow to perception (Harnad, 1990).[3] Resolving REs to visually present objects is an excellent starting point for this: words that are used to refer to (i.e., describe) objects have some kind of relation to the object itself; otherwise, one must assume, a rational speaker wouldn't have said those particular words to describe that particular object. A component that models that relation can potentially capture the meaning of that word (i.e., *meaning* as words are used) in observing how a word is used in a specific context to refer to a particular object. Of course, this only considers the meaning of words and REs that have some kind of link to perception; other more abstract notions that can be talked about with language (e.g., *health* or *intelligence*) which arguably do not link directly to visual perception, is not the focus here. Yet this "low-hanging fruit" of REs allows us to capture perceptually-based meanings of words and how they relate to the objects with which we interact. We consider these ideas in greater detail in Chapters 2, 5, and 6.

1.3.4 Reference in Child Development

In Wittek and Tomasello (2005), it is claimed that:

> Among children's earliest communicative attempts are acts indicating objects for other people, for example, pointing to an object or holding up an object to show it. Once language begins, children rapidly acquire a host of additional linguistic

[3] Not to be confused with *building common ground* (Clark, 1996) which is also referred to as *grounding*, though they are related.

means for indicating objects, mainly in the form of various kinds of nominals (noun phrases).

Human babies often use gestures to communicate before they use words, particularly pointing gestures (e.g., with an index finger or, in some cultures, adults point with lips) which are produced to direct attention to a particular object (Butterworth and Morissette, 1996; Goldin-Meadow, 2007). It can be argued that this gives human babies the ability to focus on particular objects that are made salient by means of pointing. Thus when a pointing gesture occurs thereby directing a baby's attention to an object, and a co-occurring description of that object is uttered, the baby recognises (to some degree) that the sounds of that utterance have some kind of relation to the object that was being indexed by the pointing gesture. With enough stimuli, that baby can begin to pick out sound patterns that denote certain aspects of objects (e.g., the sound pattern represented by the word *ball* co-occurs highly with seeing a round shape or the action of smooth rolling). Further down the developmental road, certain features can be grouped into types (e.g., *blue* and *red* are grouped as *colours*), allowing further abilities to use words that mean certain things in different contexts. For example, a child that has seen a green ball before, with a co-occurring RE that refers to it (e.g., *the green ball*) can extrapolate later when she sees a red ball that it is still referred to as a ball, though with a different colour term.

To illustrate more concretely that small children use REs early in their spoken language development, we turn to CHILDES, a corpus of collected audio (and video) of interactions between children and their caregivers. The following is an excerpt from one interaction (C denotes the child–between 1 and 2 years of age–P denotes the caregiver; at the time of the interaction, the caregiver is taking purchased groceries from bags and putting them away):[4]

(7) a. P: What's that?
b. C: bread bread
c. P: That's bread! Let's see, oh, what's this?
d. C: cook
e. P: Cookies, that's right. Hmm, let's see here, what else. What is that?
f. C: nana

As (7) illustrates, P referred to visually present objects using demonstratives. P was eliciting C to name the objects (which isn't necessarily drawing the child's attention to the object, rather the child is naming the object that is already under discussion); the focus here is that visually present objects are important when children are learning how to use language in that the visual aspects of the objects were available as features when the word that names a particular object

[4]CHILDES/Eng-NA/Davis/Rebecca

was being used.

Without this kind of interpersonal verbal interaction, a child can have difficulties in language acquisition and speech development. For example, in Sachs et al. (1981) the authors studied two children whose only exposure to spoken language was by watching television (their parents were both deaf, but the children had normal hearing capabilities). The authors evaluated the abilities of each child and found that their abilities to produce and understand speech were far below average for their respective ages.

The type of interaction between child and caregiver that is necessary for normal development in the language capabilities of children is a type of situated dialogue as we have described earlier: the two interlocutors are co-present and non-linguistic pointing gestures are observable by the child. Importantly (though not a direct focus of this thesis), the caregiver can provide positive feedback when a child produces some kind of notion of understanding by action (e.g., grasping the object that was referred, or by repeating part of the RE; as in Example (7)), or negative feedback when the child utters a word that is unrelated to the referred object (knowing about "positive" and "negative" examples like this are also important for a component to learn how certain words co-occur with certain visual features, as shown in Chapter 6).

Though this developmental aspect of learning REs is not the main focus of this thesis, clearly a closer look at objects, their descriptions, and (optional) co-occurring pointing gestures is a good starting place for learning how to build a representation on how words are used in context, which has implications to their individual and composed meanings–which, it appears, is not unlike how humans begin to learn how language is used.

1.4 Proposed Solution: Two Models for Resolving References

Previous work in reference resolution (presented in Chapter 3) have in many cases assumed a symbolic representation of word meaning; i.e., have not learned a grounded semantics. For those that do, they are generally not modelled or implemented to work incrementally, as described above. Moreover, those that are grounded and incremental, have not attempted to fit their models into a formal framework. The models presented in this thesis attempt to overcome these shortcomings.

Chapters 5 and 6 explain two approaches that address the above-stated problem: how to model the resolution of REs to the objects which they are intended to refer. A sketch of these two approaches is presented in this section. Both models are probabilistic; that is, they are given data (i.e., examples of a set of representations of objects, REs, and the objects that are referred), then the models learn, each in their own way, patterns between how words in the REs are used to refer to certain objects. Details, including evaluations and discussions, are given in

their respective chapters.

1.4.1 A Generative Model

The first model is a generative approach with the following generative story: given a representation of the set of objects and the set of properties of all of those objects (e.g., colours and shapes), a speaker will produce a RE to a particular object by picking out properties that belong to that object, and uttering words that have been observed to co-occur with those properties. Using a generative model for comprehension (i.e., as a listener that resolves) amounts to "turning it on its head" by observing the words of the RE and working backwards.

In order for this model to work, there are some important assumptions. First, a *rationality* assumption that the speaker will only utter words that highly co-occur with properties that the object has (as perceived by the speaker); taken together those words should allow the listener to uniquely identify the referred object. Second, that the set of properties is pre-defined; i.e., all of the possible colours, shapes, and spatial positions that objects can have are known beforehand.

Chapter 5 shows, through a series of experiments, that the model works well under varied circumstances. The model learns a mapping of words (i.e., grounds) in REs with objects via the visual properties of those objects, which act as a mediating variable. Importantly, the model can make use of non-linguistic information from pointing gestures and gaze (of the speaker), allowing the model to handle demonstratives, and it can also handle REs that use pronouns (which refer to objects that were previously referred via a definite description). The set of properties is a two-edged sword: the properties that an object has can be very dynamic across a dialogue (e.g., an object that begins on the left could be moved to the right, changing its properties), but the need to specify beforehand the set of properties requires time and development. The model works well when there is a symbolic representation of the objects (e.g., a virtual game board). More qualities and shortcomings of the model are explained and discussed in Chapters 5 and 7. This model is called the *simple incremental update model* (SIUM).

1.4.2 A Discriminative Model

The second model follows directly from the shortcomings of the first generative model, but builds off some of its strengths. Perhaps most importantly, it doesn't need a set of properties to map between objects and words in REs. As its basis, the model learns how well the low-level features that make up an object (e.g., RGB values) "fit" a particular word in a RE; for example, given the low-level features, how well does a particular object fit the word *red*? This model treats each word in a corpus (i.e., a set of REs, the objects in the scene in which the REs were used, and knowledge about which object was referred) as an individual classifier that learns

(i.e., grounds) from the positive and negative examples which low-level object features are the most distinguishing for that word. For example, the classifier for the word *red* should learn (specifically, perform feature selection) that high R values in the RGB colour values is more discriminative than other features. The model is discriminative because, given information about the objects, the model resolves the intended object directly (i.e., a conditional probability). This model is called the *words-as-classifiers* (WAC) model because each word is treated as an individual, logistic regression classifier.

Chapter 6 shows that the model is able to resolve objects in most cases, even when given a low-level representation of real, tangible objects (i.e., not virtual) in a noisy environment. It improves over the generative model when the representation of the objects contains uncertainty.

1.5 Scope and Aims of this Thesis

The goals of this thesis are itemized in the following bullet points in order of significance:

- Model the task of reference resolution such that:
  - The model can resolve referring expressions incrementally (i.e., word by word).
  - The model learns, given data, a mapping between visually present objects (represented either directly or by some mediating variable) and words in referring expressions.
  - Given novel REs and novel scenes, the model can generalise and resolve under new circumstances.
  - The model can handle definite descriptions and demonstrative references in a single framework.
  - The model can be implemented as a component in a spoken dialogue system.
  - The model can be evaluated to show that it resolves references despite possible issues (e.g., noisy conditions) with the representation of the referring expressions or the objects.
- Formulate the model in such a way that word meanings can be accounted for (what is meant by word meanings is explained in Chapter 2).
- Fit the model into a semantic framework (e.g., formal logic and grounded).

In short, the primary goal is to produce a component of reference resolution that works incrementally and can be used in an interactive dialogue setting. To a lesser extent, but of no

less importance, is that doing so requires, to some degree, that the meanings of the words that make up the REs are in some way learned and represented. With such meanings in hand, it could prove useful to fit them into a larger semantic framework, such that they can be applied in other situations. Though not the main focus of this thesis, we return to this throughout in subsequent chapters. As noted, this thesis proposes two such models. How they compare and how well they fit the above goals is explained in the final chapter.

1.6 Outline of the Chapters

This introductory chapter ends with an outline of this thesis, complete with an overview of each chapter and notes to the reader.[5]

Chapter 2: Background

The chapter on background is broken into three main sections. First, incremental spoken dialogue systems. This section is dedicated to spoken dialogue systems, what they are specifically, what it means for a spoken dialogue system to be situated and incremental, and how the models described in this thesis could fit into a spoken dialogue system. We then see a definition of the types of REs that we will consider (and some examples of those that we don't). We then see and initial attempt at modelling the resolution of REs using a logical formalism and the shortcomings of that formalism. Fourth, grounding. In this section, we see what is meant by grounding and give a review of some other tasks and models that have applied grounding and how that can possibly overcome the shortcomings explained in the prior section. Finally, meaning and reference: What do semantics and pragmatics have to say about reference? Who are the major contributors to the theories? What specific phenomena are we going to work with and what do they have to do with resolving REs? These questions are addressed in this background section.

Readers who are already familiar with the background of spoken dialogue systems, semantics and grounding, as well as the philosophy of reference and meaning can skip this chapter, but note that subsequent chapters refer back to ideas in this chapter.

Chapter 3: A Review of Reference Resolution Literature

In this chapter we look at other work in reference resolution, in particular those who have implemented and evaluated components in spoken dialogue systems. We also determine

[5] While the tone of this introductory chapter has been folksy, I use "we" in many cases to give credit to colleagues with whom the theories, derivations, and experiments were produced. I only use "I" in cases where I assert my own views.

what kind, if any, of grounding these other approaches are using and if their approaches are incremental. We see that there is ample work in automatic reference resolution, giving us a solid foundation upon which to build. However, we see that there are areas for obvious improvement particularly in the areas of learning grounded meaning and incremental processing, which motivates the need for the work described in this thesis.

Readers who are familiar with research in reference resolution and NLU can skip this chapter.

Chapter 3: Data

In this chapter, we look at the data that are used in the experiments described in Chapters 5, 6, and 7. This includes several corpora that make use of Pentomino puzzle pieces (including multi-modal data from eye gaze and pointing gestures), tangram puzzle pieces (with accompanying eye gaze), and the well-known ATIS corpus.

Readers who are interested in either of the models presented in Chapters 5 and 6 should read this short chapter first, as they are evaluated using the data described in this chapter.

Chapter 5: The Simple Incremental Update Model: A Generative Model of Incremental Reference Resolution

This chapter contains an explanation of the *simple incremental update* generative model of incremental reference resolution. The model works incrementally, learns a grounded mapping between words in REs to visual properties of objects, can take contextual priors into account, can handle definite descriptions, demonstratives, and pronouns in a single framework, and is robust to noise coming from the automatic transcriptions of REs. The model is extensively evaluated in German, Japanese, and English tasks under varied conditions to test the separate parts of the model. Though the model is robust when resolving references to scenes that have no uncertainty (e.g., virtual scenes on computer screens), its ability to handle uncertainty in the representation of the objects leaves room for improvement. The strengths and weaknesses of the model are presented as well as discussion on where the learned word meanings fit into formal semantics.

Chapter 6: The Words-as-Classifiers Model: A Discriminative Model of Incremental Reference Resolution

This chapter contains an explanation of the *words as classifiers* discriminative model of incremental reference resolution, how it builds upon the generative model, and how it improves beyond the generative model. The model does not need a set of properties as

was the case with the generative model; the meaning of words is represented and applied in the words of the REs, and the model can handle more complex REs that pick out a referred object by using a landmark object as a reference point (e.g., *the red circle next to the green one*). The model is robust to noise in both the transcription of the REs and the representations of the objects. The model is evaluated in two reference resolution tasks. Some investigation shows that the meaning (i.e., semantics) of some words is indeed learned and represented. Though at the writing of this thesis, the model does not handle reference via demonstratives directly, we show that it can be fused with another model that does, and further explore some possible ways that the model could be made to handle demonstratives directly.

Chapter 7: Closing Remarks: Comparisons and Outlook

This chapter compares the two models presented in Chapters 5 and 6 in general, within a reference resolution task, how they perform incrementally, and how they fit into the semantic framework set forth in Chapter 2. The latter part of the chapter concludes the thesis, gives a recap of what was shown, and offers some parting thoughts on reference and meaning.

Appendix A: Implementation of SIUM

This appendix explains how the generative model explained in Chapter 5 is implemented in Java and as a component of a spoken dialogue system framework.

Appendix B: Implementation of WAC

This appendix explains how the discriminative model explained in Chapter 6 is implemented in Java and as a component of a spoken dialogue system framework.

2
Background

Semantics is about the relation of words to thoughts, but it also about the relation of words to other human concerns. Semantics is about the relation of words to reality–the way that speakers commit themselves to a shared understanding of the truth, and the way their thoughts are anchored to things and situations in the world.

- Steven Pinker

To begin this chapter, suppose that two people are walking together along a path and find themselves among some interesting looking objects, as depicted in Figure 2.1.

The speaker (indicated by S in the figure) finds one object particularly interesting (indicated by I in the figure) and wants to talk about it with his friend, the listener (indicated in the figure by L). In order to begin talking about the object, S must first draw L's attention to that object with some kind of referring expression (indicated by U in the figure). The overall progression of what must happen in order for S to draw L's attention to I is outlined in the following:

1. S perceives object I
2. S forms the intention of talking about I with L

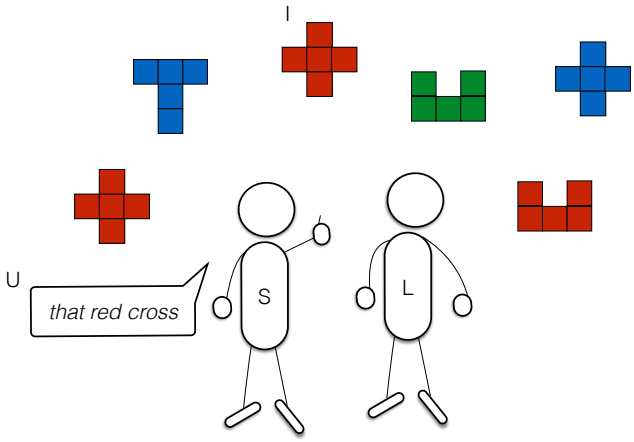


Figure 2.1

3. S initially indicates I to L in one of the following ways (via an indication event U):
   (a) a descriptive phrase (e.g., *the red cross*)
   (b) a demonstrative phrase (e.g., while pointing, *that*)
   (c) a combination of descriptive and demonstrative phrases (e.g., while pointing, *that red cross*)
4. L perceives the objects in S and L's immediate vicinity (including I)
5. L hears U (and visually perceives the optional pointing gesture)
6. L resolves U as it unfolds to isolate I from the other objects as the one indicated by S

The overall goal of this thesis is to produce a model and implement a component of that model that can perform L's part of the task as illustrated in the final three steps above. This chapter breaks down the steps in the above example and examines specific parts to form an interdisciplinary background. To begin, the model/component needs to be approached from the broader setting of where it is used, as in the example above between S and L: situated dialogue; in particular, *spoken dialogue systems*, and where such a component would fit into a dialogue system. This is explained in Section 2.1. In the remaining sections, we turn our attention to what is necessary for L to perform the task of resolving a reference made to an object. In Section 2.2 we will see the types of referring expressions that this thesis focuses on, namely first mention REs: definite descriptions and demonstratives. In Section 2.3 we focus on

L's final two steps and look at a more formal approach to resolving references using a semantic formalism: *first-order logic*, which allows us to bridge the gap between a conceptualisation of the model and an implementation of it. We will see that first-order logic has some limitations when applied to a practical model/component, but those limitations are addressed in the section that follows by ideas from *grounded* semantics–thus all three of L's steps are addressed and a component can then be modelled.

Of lesser focus in this thesis, but not necessarily of lesser import, is Section 2.5 which explores relevant philosophical literature on reference and meaning, which will help the models fit into the larger scheme of ideas. We return to some of these ideas throughout the thesis. As is shown in Chapters 5 and 6, the presented models of reference resolution give credence to some of the fundamental theories on reference and meaning presented. Finally, this chapter concludes with a listing of several assumptions we must make in order for the model/component to be realised.

2.1 Incremental, Situated Spoken Dialogue Systems

In this section, we look at the context in which reference takes place: spoken, situated dialogue. In the example at the beginning of this chapter, this includes both S and L's contribution to the (albeit short) dialogue.

2.1.1 Spoken Dialogue Systems

As explained in Chapter 1, a *spoken dialogue system* (SDS) is a computational agent that can converse with human beings through everyday spoken language (see Lison (2013)). The most basic form of a SDS requires some way of representing the human user's utterance, usually by attempting to transcribe the speech into written words through an automatic speech recogniser (ASR). That (attempted) transcription is then fed into a natural language understanding (NLU) component that abstracts over the transcribed utterance (e.g., via a syntactic or a semantic representation). That representation is then given to a dialogue manager (DM) which determines the next action to take (e.g., ask some kind of clarification request, or look up requested information in a database). The action often results in the need to utter something back to the human user, which is the job of the natural language generation (NLG) component (e.g., generation of the response utterance, then the actual synthesizing of that utterance). This procedure is visually represented in Figure 2.2.

This kind of SDS setup is typical of the kind of dialogue that would take place over a phone where the only modality of communication between two participants is speech, which is what is typically under focus in SDS research. There is a connection between the dialogue manager

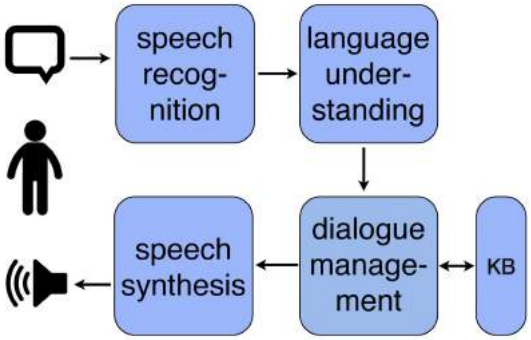


Figure 2.2: Example of a standard spoken dialogue system: when a user speaks, a speech recogniser provides input to a language understanding module, which creates some kind of abstraction over the input that is given to the dialogue manager, which makes a decision on an action to take (potentially based on information from a knowledge base), which generally involves the generation of speech to the user.

and some kind of *knowledge base* (KB) which is a set of facts about the world that the SDS can say something about. For example, if the SDS is a phone-based system that aids users in finding information about booking airline flights, the knowledge base would need to have flight information; e.g., origins and destinations, flight times, information about airports, etc. When a user speaks into his phone, that signal is transmitted and received by the system and passed through the ASR, which gives its hypothesis to the NLU which attempts to form a meaning representation over the utterance. That is given to the DM which determines the best course of action given that utterance, for example, if a request for flight information from Chicago to Atlanta is asked for, the DM would need to look up that information in the KB and then inform the NLG to produce an answer based on the results of that information. These types of systems, and the individual components that make up these kinds of systems, have been well-studied. For example, the *Air Travel Information System* (ATIS) corpus (Hemphill et al., 1990; Dahl et al., 1994) was produced to develop and evaluate NLU. In Chapter 3, we look closer at the NLU component, related literature, and how resolving references relates to NLU.

2.1.2 The Reference Resolution Component

In terms of a SDS component, reference resolution (RR) is the task of resolving REs to the referent. At its highest level of abstraction (that takes all three of L's steps as outlined in the

example in the beginning of this chapter), this can be formalised as a function f_{rr} that, given a representation U of the RE and a representation W of the (relevant aspects of the) world (which can include aspects of the discourse context), returns I^*, the identifier of one the objects in the (non-visual) world that is the intended referent of the RE:

$$I^* = f_{rr}(U, W) \tag{2.1}$$

This function f_{rr} can be specified in a variety of ways. Recent work has used stochastic models using the following approach: given W and U, the goal of RR is to obtain a distribution over a specified set of candidate entities in that world, where the probability assigned to each entity represents the strength of belief that it is the referred one. The referred object is then the argmax of that distribution:

$$I^* = \underset{I}{\operatorname{argmax}}\, P(I|U, W) \tag{2.2}$$

A RR component could replace the NLU component depending on the task, or it can be a sub-component of NLU, performing the difficult task of resolving references while NLU handles other processes that produce semantic abstractions over utterances (which could also be useful to a RR component).

As useful as these systems can be practically, as well as in terms of researching how language is used, they aren't sufficient to handle the types of phenomena in dialogue that this thesis explores. As noted in Chapter 1, we need to handle the fact that (1) the space is shared as the participants are co-located, and (2) the time is shared as the participants fluidly take turns and comprehend utterances as they unfold. The standard SDS in Figure 2.2 is not fully amenable to these conditions. In the following sections, we look at the variants of SDS that *are* amenable to these constraints; for situated SDS, and for incremental SDS.

2.1.3 Situated Spoken Dialogue Systems

In a situated SDS, speech isn't the only modality used for communicating between the system and the human participant. As noted in Chapter 1, situated dialogue denotes co-location; the participants can see and hear each other, and they are able to see objects in their shared space. A SDS that can replace one of the participants in such a setting needs to go beyond just processing speech; it also needs to have a representation of the visual situation which, following

Equation 2.1, we call the world W, and it has to be able to observe non-linguistic, yet communicative cues from its interlocutor, namely (for the interests of this thesis) pointing gestures. In other words, the SDS needs some notion of *situational awareness*.

Figure 2.3 shows visually how such a SDS might look: as before, there is ASR and speech synthesis, but the NLU component (which is now a RR component) also has information about the interlocutor's pointing gestures (denoted as *Deixis*), and the KB is replaced by a representation of the world W (though the manner of the representation of the KB and W could be the same).[1]

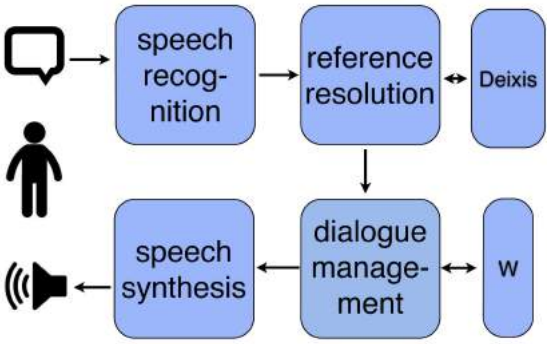


Figure 2.3: Example of a situated dialogue system that has a representation of how the world (W) is represented (similar to the KB in Figure 2.2), also there is a way of recognising the pointing gestures of the human participant.

The representation of W here deserves some additional attention. If goal of RR is to resolve REs made to an object that exists in the immediate space–a visually present object–then a component that resolves RRs needs to know about those objects. That is, W needs to somehow represent individual objects that appear in the scene (how this is done is explained in later chapters, and depends on the way the world is grounded with the language). Those objects are distinct from each other as they are represented, and some kind of visual information about them (either raw features or a computed set of visual properties) also needs to be present.

The component that gives deictic information to the RR component also deserves some additional explanation. Such a module has a fairly big job to do: it must somehow take information about the human interlocutor, and determine not only if that human is performing a pointing gesture, but to what object that human is pointing, or at least some kind of approximation thereof.

[1] One could see a situated dialogue system as one that doesn't just have ears, but also has eyes that can decipher visually present objects and decipher what the human interlocutor is doing.

Modelling W and a deictic gesture are both non-trivial tasks and add to the uncertain nature of the ongoing dialogue. However, they are necessary in a dialogue system that is situated, which is the kind of dialogue that we want the RR component to be able to handle.

Motivation Situated dialogue systems are essential to tasks that require a system to have some awareness about the immediate situation. For example, when driving a car, research shown that when the driver speaks to another person using hands-free devices (e.g., hands-free cell phones) there is a decrease in driving performance (He et al., 2013; Horrey and Wickens, 2006; Ishigami and Klein, 2009; McEvoy et al., 2005). Even just listening to speech causes increased cognitive load on the driver (Demberg et al., 2013). This is not the case, however, when drivers speak to passengers (Drews et al., 2008a), as passengers are aware of the driving situation and can adopt strategies that help the driver perform both the primary task of driving and of the dialogue with the passenger (Drews et al., 2008b). It was shown in Kousidis et al. (2014); Kennington et al. (2014b) that a situationally aware (i.e., situated in that the road and driving conditions were accessible to the system) SDS does not inhibit the driver's driving abilities because the system can respond to situations that require the driver's attention by interrupting its own speech (which also requires incremental processing, which is explained below).

Another practical area is robotics. When embodied robots interact with their environment, they can be made to interact with humans through speech. The type of SDS they would need is a situated SDS, as a robot would need to have some kind of representation of its surroundings, including information about human interlocutors. See, for example Chai et al. (2014); Kennington et al. (2014a).

2.1.4 Incremental Spoken Dialogue Systems

Another aspect of situated dialogue that a standard SDS doesn't typically handle is the fact that dialogue occurs in real-time and is highly interactive. Consider the following example:

(1) a. J: So Sarah ... I hear she has a new dog
b. K: yeah
c. K: She does. In fact, I was there when she bought it
d. J: oh?

J brings up the topic of Sarah's potential new dog. K gives *feedback* (i.e., a verbal indication of understanding; e.g., a back channel) in line (1-b) for J's utterance in (1-a) but does not attempt to take the floor (i.e., take a turn as the speaker). Such feedback is a way for K to signal

to J that she has understood his utterance until that point. J also gives feedback to K on line (1-d) for the utterance in line (1-c) for the same reason.

If K is replaced with a SDS, that system would need to produce similar behaviour as K where the feedback is produced as the utterance is ongoing. Traditional SDSs, such as one represented by Figure 2.2, don't work in this way. Traditional systems usually take as input full utterances from the ASR, requiring the ASR or some related component to determine when speaking has begun, and when a silence of specific duration has been detected (this is known as *end-pointing*). When such a silence is detected, the ASR hypothesis is then given to later modules; e.g., NLU. This results in a kind of strict turn-taking style of dialogue between a human and a SDS which can be compared to playing a game of *ping pong*.

Dialogue by its very nature is incremental in that participants in a dialogue take turns speaking (Schlangen and Skantze, 2011). However, in contrast to a traditional SDS, an incremental SDS doesn't wait until the end of an utterance to begin processing, making the increments of dialogue more fine-grained. In principle, an incremental SDS attempts to process as much as possible as early as possible, while attempting to not re-compute parts that have already been computed (more on this below). In reality, an incremental SDS processes the input utterance word-by-word, which has been shown to be a level of granularity in which humans interpret utterances (Brennan, 2000; Schlesewsky and Bornkessel, 2004).

Motivation On a practical level, dialogue systems that process incrementally produce behaviour that is perceived by human users to be more natural than systems that use the traditional turn-based approach (Aist et al., 2006; Skantze and Schlangen, 2009; Skantze and Hjalmarsson, 1991; Asri et al., 2014), offer a more human-like experience for the human users (Edlund et al., 2008) and are more satisfying to interact with than non-incremental systems (Aist et al., 2007). Psycholinguistic research has also shown that humans process (i.e., comprehend) utterances as they unfold and do not wait until the end of an utterance to begin the comprehension process (Tanenhaus and Spivey-Knowlton, 1995; Spivey et al., 2002). This has ramifications for resolving REs: as a RE unfolds, a component that resolves REs should attempt to resolve the referred object at each word increment prefix, updating the belief over candidate referred objects as additional words are added to the prefix.

Work has been done in incremental processing in many areas of dialogue systems: speech recognition (Baumann et al., 2009), speech synthesis (Buschmeier et al., 2012), and dialogue management (Okko et al., 2010; Selfridge and Arizmendi, 2012). Architectures for incremental dialogue systems have been proposed (Schlangen and Skantze, 2009, 2011) and incremental toolkits are also available (Baumann and Schlangen, 2012). More relevant to the work in this thesis is a recent attempt to identify the requirements for incremental semantics in dialogue

processing (Hough et al., 2015), as well as work in incrementally processing utterances to produce syntactic as well as semantic abstractions (Demberg and Keller, 2008; Purver et al., 2011; Peldszus et al., 2012; Peldszus and Schlangen, 2012; Beuck and Menzel, 2013). A review of work in incremental RR and related tasks is given in Chapter 3.

2.1.5 Incremental Computation

Comprehending an utterance (or, more specifically, a RE) incrementally is more than just applying a particular component on finer-grained prefixes, such as words. Following Schlangen and Skantze (2009, 2011) we distinguish between two kinds of incremental processing: *restart* incremental and *update* (i.e., fully) incremental. In a restart-incremental system, all internal state is thrown away between updates and output is always (re-) computed from scratch using the current input prefix–not just the newest increment of it. An update-incremental system keeps its internal state between incremental update steps, enriching the internal state at each incremental update with the delta between the current and the previous increment.

The difference between restart- and update-incremental approaches is illustrated in the following two examples (as an incremental ASR component might produce):

(2.3) the

(2.4) the red

(2.5) the red circle

(2.6) the red circle next

(2.7) the red circle next to

(2.8) the red circle next to the

(2.9) the red circle next to the green

(2.10) the red circle next to the green circle

In the above example, as input is received incrementally, a restart-incremental RR component would use the prefix at each increment and recompute what has already been computed. There is no maintenance of internal state. For example, a SDS component described in DeVault et al. (2009) is a NLU component that produces an entire semantic representation (in this case, an expected frame), even if it is only from partial input and no internal state between update steps is kept. Contrast that with the following:

(2.11) the

(2.12) red

(2.13) circle

(2.14) next

(2.15) to

(2.16) the

(2.17) green

(2.18) circle

The above example maintains an internal state and updates that internal state based on new information, without recomputing information that has already been computed.[2]

Of course, processing update-incremental SDS is not a trivial task. For example, the input given by the ASR might not be reliable as it processes incrementally, e.g., it produces output in the middle of a word, and would need to somehow undo the fact that it produced an early hypothesis and then produce the output that is more informed. There are other details that need attention when approaching incremental dialogue that works update-incrementally. In the following section, we look into a recently developed framework of incremental dialogue that addresses these concerns in a systematic model introduced in Schlangen and Skantze (2009, 2011), which plays an important role in deriving the model of RR in Chapter 5, and gives us some concepts and notations that is used in forthcoming chapters.

2.1.6 The IU Approach to Incremental Dialogue Processing

The basis of the model presented in Schlangen and Skantze (2009, 2011) is the *incremental unit* (IU) which is a minimal amount of 'characteristic input' that modules take in, update their internal state based on that input, and in turn produce their own IU output. (This model is often called the IU-model of dialogue processing, and we will henceforth refer to it as such.) This 'characteristic input' can be defined to be anything that is necessary to a particular module; such a definition implies that the granularity is also specified.

[2]The update-incremental approach has obvious benefits such as only needing to update the internal state based on the delta between an increment and the previous increment. However, in Khouzaimi et al. (2014) the authors show that non-incremental components can be made incremental, albeit restart-incremental, which is sometimes preferable over re-modelling and re-implementing a component from scratch to work update-incrementally.

For example, a typical, traditional SDS would define the characteristic input from an ASR module to a RR module to be an entire utterance. Thus, an IU that is outputted from ASR which would then be input to RR would be some kind of representation (e.g., transcription) of an entire utterance. An incremental SDS which typically works on the word level would be finer grained: an ASR module would produce IUs on the word-level; as words are recognised, they are passed to the RR module which would need to be able to process (e.g., update its belief state as to which object is being referred) at each word. This simple, yet important difference is illustrated in Figure 2.4.

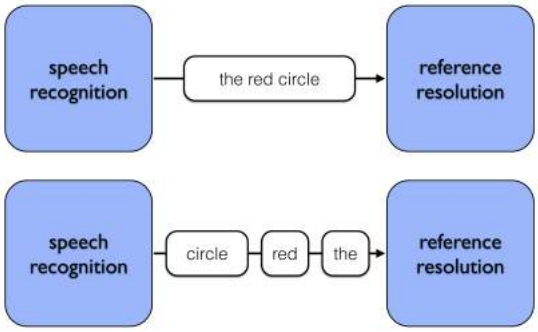


Figure 2.4: Example of the difference between (top) RE- and (bottom) word-level increments being sent from an ASR to a RR module.

How these IUs are defined is up to the system designer. For example, one system's ASR might produce an IU when it determines that a recognisable word has been uttered, whereas another system might produce an IU at specific time increments (i.e., by something other than linguistic units).

Additionally, the words in Figure 2.4 that make up the incremental input are part of a larger utterance and that utterance, in turn, is part of the larger dialogue. This whole-part relation is considered in Schlangen and Skantze (2009, 2011). That is, IUs that are created by the same module can be connected via a relation called *same-level links* (SLLs) which give a particular IU a direct link to its successor and a link to the IU that is its successor. For example, in Figure 2.4, the IU with the word *red* as its characteristic input is the successor of the IU for *the* and the IU for *circle* is the successor of IU for *red*.

Each module in the SDS take a specific type (or multiple types) of IU as input and produce, in turn, its own specific type of IU as output. Thus one module might receive very fine-grained input, but produce output at more spaced intervals. For example, an ASR module might produce IUs at each word, and the NLU module might update its semantic abstraction over the words it

has already received at each word. However, the DM, which receives IUs from the NLU, might not produce an action at each individual word. Rather, it might produce a back channel to signal ongoing comprehension (e.g., *m-hm*) after certain words, and produce an action when there is enough information to justify it (e.g., looking up an answer in a knowledge base, or extending a robot's arm to reach for an object that the RR module thinks being described).[3] This gives rise to another important relation: how do the IUs that a module outputs relate to the IUs which that module took in as input? In the IU-model framework, such a relation is the *grounded-in* (GRIN) relation.

A more complete illustration of SLL links and GRIN links is given in Figure 2.5. In this example, there are three modules: ASR, a part-of-speech (POS; i.e., produces a linguistic tag for each word) module which receives ASR output as input, and a module that determines when to produce feedback which receives POS output as input. For each word IU produced by the ASR module, there is an individual corresponding POS IU, but it took three of these POS IUs before the feedback module produced an IU; the feedback IU is GRIN to all of the POS IUs.

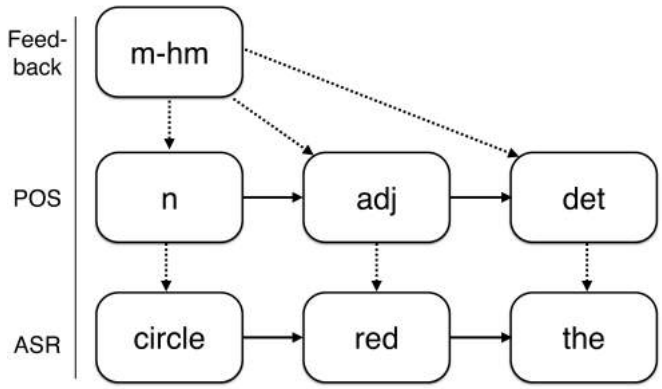


Figure 2.5: Same-level links (solid arrows) provide links between IUs of the same module and grounded-in links (dotted arrows) provide links to the IUs that justify that IU. Time increases to the left.

It is important to note that there are now two networks at play: a network of processing modules, and a network of IUs that were produced by those modules. These will be distinguished as the IU-modules and the IU-network, respectively.

One criticism against using incremental units that are finer grained than utterance level is that, often, waiting for more input means more informed hypotheses. This is certainly the case

[3]Though the decision to produce a back channel certainly is an action.

for ASR, where a word could be recognised, but at some point it is actually found to be a prefix of an ongoing word. This, however, isn't a strike against processing dialogue incrementally as long as there are provisions for handling these cases. The IU-model of Schlangen and Skantze (2009, 2011) defines provisions for such cases where a module can 'change its mind' in light of new information. Specifically, following Baumann and Schlangen (2012), there are three types of operations that a module can perform on an IU (as it pertains to the IU-network), namely `ADD`, `REVOKE`, and `COMMIT`. Each will be explained. Before we do, however, we need to establish how IUs are communicated between IU-modules.

Each module in the IU-modules has three parts: a *left-buffer*, a processor (i.e., an internal state) and a *right-buffer*. An IU-module takes in IUs on its left buffer. When IUs appear on its left buffer, it takes those IUs and processes them, updating its internal state, and produces (when necessary) corresponding output IUs which are put onto its right buffer. Two modules are connected by their buffers: a module's right (i.e., output) buffer can be connected to another's left (i.e., input) buffer. Thus when a module produces new IUs, those IUs are simultaneously put onto its right buffer, and any other module's left buffer that is connected to that module's right buffer. This set of IU-modules and the connections between them form the IU-module network, whereas the IU-network is a network of the IUs themselves. We now return to the explanation of the operations on IUs that modules can perform.

ADD Adding an IU is the operation that adds an IU onto the IU-network. When a module takes in new input on its left buffer and processes that input, it may produce its own characteristic output which is packaged into an IU. A module has the obligation to do several things when an IU is produced. First, the SLL and GRIN links must be established, giving the IU place in the IU-network. Second, the modules that are dependent on that module's output need to be informed that the IU has been added to the IU-network; i.e., the modules whose left buffers are linked to that module's right buffer need to be informed that an IU has been added. Those modules can in turn act upon that new input.

REVOKE As a module receives new input, it may determine that an IU that it had previously produced is no longer valid, and as a result should no longer be a part of the IU network (and, usually, would be replaced with another IU). When this occurs, the module has an obligation to inform the modules connected to its right buffer that a particular IU has been revoked. It could be the case that the IU has already been processed in later modules, so a module that receives a notification that an IU has been revoked needs to update its own additions to the IU-network that were based on that IU, which would result in the obligation for that module to inform the modules that are connected to its right buffer that an IU has been revoked, an so on. In general,

once something has been revoked, the new IU that "takes its place" can be added to the network as described above.

COMMIT When a module determines that an IU has been added to the network will not, by any circumstance, be removed from the network, it can inform modules that are connected to its right buffer of this decision. For example, when an ASR module determines that an interlocutor has stopped speaking, there will be no additional input, so the hypothesised transcription in the form of IUs that it has produced will not be revoked (i.e., there will be no new information to inform such a revoke). No additional operations are needed to augment the IU-network, but the IU that has been committed now has the state of being committed, and the other modules must be informed of this change in the IU-network. This might be useful information to later modules that have to make a decision; it would be informative to know that a decision can be made based on what is already in the IU-network–waiting for more information would not be beneficial.

Example The ADD and REVOKE operations are illustrated in Figure 2.6 for the RE *the red circle*. As the ASR module adds new words to the IU-network, the POS module is informed of each added IU which gives rise to that module's corresponding input POS IUs. During processing, the ASR adds the word IU for *sir*, but determines later that *sir* was actually the beginning of the word *circle*. The IU for *sir* is revoked, and the ASR module informs the POS module that a revocation of the IU for *sir* took place. The POS module then revokes its own IU for *n* (noun) which was produced (i.e., GRIN) by the word IU for *sir*. The IU for the word *circle* is then added to the network, which informs the POS module, which produces an IU that grounds into the IU for the word *circle*. The rest of the process continues with ADDs (and COMMITs if the ASR detects a certain amount of silence).

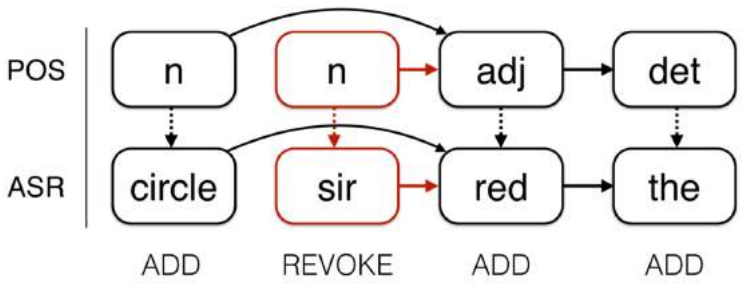


Figure 2.6: Two modules perform ADD operations until *sir* is replaced by *circle* using a REVOKE, then an ADD operation. The ASR module informs the POS module of the changes, and the POS module updates its corresponding IUs accordingly.

Summary of the IU-Model Following the explanation given in Kennington et al. (2014d), SDSs based on the IU-network approach consist of a network of processing *modules*. A typical module takes input from its *left buffer*, performs some kind of processing on that data, and places the processed result onto its *right buffer*. The data are packaged as the payload of *incremental units* (IUs) which are passed between modules. The IUs themselves are also interconnected via *same level links* (SLL) and *grounded-in links* (GRIN), the former allowing the linking of IUs as a growing sequence, the latter allowing that sequence to convey what IUs directly affect it. A complication particular to incremental processing is that modules can "change their mind" about what the best hypothesis is, in light of later information, thus IUs can be *added*, *revoked*, or *committed* to a network of IUs.

More information on the IU-model of incremental dialogue processing can be found in Schlangen and Skantze (2009, 2011). The authors also note that the traditional SDS is a special case of the incremental SDS where the granularity is simply at the utterance-level instead of the word level. The IU-model has also been adapted into a practical toolkit presented in Baumann and Schlangen (2012) (which was extended to make situated dialogue more streamlined in Kennington et al. (2014d)), allowing practical systems to be built using the IU-model as a framework ranging from in-car dialogue systems (Kousidis et al., 2014; Kennington et al., 2014b) to a dialogue system that a robot would use to play a game with human participants (Kennington et al., 2014a).

The models presented in Chapters 5 and 6 and the components that they are implemented as (see Appendix A and B) make use of this incremental processing model. They work in an update-incremental fashion, which requires them to be able to handle `ADD`, `REVOKE`, and `COMMIT` operations.

This section provided an explanation of SDSs and how a component of RR fits into that, with emphasis on situated, incremental variants of SDS. With this, we now isolate what the RR component is doing (i.e., what L did in the example at the beginning of this chapter) when it incrementally resolves REs.

2.2 Types of Referring Expressions

In the previous chapter, we showed that REs are a type of NP. In this section, we look at the types of NPs that make up the REs that we focus on in this thesis, and, to make matters clearer, some types that are not considered. We begin with the latter.

2.2.1 Some Types we don't Consider

To illustrate, consider the dialogue in (2)

(2) a. J: Did you hear that Sarah has a dog?
b. K: Yes, I was there when she bought it.
c. J: Ah, so the dog is real.
d. K: Yes, yes, her dog's name is Biff.

Proper Names

Biff in (2-d) is an example of a *proper name* usage for the dog under discussion between J and K. In this thesis we do not focus on proper names. The reader is referred to Chapter 5 of Abbott (2010) and the references there for further investigation into proper names.[4]

Sentential Propositions

A proposition is a statement or assertion that expresses a judgement or opinion, as in Example (3).

(3) this sentence is an example of a proposition

In the example dialogue in (2), the statement *the dog is real* made by J in (2-c) is also an example of a proposition, as it appears during the course of a "typical" dialogue. Generally, propositions require some kind of existence verb, i.e., a statement of *identity* (e.g., A *is* B), which equates the two parts of the proposition. In the example in (2-c), J equates *the dog* with being *real*.[5]

[4]This is not to say that naming a visually present object and using that name to later refer to that object is not a phenomenon that happens in a reference task such as we are interested in. This has been looked into by others (Chai et al., 2014) and focuses on establishing common ground between dialogue participants, as in Clark (1996), which is related to learning the kind of meaning we attempt here.

[5]The pattern of a proposition is A is B, something that Burge (2010) called an *indication*, which differs from reference in an important way. Of indication, he said (p.32):

> ...functioning to refer does not constitutively carry with it a function to engage in attribution or functional application. Since attribution is a constitutive representational function of the predicate 'is red' and the concept is red, they do not refer to anything. They indicate the property of being red. A primary representational function of predicates in language and predicative concepts in thought is attribution. So predicates and concepts indicate entities—bear relations to aspects of a subject matter. Their doing so is fundamentally in the service of attribution, attributing such aspects to further entities (often entities that are referred to). In occurrences in which no logical operations, such as negation, are involved, the predicate and the concept function to attribute what they indicate. For example, in That apple is red, is red functions to attribute what it indicates—the property redness, or the property of being red—to what That apple refers to. In attributing a property, they represent something as having that property.

In this thesis, the focus is on definite descriptions that don't necessarily have propositional value. Consider the difference between (4-a) and (4-b):

(4) a. the red circle
b. the circle is red

The example in (4-a) *presupposes* (that is, assumes) that a visually present object has the properties that are being uttered, namely that it is red and has a circular shape. It is assumed that the notion of 'red-ness' and 'circular-ness' have already been learned and will be understood by the listener. This is in contrast to (4-b) where there is a proposition that an object is red, as if the notion of 'red-ness' is either not yet learned, or there is some kind of question as to the circle's actual redness that needs addressing via a proposition. For more work about propositions, the reader is referred to the papers mentioned in Section 2.5 below.[6] Included in what we will not consider are propositional attitudes where propositions can be prefaced by things like *I think* or *I believe*, e.g., *I think that I want you to look at the red circle* or *John believes that there is a red circle*. The NPs that we are interested in can occur from within those kinds of sentences, but we aren't interested necessarily on those kinds of sentences as a whole.

Indexicals

The resolution of a referring NP that depends on immediate surroundings is, by definition, indexical (Barwise and Perry, 1981, p. 33). In other words, all of the types of NPs that we are interested in are indexical. However, there are some (more prominent) indexicals that we do not consider, which we strictly refer to here as indexical. Specifically, personal pronouns such as *I*, *we*, *you*, *he*, *she* that generally refer to people and the referents of those indexicals are highly dependent on the dialogue context (e.g., who is speaking, who is listening, etc.). Though there will always be a speaker and a listener, and they could be considered as "objects" which are visually present and hence referable, they nevertheless will not be of particular focus in this thesis. For discussion on these kinds of indexicals, see Kaplan (1989); Nunberg (1993). The reader would also benefit from Chapter 10 of Daniels (1990).

2.2.2 The Two Types we do Consider

As mentioned in Chapter 1, in this thesis we are primarily concerned with NPs that refer to an entity that is visually present. Such NPs that require contextual information for resolution are

[6]The RE *the red circle* is arguably a proposition, albeit a very weak one. For example, it is the same to say *the red circle exists*, which is a kind of proposition, but this is also an attributive use (as in Donnellan (1966)) rather than a referential one.

forms of **deixis** (also termed *indexical*, as noted above). The goal of their being uttered is to direct a listener's attention to the referred object.

Examples of these kinds of NPs are in (5). Each type will be introduced in turn.

(5) a. J: What about **the red one**?
b. J: (pointing) **That** one.
c. S: I think **it** is too small.[7]

Definite Descriptions

To put (5) into context, J is referring to an actual, single (more on this singleness constraint in the next section), visually present entity (i.e., an object). He initially uses the NP *the red one*, which is a type of definite description that contains words that pick out visual properties that the object has. Descriptions such as this don't name an object directly, rather aspects of an object. In order to resolve such a reference, the entire RE must be resolved and the lexical meaning of the words that make up definite descriptions must be somehow represented and composed.

Demonstratives

The utterance in (5-b) is an example of a *demonstrative*, deictic word, where some kind of non-linguistic cue is used to refer to the object (e.g., a pointing gesture, or by directing gaze to the object) and the utterance of the demonstrative word (such as *that* or *there*) indicates to the listener that a non-linguistic cue is also presented to aid in the resolution of the reference.[8]

[7]Because it is not a first mention, we are to a lesser degree interested in pronouns, e.g., the pronoun *it*, but we are nonetheless interested in that such a pronoun usage refers *exophorically* to the object. Though it does refer *anaphorically* (i.e., a reference to a previous linguistic entity) to the NP in (5-a), we are more concerned that it refers to a visually present object rather than a linguistic antecedent. In most cases, the kinds of pronouns we are interested in will indeed have a linguistic antecedent, like a definite description as explained above where, having been described sufficiently to direct the listener's attention to an object, subsequent discussion about that object indexes it by a pronoun. However, in all cases the object that is being referenced by a pronoun does physically exist and is visually present, so we will treat it as an exophoric pronoun. Pronouns like this do not draw a listener's initial attention to an object indicated by a speaker as definite descriptions and demonstratives do, and therefore function in a somewhat different way. However, a component of RR should be able to handle them under certain circumstances.

[8]These types of REs also fit into the *Giveness Hierarchy* set forth in Gundel et al. (1993), though how they are used in these differing ways will not be explored. Suffice it here to say that these different types indicate different cognitive statuses (or belief of status) of the speaker, namely, that definite descriptions are uniquely referential, demonstratives are familiar (e.g., both participants can see the object), and pronouns are already "in focus". We can assume that we are focusing on the most restrictive types (4-6), that the other types are subsumed. They found that,

2.3 Modelling Reference with First Order Logic

Section 2.1 set up the context where RR takes place and the previous section isolated the types of REs that will be our focus. In this section, we move closer to actually modelling RR by appealing to a more formal system that has been used in semantics: first-order logic (FOL). We explain how the formalism works, give examples of its application to the types of REs we are considering, and show some of the shortcomings in the kind of model/component that we wish to construct.

2.3.1 Syntactic Assumption

Though we are primarily concerned with the NPs that make up three specific types of REs, they often occur within larger utterances, e.g. (the target REs are shown in bold typeface):

(6) a. I wanted to ask if you can see **that red circle** and tell me what you think of **it**.

Here, and in subsequent chapters, we assume that a component exists that removes NPs that refer from their greater context (e.g., this could be done syntactically by finding NPs and then determining if they refer; see, for example, Friedrich and Pinkal (2015)). Given these NPs that refer to visually present objects, we can now take a closer look at how they could be resolved with FOL.[9]

2.3.2 Definite Descriptions

FOL consists of *terms*. There are several types of terms; the ones we use are *variables*, *predicate* symbols, and *formulas*. We begin with formalising definite descriptions. To illustrate, to say *the red circle* in a more formal way is to say that a circle exists and that circle is red. In other words, there is an entity that exists (more formally, a variable) and it is in the class of circles (a predicate) and it in the class of red things (another predicate). Let's focus first on the circle class:

(7) $\iota x.circle(x)$

The FOL representation in (7) is read as follows: that unique entity x such that x is a circle. In other words, that entity x is in the class of circles. This nicely frames the problem of

the more restrictive the type, the shorter the linguistic content (hence pronouns are generally most strict and have less content). This is not of particular focus of this thesis, but useful for drawing potential insight.

[9]Note that we also assume that the scope of the NP makes up a complete RE. We further assume that we are referring to single entities. More general quantification in the models presented in Chapters 5 and 6 is left for future work, but we provide some thoughts on the matter in the final chapter.

resolving references to resolving the object which can replace x which satisfies the formula; i.e., in which the formula evaluates to true. The variable x can range over all of the objects in a domain (defined shortly) and the ones that satisfy the formula evaluate to true. Clearly this assumes that such a mentioned object exists, and that assumption (indeed, constraint) is applied to the model here; in other words *existence* is presupposed. The ι (iota; read *unique x such that*) symbol expresses the contribution of the definite article which restricts the NP to denote a set with a single entity (Peano, 1897). This is a hard constraint that we are putting on the model; it resolves to one and only one entity, known as the constraint of *uniqueness*. Now that we have a formal representation (i.e., abstraction) over the utterance, it needs to somehow be "connected" to the world. To this end, consider the example in (8):

(8) a.

b. $\iota x.circle(x)$

If the context contains only the two objects in (8-a) whose suffixes represent their object identifiers, we can then range over the object identifiers of the context in (8-a). That is, following model theory (Tarski, 1956; Chruch, 1940), we have a model $M = \langle D, F \rangle$ where $D = \{a, b\}$ (the two object identifiers), and there is a characteristic function which maps the objects in D to the set of circles, $F(circle) = \{b\}$ and another which maps to the set of squares $F(square) = \{a\}$. What we want is to know the unique object identifier $d \in D$ such that the denotation (represented as $[\![]\!]$) of the formula $[\![circle(x)]\!]^{M,g[d/x]} = 1$, where g is a special function that assigns x to d for each possible value of d in the model M. We see that when x is replaced by b, the formula evaluates to true, as in (9) where the uniqueness assumption imposes the constraint that *no more than one* of the formulas evaluate to true; the existence assumption imposes the constraint that *at least one* of the formulas evaluate to true.

(9) a. $[\![circle(x)]\!]^{M,g[d/a]} = 0$

b. $[\![circle(x)]\!]^{M,g[d/b]} = 1$

This works fairly well for single-predicate NPs such as *the circle*. The slightly more complicated FOL formula for *the red circle* given in (10) makes use of the logical *and* operator $\wedge$:

(10) $\iota x(circle(x) \wedge red(x))$

(10) evaluates to true if there is a unique entity in both $circle$ and red that makes the formula evaluate to true, i.e.: $[\![\iota x(circle(x) \wedge red(x))]\!]^{M,g[d/x]} = 1$ iff $[\![circle(x)]\!]^{M,g[d/x]} = 1$ and $[\![red(x)]\!]^{M,g[d/x]} = 1$ and there is only one value that simultaneously satisfies both $circle$ and red. With this, we can derive a simple way of composing the words (which, for the most part,

map directly to predicates) of definite descriptions via the $\wedge$ operator in definition (11):[10]

(11) $[\![a_1 \wedge a_2 \wedge \cdots \wedge a_n]\!]^{M,g[d/x]} = 1$ iff $[\![a_1]\!]^{M,g[d/x]} = 1$ and $[\![a_2]\!]^{M,g[d/x]} = 1$ and ... and $[\![a_n]\!]^{M,g[d/x]} = 1$

This is an *intersective* approach to composition, which is explained in further detail below. We showed that this works for definite descriptions with just the head noun (e.g., *the circle*) and when an additional adjective modifies the noun (e.g., *the red circle*). But the definition in (11) allows us to construct formulas of arbitrary length. Some examples with corresponding FOL representations:

(12) a. *the small red circle*
b. $\iota x(small(x) \wedge red(x) \wedge circle(x))$
c. *the red circle on the left*
d. $\iota x(red(x) \wedge circle(x) \wedge on_left(x))$

This even covers cases where relational prepositions are used to denote a relation between objects, e.g.:

(13) a. *the circle above the square*
b. $\iota x[circle(x) \wedge \iota y(square(y) \wedge above(x, y))]$

Where that the $above$ predicate takes two arguments, but still maps those arguments into an individual pair (e.g., $F(above) = \{(a, b), (d, c)\}$ where a is above b, and d is above c).

2.3.3 Demonstratives (and Pronouns)

The framework so far works well enough for the kinds definite descriptions are considered throughout this thesis. We now look at formalising pronouns and demonstratives, which aren't completely straightforward. For example (here we are reminded that both of these types of NP assume existence and uniqueness as described above):

(14) a. *that*
b. $\iota x.that(x)$
c. *it*
d. $\iota x.it(x)$

[10] This assumes an identity relation between the application of x that the same value for x is applied to all of the individual functions.

The FOL representations for *that* and *it* in (14) seem intuitive, but the characteristic functions aren't as easy to explain. For it (or other like-pronouns), we can define the characteristic function to map from entities in the domain D to the set of entities that were most recently referred.[11] For *that*, the characteristic function maps from D to the set of entities that are currently being pointed at. Perhaps somewhat more intuitive to these meanings are the following FOL representations, (15-a) for *it* and (15-b) for *that*:

(15) a. $\iota x.recently_referred(x)$
b. $\iota x.pointed_at(x)$

2.3.4 Functional Application of Objects

We use an additional operator, the λ-operator, to abstract over variables which are useful when we wish to apply the variables (in this case, objects) in a domain D to a FOL representation (applying objects to formulae in this way is a more direct method of applying a world representation of a formula than using model theory).

$$[\![w]\!]_{obj} = \lambda x.\phi_w(x) \tag{2.19}$$

Where a symbolic concept ϕ_w of a word w (e.g., red or $circle$) have corresponding predicate concepts in FOL and a representation of an object x in D can be applied to that concept via the $lambda$ operator.

With these representations, we can semantically represent the kinds of NPs we are interested in in a simple FOL form, assuming that the domain D and the functions F are fully specified.[12] That is, we assume, if the concepts ϕ_w can be defined and their functional application learned, then the task of fitting application of objects into FOL is more or less complete. The task for this thesis, then, is to couple objects x with language concepts ϕ_w. This is a form of *lexical* semantics, where the "meanings" are considered. More on this below.

[11] This is of course a gross simplification as to what pronouns actually do, including their pragmatic constraints, but we aren't quite concerned with that here; a pronoun will always refer to a visually present object that is somehow salient. We will see examples later when this simple characteristic function doesn't work.

[12] Functional lambda operations were a big part of Montague's (explanation below) semantic (and syntactic) framework (Hobbs, 1983; Cotelli et al., 2007). Here, the only thing applied functionally with lambda are the objects to single predicates. Other operations, such as those connecting those single predicates with other predicates, are defined above in FOL.

2.3.5 Limitations of Intersection

The FOL approach explained above assumes an *intersective* mode of interpretation. That is, as we have defined the *and* $\wedge$ operator, the method of composing the results of the application of two different characteristic functions (e.g., *red* and *circle*) is to find the unique object that exists in both sets (i.e., the intersection; e.g., the object that is both in the set–the class–of red things and in the set of circles). This has its limitations, because some sets are determined relative to each other. Consider the following examples:

(16) a. the small elephant
b. $\iota x.small(x) \wedge elephant(x)$
c. the big mouse
d. $\iota x.big(x) \wedge mouse(x)$

Clearly, these are not composed in the same way as described above, because even a small elephant is larger than a big mouse; i.e., the class of small things doesn't necessarily include the elephant as described; rather, an elephant can be considered small by comparing it to other elephants–the same is true for big mice. This is a limitation of this approach, but we show in Chapters 5 and 6 (and particularly using the data presented in Chapter 4), the assumption that classes are composed in this intersective way works well in practice. We leave more involved methods of composition that handle these kinds of phenomena for future work.

Some Key Shortcomings of First-Order Logic

Even though we are interested in fairly limited phenomena that can be captured with FOL, the work of resolving REs is not yet done. Theoretically, we have a well-established system that we can use to compose a RE in order to determine which object has been referred. However, there are some issues with this approach, particularly when applied to a practical component.

First, is the practical representation of the domain D and the set of functions F in the model M. In order for FOL to work, both must be fully specified. How this is to be done in a practical component constitutes one of the main parts of the task. For D, each object needs to be identified and represented.[13] Perhaps a greater difficulty is in F, that is, defining the characteristic functions that map the objects into specific classes (e.g., red, $circle$, or on_left). In traditional work in RR, there is a simple text-match mapping between a word, e.g., *red* and a

[13] In a system that uses a virtual scene, this is usually quite straight-forward, as objects are already symbolically represented and can be given object identifiers. For a scene that has real-world objects, some kind of vision processing needs to be done in order to segment and represent the individual objects. We look at various approaches, including the approaches presented in this thesis in later chapters.

corresponding predicate red, and the objects in $F(red)$ are defined beforehand. In short, there needs to 1) be a way of determining the set of F, which we also term the *classes* that objects can belong to (e.g., the class of *red* things), and 2) how to determine whether or not an object fits into those classes, i.e., the actual application of each function in the set of F.

Another shortcoming is the assumption (in fact, the requirement) that there is a single referent, *without uncertainty* (i.e., existence and uniqueness). While these are assumptions that we rigidly make here, this poses a problem in practical components which need to be able to handle uncertainty in both the recognition of the RE, and the representation of the world that makes up D and F. For example, someone might describe something as vaguely red, when it doesn't fit the prototypical red that would put that object into the class of red things. Something that isn't prototypically red, but still redder than the other visually present objects should be more likely to be the referred object.

A third shortcoming is the way the REs are composed in a FOL framework. Typically, an entire RE must be specified in FOL before it can be computed and must be "unpacked" from the bottom up, but we noted earlier that humans process the resolution of REs incrementally, i.e., they don't wait until the end of an utterance before processing, rather they process as much as possible as early as possible (which is how we want to model the resolution of REs in this thesis). Example (13) above is a good example of this: standard FOL would need to completely resolve y before it can resolve x. However, in an incremental system each word in the RE should contribute to the resolution to the referred object as the RE unfolds without needing to wait for the resolution of one of the variables (i.e., both are resolved simultaneously, given the unfolding RE).

To recap, the shortcomings that need to be addressed for a practical system are enumerated as follows:

1. the set of classes must be determined
2. the functions that assign objects to those classes must be determined
3. incremental composition

These shortcomings constituted the limitations of FOL, but they also constitute the shortcomings of previous approaches to RR (discussed in greater detail in the next chapter). Overcoming these shortcomings is one of the main contributions of this thesis. How two these shortcomings might be overcome is explained throughout subsequent sections of this chapter.

2.3.6 A Brief Survey of Other Semantic Approaches

FOL has been used in logic and semantic theories for a respectable amount of time, and it is clear that FOL doesn't capture all of the phenomena in language use. However, for the purposes of this thesis it is sufficient. Other approaches to representing language semantically exist; indeed ones that are designed for use in dialogue, so why not use those instead of FOL? In this section, we address this important question by briefly describing other approaches to semantics and provide reasoning for using FOL.

Montague Semantics

Richard Montague developed a system of symbolic logic that was introduced in Hobbs (1983); Cotelli et al. (2007), promoted in Partee (1975). Among other things, Montague attempted to solve intensional constructions (see below) and quantificational NPs, however, we don't really have need of generalised quantifiers when dealing with NPs that refer to a single object. That, of course, is *not* to say that they aren't important. We leave other types of quantification to future work. Indeed, many have followed Montague and looked into generalised quantification, see, for example Barwise and Cooper (2008); Hintikka (1973). He also formalised a way to handle the syntax of natural language using categorical types which we could use, but instead we will assume some simple syntactic structure that plays directly into the FOL representations and work from there.

Discourse Representation Theory

(DRT) is a theoretical framework for discourse phenomena such as anaphora and tense (Kamp, 1993). It goes beyond the sentence level, parting ways with formal semantics (e.g., FOL) but does continue to use model-theoretic tools to represent a discourse. Such a framework could have use here, as we are interested in a dialogue setting. However, we aren't interested in the interactive, multi-sentence, discourse aspects of dialogue per se; we are interested in individual REs, the words that make up those REs, how they refer and what they mean. Certainly, DRT can handle phenomena which are difficult to represent in FOL, such as anaphora (in particular, donkey sentences), but they are not needed in this thesis.

Situation Semantics

Situation semantics is a framework that represents the situation of a given speech event. Situations consist of 'individuals having properties and standing in relations at various spatio-temporal locations' (Barwise and Perry, 1981). Those situations can be real or abstract; the

former are real situations (like the ones we are interested in) whereas the latter are akin to set-theoretic objects that are constructed from individuals. Closer to what we are interested in, Kratzer (2011), focuses on situations, which has been extended in work by Paul Elbourne (2001, 2008); Daniels (1990).

Of particular interest are his lexical entries for *the*, *it* (Daniels, 1990), and *that* (Elbourne, 2008).[14]

(17) a. $[\![the]\!] = \lambda f.\lambda s : s \in D_s \;\&\; x.f(x)(s) = 1.\iota x f(x)(s) = 1$
b. $[\![it]\!] = \lambda f.\lambda s : s \in D_s \;\&\; x.f(x)(s) = 1.\iota x f(x)(s) = 1$
c. $[\![that]\!] = \lambda x.\lambda f.\lambda g.\lambda s.\iota z(f(x)(\lambda s'.z)(s) = 1 \;\&\; g(\lambda s'.z)(s) = 1$

Note that the entries for *the* and *it* are identical. In other words, in terms of situation semantics, there is no difference between a definite phrase and a pronoun. That is, there is a situation s in a domain D_s, where x is an entity that exists and there is only one such entity (hence ι). These take a NP (minus the determiner) as an argument (e.g., *red circle*). The entry for *that* is trickier, but basically shows the same thing: that there is an entity that exists, that there is only one such entity, and that entity has a gesture (similar to our *pointed_at* predicate in Example (15-b)). The entry for *that* presented in this framework does account for the types of usage in which we are interested, such as using *that* with a pointing gesture, or *that* followed by a definite description (with optional pointing gesture). Models of reference resolution using situation semantics have also been proposed (Poesio, 2011). Situation semantics have been shown to handle phenomena in language that FOL and traditional Montague Semantics cannot on its own (e.g., donkey sentences, see Kratzer (2011)).

Type Theory with Records

Type theory with records (TTR; (Cooper, 2005, 2012)) in a way takes up the intuitions of situation semantics with a different (and arguably less problematic) formalism. TTR is an integration of Montague semantics, DRT, as well as frame semantics (Fillmore, 2006). TTR is well-suited for dialogue. It was the principle framework used in Ginzburg (2012), which was concerned with the issue of how to describe certain linguistic features of interactive conversation. TTR works well because it represents aspects of semantics as we've discussed them, but also utterance types which are aspects of language that are more pragmatic, such as speech acts (Searle, 1976); e.g., clarification requests in dialogue, and other dialogue moves. While these are all important aspects of a fully-functional dialogue system, and we are certainly interested in resolving references within a dialogue framework, the phenomena of dialogue are beyond the

[14] Some aspects of these entries are left out for simplicity.

scope of this thesis.[15]

Discussion

While these other formalisms and frameworks yield richer semantic representations (as well as, in some cases, pragmatic), all build upon FOL in various ways. It is therefore assumed that, if the models presented in Chapters 5 and 6 fit into a FOL framework, then they can be used in extended frameworks such as DRT, situation semantics, and TTR. Importantly, none of these approaches directly overcome the shortcomings of FOL listed above. To begin overcoming those limitations, we now turn to grounded semantics.

2.4 Reference and Grounding

In this section we take a look at the background on *grounded semantics* which provides a foundation upon which we can build as we overcome the some of the above-mentioned shortcomings of FOL. We begin by asking a question that has been asked before (e.g., Roy and Reiter (2005); Larsson (2015)), namely how does language relate to the non-linguistic world and how does linguistic meaning relate to *perception*? How do we as humans learn words and agree on their meaning such that we can use those words to convey and understand our intentions with each other? The formal approaches in the previous section aren't quite able to answer these questions directly (Steels and Kaplan, 1999), and we will see that the answer to these questions help address the shortcomings of FOL outlined above.[16]

In this section, we first look at some approaches to meaning from the field of cognitive science. We then look at how the ideas reviewed in this section fit into reference, particularly in definite descriptions, and demonstratives. We then look at how the ideas in this section can

[15] It should be noted here that attempts have been made at using TTR as a formal basis for lexical semantics (Larsson, 2015). We discuss this further below.

[16] As an aside, the claim that manipulation of symbols, be they computational or logical symbols as presented in the previous section, does not equate to intelligence (in which meaning plays an important role) has a long and interesting debate. One well known contribution was made by John Searle (1980) in his famous "Chinese Room" *Gedankenexperiment*. Searle attempted to make the case that a program cannot give a computer a mind, understanding, or consciousness in the same way that humans do regardless of its behaviour, even if that behaviour seems to be intelligent. In artificial intelligence (AI) research, there is a view of *strong* AI (in which a correctly functioning computer program that behaves human-like is, in fact, intelligent and has a mind in the same way humans do) and *weak* AI (where computer programs simply simulate human behaviour, but don't have human-like minds). In this thesis, an attempt is being made to address some issues that are directly related to AI such as capturing the meaning of words, but we are not making any claim that the models presented here are in fact models of how things are done in a human's mind, nor are they by themselves intelligent. The goal, rather, is to build a component that would be used in a system that a human would interact with in a natural way.

be reconciled with semantics, and extend our formal approach to work with these ideas. We then revisit some discussion from the previous section in light of the ideas explored here.

2.4.1 The Symbol Grounding Problem

Roy and Reiter (2005) make the following claim:

> There is sometimes a tendency in the academic world to study language in isolation, as a formal system with rules for well-constructed sentences; or to focus on how language relates to formal notations such as symbolic logic. But language did not evolve as an isolated system or as a way of communicating symbolic logic; it presumably evolved as a mechanism for exchanging information about the world, ultimately providing the medium for cultural transmission across generations.

Indeed, in the previous section full attention was given to how REs can be represented in a (albeit simplified) formal logic. However, language isn't used in isolation–certainly language isn't isolated when being used to refer to visually present objects. When dealing with language and meaning, we must look at how everyday language is used by humans. The human brain is physically embodied and can interact with its environment through percepts such as sight, sound, touch, etc. Humans also do not exist alone, but are surrounded by objects (as noted in the Chapter 1) with which we can interact, and, importantly, humans also come into contact with other humans. If human A wishes to draw human B's attention to an object that is external to both of them (for example, a piece of fruit, or as in the example at the beginning of this chapter), how can such an intention be performed? Regardless of how language came to be what it is now, the fact of the matter is that humans do use spoken language to communicate with each other in such a circumstance as drawing attention to a piece of fruit. Moreover, the choice of words is important: the individual words in the RE produced by human A must "pick out" properties which human A perceives that particular piece of fruit to have (e.g., its colour, shape, or size) knowing that human B would understand those particular words to pick out the properties of the intended referent. That is, both A and B have a representation of words that pick out visual properties and those words are agreed upon by both. Without such an agreement, communication via spoken language could not occur at all.

With this, we can define what is meant by **grounding**: the agreed-upon meaning of a word is based on perceptual experiences with which that word associates. That is, the meaning of (many) words is *grounded* in perceptual experience. For example, one cannot really know the meaning of the word *red* if one has not seen a red object and experienced that colour word being associated with a visual stimulation in the form of redness.

This seems to be the form of meaning we are after when dealing with reference to visually present objects, but can such a grounded meaning be somehow represented in a symbolic system such as FOL? Harnad (1990) notes that indeed symbolic approaches, such as our formal approach in the previous section, are autonomous functional modules that simply need to be "hooked up" to the world to work. Beginning with a symbolic system such as we have discussed above and hooking it up to the world amounts to representing the world in some way, i.e., mapping from events and entities in the world to a representation that can then be somehow useful to the symbolic system. In our above example, repeated here

(18) $\iota x(circle(x) \wedge red(x))$

resolving the RE which gave rise to the FOL in (18) amounts to finding the entity in the domain D that makes the statement true. But if we are referring to visually present objects, how can we represent those objects such that they can be in D (as explained above)?

This is the problem that we eluded to above: How is it decided that an object fits into the class of red things, or that an object has the property of being red? A rule (i.e., a pre-defined function) could be made where, for example, if an object is perceived as having a value in the RGB scale that falls in the range of things that are generally called red, then that object has the property of being red. This seemingly top-down approach to linking symbols to percepts by defining functions has several problems. First, there are potentially exceptional cases where a non-red object is described as being red, albeit not in the prototypical range of red as defined by the function. This is a problem in robustness. The second problem is that the meaning of a word is encoded in the function–the function is not learned by experience, which is how humans seem to learn meanings of words.[17]

Grounded semantics addresses this problem by learning the functions "by experience" from data. For our purposes, given *training* examples of referred objects and the REs that were used to refer to those objects, a model of RR would need to learn a mapping between object

[17]These problems have been looked into by *connectionist* approaches to cognition (again, for our purposes, we are only interested in cognition as it pertains to word meanings), where neural networks have been introduced as mimicking, to some degree, how the brain works. (See earlier work in connnectionist models (Hinton et al., 1986) and comparisons between symbolic and neural learning approaches (Shavlik et al., 1991). See also Pinker and Prince (1988); Fodor and Pylyshyn (1988) for criticisms to early connectionist approaches.) In such a network, a multilayered network of nodes encodes patterns of behaviour. Nodes on one part of the network (e.g., the bottom) would link, for example, to percepts, while higher-level nodes represent more abstract notions. For example, a low-level node would be able to read a colour value and a higher-level node or subnetwork of nodes would be able to interpret that as red, where red is a linguistic concept. Early approaches to this have worked for toy examples (see Harnad (1990); Hinton et al. (1986)). More recently, neural networks have used more principled approaches on how the nodes function, how the networks are constructed, and how they learn from data.

properties (or features) and aspects of language. It is through this grounded learning that such a model can learn the "meaning" of visual words, such as colour and shape terms. This is presently explained for definite descriptions and demonstratives in greater detail.

Grounding and Definite Descriptions

Definite descriptions that refer to visually present objects presumably are made up of words that help a listener distinguish an object from other objects by expressing properties that the referred object has. Thus words denoting colour, shape, size, spatial placement, etc., are used. The meaning of these words are not abstract (compared to, for example, the word *love*) where the meaning, it seems, is directly related to how those concepts are perceived visually. For example, if one attempts to explain the meaning of the word *red* to an individual who has never visually experienced red, it would be very difficult without pointing to objects that are red.[18] Rather, a word with a grounded meaning representation is based on perceptual information.

In the RE *the red circle*, the meaning of the word *red* is more of an ability to determine the redness of an object, given visual features of that object. In other words, given the features of an object, a meaning representation of *red* would be able to tell if that object is red or not. Likewise, the meaning of the word *circle* is the ability to determine something's roundness; the entire RE gives both words a vote as to how well they fit a particular object–that it is red enough and round enough to be distinguished from other objects.

This can be learned from data: given enough REs containing the word *red* and the objects to which those REs referred, a range of features values can be observed as being described as red with some potential uncertainty. The model of RR would need to somehow select the features and the range of values that denote what is referred to as being red.

Grounding and Demonstratives, Pronouns

Are words such as *it* and *that* grounded as descriptive words are? Pronouns refer to linguistic antecedents, but the exophoric kind we are interested in can also be grounded in a similar way as descriptive words. Recall our formal representation for pronouns and demonstratives, repeated in (19):

(19) a. $\iota x.recently_referred(x)$
b. $\iota x.pointed_at(x)$

[18] Red is, of course, a primary colour and cannot be derived from other colours. However, even though a colour that can be described in terms of others, such as purple which falls directly between red and blue, it still isn't quite sufficient to explain the meaning of purple as being a mix of red and blue.

Here grounding would be the same as it was with descriptive words: learn a function, either $recently_referred$ or $pointed_at$, and map between those concepts to the respective objects that fit into those classes. That is, learn a function that picks out the objects that were either recently referred or are currently being pointed at, respectively. Such a function for pronouns is fairly straight-forward: an object that is referred receives some kind of property that it was the last one that is referred, though pronouns aren't necessarily grounded in visual information, rather in contextual discourse information (though the referent is visually present). In contrast, grounding demonstratives is based on visual features, namely, knowing to which object the pointing gesture is indicating.

2.4.2 Grounding, Semantics, and Probabilities

We now look at how words grounded in perception fit into our formal framework. Following Equation 2.19 above, instead of representing an object x an abstract identifier, we can represent the object directly as some kind of feature vector, where the features represent visual aspects of that object. Equation 2.20 shows a slight modification to Equation 2.19, where x is now the vector $\mathbf{x}$:

$$[\![w]\!]_{obj} = \lambda \mathbf{x}.\phi_w(\mathbf{x}) \tag{2.20}$$

That is, we can represent a predicate ϕ_w using some kind of function that *learns* (i.e., not pre-defined) a mapping between object features and a decision that the features of the object are a good fit to what the concept represents. For example, a grounded function for red, i.e., $\lambda \mathbf{x}.red(\mathbf{x})$, would return 1 if the features x that represent some object are deemed by that function to sufficiently represent the concept of redness. In fact, instead of returning 1 if the object is deemed red enough and 0 if it is not, the grounded function for red can return a score, e.g., a probability between 0 and 1, where the closer to 1 the score is, the more red it is deemed to be. This is a type of probabilistic/stochastic approach to learning the grounded functions of words. Such a learning can happen by using example data of interactions, e.g., observing referring expressions and the features of the objects to which they refer, where the grounding function learns what features distinguish objects from being a good fit to that word and those features that do not.

The difference between a symbolic approach to meaning and a grounded approach to meaning is illustrated in Figure 2.7. For the symbolic approach, the world is represented as a set of classes (i.e., properties, both work in this example) where a particular object is denoted via a rule as being a member of a particular class by the judgement of a human designer. Often, this

is done by a direct mapping between the word and the name of the class (e.g., some kind of orthographic distance function). Meaning in the grounded approach requires that a function between a word and a class be learned through data; this is illustrated in that there is a line between every word and every class, where the thickness of the line represents the degree as to how much that particular word belongs to that particular class. Furthermore, the class names are arbitrary as long as they are consistent.

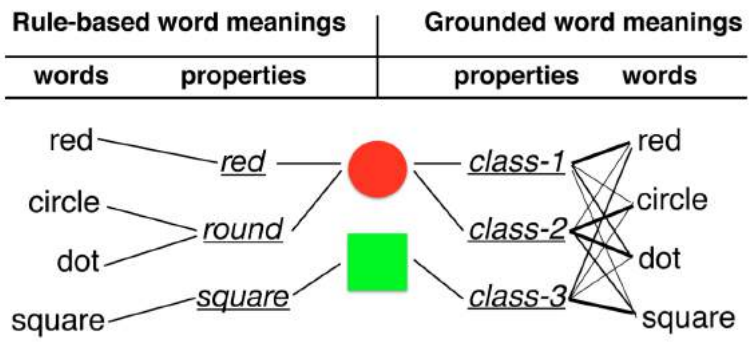


Figure 2.7: Symbolic, rule-based word meaning vs. a grounded word meaning. The rule-based approach requires words and class names to be defined, the grounded approach learns these mappings from data.

The models presented in Chapters 5 and 6 are grounded models. In the former model, a mapping between words and pre-defined properties is learned. In the latter model, the need for properties is eliminated and words are mapped to low-level object features. Both models are probabilistic and learn from example data.

Overcoming the Shortcomings

Recall the three shortcomings with FOL explained above

1. the set of classes must be determined
2. the functions that assign objects to those classes must be determined
3. incremental composition

When coupled with grounded semantics, these shortcomings aren't quite as constrained as they were before. We can represent objects in various ways, for example using low-level features or a set of defined properties. Using stochastic methods, we can use machine learning techniques to learn the functions that map words and concepts in FOL. This is explained in greater detail in later chapters. A stochastic approach also allows us to relax the uniqueness

constraint in that a probability distribution over all objects is produced instead of a singleton set as a result of a computed truth value. The final constraint, that of bottom-up composition, has not been addressed through our discussion of grounding, but is of high importance, as we saw in Section 2.1. Incremental considerations are taken into account within the framework of each model and explained in their respective chapters.

The models presented in Chapters 5 and 6 link in their own way to FOL and perform a kind of grounding. Each also resolves REs in an update-incremental fashion, addressing the issue of bottom-up composition. The models are explained and evaluated in their respective chapters.

2.5 Reference and Meaning

Grounded semantics and FOL have both been used as approaches to approximating meanings of words, expressions, and sentences. The purpose of this section is to provide an overview of some philosophical ideas of meaning as it pertains to reference. Though not the direct focus of this thesis, the meanings of words and expressions are necessarily at issue when grounded semantics meets a semantic formalism like FOL, as we are attempting to do here. This section provides us with some additional terminology, in particular that of *intension* and *extension* which we use throughout this thesis.[19]

2.5.1 Philosophical Background

To begin, consider the following:

(20) a.

b. the red circle

At first glance, it may seem that the red circle represented in (20-a) is in fact the *meaning* of the RE found in (20-b). This could be generalised into a simple definition of meaning; that the meaning of a RE is, in fact, the object to which it refers. This, of course, doesn't really capture meaning at all, neither for the RE, nor for the words that make up the RE. The difference between the referent of a RE and the meaning of a RE can be distinguished more clearly in an example given in Chapter 1, repeated here:

[19] A more in-depth analysis of the philosophy of reference and meaning has been handled elsewhere; particularly as it pertains to proper names and indexicals, which we are not concerned with in this thesis. The reader is referred to Kellerwessel (1995) and Abbott (2010) for a more in-depth analysis of theories of reference (at the writing of this thesis the former, originally written in German, does not have an English translation). A more psychologically-motivated, yet philosophical stance on meaning and reference can be found in Burge (2010).

(21) The morning star is the evening star.

Where *the morning star* and *the evening star* are two different REs (albeit not the attention-directing type we are interested in) with distinct meanings, but they both in fact refer to the same object, namely the planet Venus. This example was explained by Gottlob Frege (1892) ("Über Sinn und Bedeutung" / "On sense and reference"), where, among other things, he explained the difference between *sense*, what we would call the meaning or notion of a word or expression, and the *reference*–that is, the referred entity itself. The entity itself isn't the meaning, rather it instantiates something to which the sense can refer. This distinction is made clear by another example using a triangle, with three additional lines, each connecting the midpoint of each line on the triangle with the opposite corner. These three lines inside the triangle intersect at a single point. The intersection of any two of those lines is the reference, but the two lines themselves that intersect at that point are the sense. Thus there are two senses, but one reference.

Before Frege, Mill (1846) made a similar distinction between *connotation*, the properties or attributes that are implied by a word or expression, and *denotation*, what an expression applies to in the world (i.e., the referred entity). This distinction is illustrated in Example (2) in Section 2.2, where in (2-a) no particular dog is being referred; the usage of the word *dog* connotes a type of entity that has properties belonging to dogs, and (2-b) where K is denoting a particular dog (which has all of the properties that the word *dog* connotes).

Russell (1905) asserted that properties (or, in his words, universals) are meanings of words (which, according to Abbott (2010) is getting at a Fregean sense). Definite descriptions, such as REs, aren't constituents, but rather quantificational NPs. Some definite descriptions are denoting descriptions, though they don't necessarily denote actual things, for example:

(22) The king of France is bald.

has a meaning, but does not denote an entity in the actual world because, at the writing of Russell's paper, there was no king of France–yet the definite description is clearly understood. Russell has a somewhat different view, then, on meaning when compared to Frege and Mill.[20]

A more pragmatic view was taken by Strawson (1950) claiming that there is a difference between the usage of a definite description and the definite description itself. Put another way, definite descriptions don't refer to things, *people* refer to things when they use definite descriptions. He further asserted that, when a definite description is used (preceded by the word

[20]Masheb et al. (2011) was quite critical of Russell's theory, distinguishing between denotation (i.e., the class of stuff denoted by a connotation) and reference, which is an individual or set instantiation of an entity of a particular class (e.g., *red* denotes red things, while reference refers specifically to an object that is red; i.e., in the class of red things).

the) the listener presupposes that the object exists, and that there is only one entity that is being referred (as explained above). Kripke (1977) also distinguished between speaker reference and semantic reference in that though the two are related, they perform different tasks.

Another important distinction can be made between REs that actually refer to real entities, and REs that are *attributive*. From Donnellan (1966):

(23) Smith's murderer is insane.

The example in (23) can be used attributively if it follows an utterance such as *anyone who would kill Smith must be crazy*, without particular care as to the individual who did the murdering, the point of the RE was to assign attributes to Smith's murderer–whoever it may be. Contrast that to *which one killed Smith?* to which comes the reply in (23), which helps the person who asked the question pick out the person who is the murderer.

Discussion These distinctions between sense/connotation, reference/denotation; speaker vs. semantic reference; attributive vs. attentional reference are things to consider when modelling a dialogue system component that resolves references to visually present objects, particularly using the above FOL as a modelling framework. Abbott (2010) notes that the semantics of attributive and referential REs are identical, at least in terms of a semantic formalism like FOL. In a dialogue setting, the speaker that utters a RE is certainly making a speaker reference to an object, but the listener must take the words of the RE, and (via semantic reference) resolve the object which the speaker referred.

Here, we follow Strawson (1950) that the speaker performs a RE in order to (also following Quine (1980)) draw the listener's attention to a particular instantiation of an object that fits the description of the RE. These are pragmatic considerations when references are being made. However, here we make the explicit assumptions that these are in fact taking place–all other types of reference (attributive, or non-referring) are not considered here.

Another focus of philosophers of meaning and is propositional truth values. For example, that *the King of France* has no referent and therefore the truth value must be false. But if one were to say, *John believes that the King of France is bald* must necessarily be false because *the King of France* has no truth value, but the fact that John believes it could very well be true (see Footnote 22). In this thesis, focus is not put on truth values of propositions. It is assumed that a RE does in fact refer and is therefore true. In terms of FOL, it is assumed that there is an object that satisfies the formula, and that there is only one (i.e., existence and uniqueness). Holding these things constant, we return again to sense/connotation and reference/denotation distinction.

2.5.2 Intension and Extension

As we are interested in resolving REs to visually present objects, it seems intuitive that we simply focus on the latter and forget the former. This would work if every object had a unique name that referred particularly, without ambiguity, to that object (and the philosophers mentioned in this section had a lot to say about names as they pertain to meaning and reference). However, when dealing with definite descriptions, such as *the red circle*, the referent is determined by the constituents of the RE, namely the words and their composition (another term from Frege where the meaning of sentences–or utterances–are composed by the meanings of their constituent parts).

Both Frege and Mill make a distinction between the meaning of a RE and the *designation*, or the thing to which a RE can be applied. Another, more technical term that is used to represent the notion of meaning (i.e., sense/connotation) is *intension* and the term used for that which is designated (i.e., reference/denotation) is *extension*. Indeed, the FOL framework that we opted for above is an intensional system of semantics in that it attempts to represent meaning by intension rather than extension.

Rudolph Carnap (1988) explained intension and extension in the following way in order to explain his semantic system using *properties* and *classes* (the connections between this and our above FOL explanation will become clear).[21] A property is something an entity has, whereas a class is something to which an entity belongs. The extension of something is akin to the class to which that something belongs; the intension of something is the corresponding property which that something has. Carnap gives the following examples (p. 18):

(24) a. The class Human is the same as the class Featherless Biped.
b. The property Human is not the same as the property Featherless Biped.
c. The property Human is the same as the property Rational Animal.

That is, Humans and Featherless Bipeds are extensionally equivalent. But they are not intensionally equivalent (i.e., they don't have the same Fregean sense). Carnap then continued to provide a semantics in which model-theoretic entities are identified with intension. This semantic system has been influential on later work in formal semantics, forming a group of *intensional logic* like our chosen FOL above.[22]

[21] He admits and later shows that the terms 'properties' and 'classes' aren't completely necessary in order to describe his semantic system, but he spans several pages using those terms as scaffolding to help the reader understand what he means by intension and extension, then discards them (as it were).

[22] Another technical explanation of intension uses the notion of *possible worlds*. Abbott (2010) explains this idea in the following way (p. 53-54):

> The value of incorporating possible worlds into one's semantics is that we can recognise a reference

For the formula in 2.20, ϕ_w is the intension, whereas the functional application of $\mathbf{x}$ to that intension forms the extension–i.e., the degree of belief that $\mathbf{x}$ fits ϕ_w. The question now is how to learn the intension, but the ideas presented thus far don't give us any idea of how to do so. The remainder of this section addresses this.

2.5.3 Intension via Extension

Kathleen Dahlgren (1976) makes the following claim in criticism of semantic theory (i.e., intensional logic):

> Semantic theory for natural language is faced with the following problem. It is relatively straightforward to formally state the semantic properties of whole sentences (such as ambiguity), and the relations between words and sentences (such as "not S is the negation of S"). It is even possible to give formal accounts which seem to be somewhat accurate, for how the meaningful parts, that is, the morphemes, phrases, and constituents of sentences combine to produce meaningful generative syntax or logic, some semantic properties of human language can be described. But no explanatory account of the semantics of individual words has been achieved using such methods. The theory of the lexicon has proven a difficult

---

or extension for expressions not only in the actual world but also in other possible worlds. The possible worlds formulation of the notion of the INTENSION of an expression brings these extensions together, and give us something like a Fregean sense. In the most straightforward system, intensions are uniformly functions from possible worlds to extensions.

The idea behind possible worlds is that there is a set of (an infinite) possible alternatives to the way things are. Certainly, the world (i.e., the universe) is the way it is. The words we use to describe objects are what they are, e.g., the word *red* has a sense of 'red-ness'. But we could imagine another world where everything is exactly the same as the one in which we currently live, except instead we use the word *derf* to refer to what we know in this world as 'red-ness'. Words aren't the only thing that could be different in an alternative possible world; historical events might have turned out differently, e.g., Julius Caesar might have avoided his death on the Ides of March, or Columbus might have gotten completely lost and not ended up in what we now call the Americas. It is important in language partly because we can entertain concepts and ideas that are not necessarily ground truth in the world that we perceive. For example, I can say something quite absurd such as *I believe that everyone has the same favourite colour*. This is certainly not the case in the real world, but in a different world, i.e., a possible world that I have constructed in my mind, this statement could very well be the case.

The intension of a concept (i.e., a word, term, or expression), then, is a function from all possible worlds, where that concept applies, to the extensions. For our purposes, the intension is a function that picks out of each possible world whichever visually present object fits that description–in that particular world.

Though a seemingly elegant explanation of intension that takes propositional attitudes into account, it isn't completely necessary to hold this view for our models of RR to visually present objects. Resolving an extension using a characteristic function doesn't require us to entertain the notion of possible worlds.

subject for philosophers and linguists alike.

This is more or less another way of putting the shortcomings to FOL we listed above. She continues (p.7):

> The meaningfulness of language lies in the fact that it is about the world.

Which follows from the above section on grounded semantics. She then makes the claim that extensionalism is preferred over intensionalism because, (following Putnam (1973, 1975)), natural language is not the property of individuals, but rather, is a social tool for communication. This focus on meaning in society was also taken up by DeVault et al. (2006) in that the meaning of a word (or, for our purposes, RE) is agreed upon by linguistic communities (see also Section 2.4 on grounding) by their usage, and not by individual speakers. Yet, somehow individuals need to be able to use language with other members of a language community and must somehow have some kind of "mental" approximation of what a word or RE means. In other words, language and meaning is established on societal level, but meanings do need to be represented in individuals somehow (i.e., meaning is to some degree approximated in the head of an individual (Chomsky, 1986; Pietroski, 2003; Daniels, 1990)). It is this interaction between individual and language community where, for the purposes of this thesis, intension should be explored.

Thus (continuing with Dahlgren (1976) p.14),

> Extensions determine intensions, though in a complex way, and not the other way around.

In other words, we can learn ϕ_w (i.e., the intension) by exposure to extensions and REs that have ϕ_w as a concept. To illustrate, suppose, for example, two friends, A and B, are walking down the street together. A is the member of a particular linguistic community (e.g., that of Chinese speakers), and B is not. As they walk down the street, A points to an object and utters something that B has never heard before. B can perceive what is being pointed at, in this case a stationary object, and get an idea of what A meant by the utterance. They continue walking down the street and A points to another object and says the same thing. This continues for some time and B begins to notice that all of the objects which A pointed to had very different features of shape, size, spatial placement, distance from them, etc. However, they all had the same colour, namely what B would call *red*. Later, B points to a red object an utters the word that A had uttered, with positive feedback (e.g., nodding) from A that tells B that the word was used correctly. One might say that, through interacting with someone who is a member of a linguistic community and perceiving objects (i.e., extensions), B was able to learn the word that

picks out the concept for redness (i.e., the intension) in that language community, via reference. As noted in Chapter 1, the same is the case for children learning their first language (Wittek and Tomasello, 2005).

This example illustrates, at least conceptually, what Dahlgren claimed; namely that it is through exposure to extensions that intensions can be learned (i.e., approximated as to how they fit with a linguistic community). The intension, then, becomes the mechanism that assigns real-world entities to classes–something that FOL does not address (as explained in Section 2.3), but something which is essential in a practical dialogue system component that resolves references made to visually present objects. We show in Chapters 5 and 6 (but more particularly the latter), that intensions can be learned via examples of extensions, and that those learned intensions can be later used in RR tasks. It is through this procedure that we address the shortcomings of FOL using grounded semantics, thereby (following Daniels (1990), Larsson (2015), and Harnad (1990)) reconciling to a small degree symbolic and grounded semantics.

2.6 Additional Assumptions

Beyond the background given in this chapter, there are several additional assumptions that need to be established before moving on in this thesis. These assumptions allow us to focus on specific aspects such as grounded meaning and incrementally resolving REs.

Reference Perspective In this chapter, we have only seen examples where the speaker and the listener have the same *frame of reference*; i.e., their perspective on a particular scene is aligned as if standing next to each other. This, of course, isn't always the case when referring to objects. For example, a speaker that observes a scene with objects and knows that the listener is observing that scene from an opposite viewpoint (e.g., from across a table), the speaker might attempt to refer to an object based on the listener's perspective (e.g., as part of the RE, saying *on the left* would cause the listener to look to her left, which is the speaker's right). Work has been done in perspective alignment, e.g., Steels and Loetzsch (2009); Liu et al. (2010), but for this thesis it is assumed that the speaker and the listener have the same perspective of the scene.

Saliency When someone looks at a set of objects and determines to refer to one of those objects, what is it about that object, when compared to the others, that makes that person pick out certain features? For example, if there are 5 objects in the room and they all have the same colour, then using the referred object's colour won't distinguish that object from the others. Objects often have features that distinguish them from others, even features that stand out (i.e., are more *salient*) from other features. Knowing what features stand out from others could be

useful in determining which object is being referred. However, as it is not directly related to REs, we are not interested in this thesis with measuring saliency directly.[23] For recent work on using saliency information to determine visually present landmarks in a navigation task, see Götze and Boye (2013).

Objects Defined What is an object? In this thesis, an object is a visually present entity that is distinct from other entities. In general, these objects also have distinct and unique properties, such as colour (i.e., a single a colour and not made up of multiple colours), shape, spatial arrangement, and size. We are not concerned with how the real world is perceptualised in in a human's brain; whether objects are represented symbolically or as a visual abstraction. We of necessity represent objects (in order to ground language with the world), and how an object is represented does have implications on how grounding takes place. However, in this thesis an object is clearly visible and distinct from other objects and has some kind of representation of which our models make use.

Universe of Discourse / Reference Domain The *universe of discourse* (Boole, 2005) in this thesis is the set of candidate objects that could be referred; i.e., all of the visually present objects that exist in a speaker and listener's immediate surrounding. This is always a small set of objects. It is always assumed that it is one of these objects that is being referred in a given RE, and not some other object that is unseen or was referred to at a different time. The task of resolving references means choosing which object from this set is the referred one. In other terminology, we assume a *reference domain* (Salmon-Alt and Romary, 2001) which are theoretical constructs that are entities which are presupposed at each use of a RE. The reference domain in this thesis is the set of visually present objects (we look into this a little bit more in Chapter 4).

Pragmatic Assumptions There are many speech acts, such as greeting, requesting, confirming, etc. (Searle, 1976). In this thesis, we are only interested in a single speech act, that of RE, such that a listener's attention is directed to an object. We also assume that grounding (i.e., establishing common ground) as in Clark (1996) where mutual understanding between the two participants has taken place to an extent, that the speaker and the listener have learned the meaning of the words used in the REs, but as we won't be looking at REs over the course of a discourse or dialogue, i.e., we mostly look at REs in isolation; that there isn't really an ongoing establishment of common ground. Moreover, we are only interested in two dialogue partici-

[23] This isn't completely true. In Chapter 5, we show in one experiment we take contextual saliency into account–something which that particular model can easily incorporate.

pants: a speaker and a listener (sometimes called the *addressee*); as explained, the models of RR that we explain take the place of the listener.

Other Assumptions As noted earlier in this chapter, we are only interested in referring to a single object (per RE). Certainly, a RE can refer to one, two, or a group of objects, which has been looked into in other work (Sauppé and Mutlu, 2014; Gorniak and Roy, 2004), but here we leave it to future work to handle REs that go beyond a single object. However, the models that are explained in this thesis should easily be extendible to be able to handle reference to multiple objects. It is also common for reference to be made to entities that don't exist in the immediate surroundings, but do certainly exist (e.g., New York, the moon, or a speech that was given). The models presented here could be extended (as we show, one of the models can refer to abstract things, if represented sufficiently) to refer to non-visual entities, but we are primarily concerned with those that are visually present. Also important, is that the objects are visually present at the moment the RE is taking place, and not at some other point in time. The objects can be perceived visually at the same time as the incrementally-unfolding RE event.

2.7 Chapter Summary

To couch the work in this thesis in an area where the models could be implemented as practical components, Section 2.1 explained *spoken dialogue systems*, particularly situated and incremental spoken dialogue systems. The types of referring expressions that this thesis focuses on were explained in the following section, namely definite descriptions, demonstratives, and exophoric pronouns.

Section 2.3 focused on what the listener does when resolving a referring expression and used a well-established formalism of first-order logic to do so. Examples of resolving an object using first-order logic for the three types of referring expressions were given. Some shortcomings were explained, namely

1. the set of classes must be determined
2. the functions that assign objects to those classes must be determined
3. incremental composition

and Section 2.4 addressed some of those shortcomings by appealing to *grounded* semantics.

Section 2.5 looked at philosophical literature on meaning as it pertains to reference and established a small aspect where our models make a contribution to theories of meaning. In short,

the meanings of words and expressions are agreed upon by language communities, but individuals need to be able to approximate those meanings in their own heads in order to use language with other members of that community. Those meanings–intensions–are learned through interactions with real-world objects–extensions–and referring expressions that designate those objects. Learning the concepts and the mechanisms for determining class membership of an object is a contribution of this thesis.

The final section noted some additional assumptions that must be made for us to focus on the aspects of reference that this thesis addresses.

The overall goal is to model the resolution of referring expressions to visually present objects sufficiently that the model can be implemented in a practical component that would be usable in a spoken dialogue system that interacts with a human. Such a system would need to learn some notion of meaning of the words that are used as referring expressions, and it would need to be able to combine those word meanings in a practical way. Moreover, such a component would need to be situated and (update-) incremental because of the nature of the task which requires that objects be visually present.

3

A Review of the Reference Resolution Literature

In this chapter, we review relevant literature and look at how others have approached the task of resolving referring expressions. We are chiefly concerned with approaches to the *comprehension* (i.e., resolution) of referring expressions (Section 3.1), but we will also see what kind of work has been done in the *generation* of referring expressions (Section 3.3). Moreover, as resolving references is a sub-task of general interpretation of natural language utterances, we will also look at some relevant work done in *natural language understanding* (Section 3.2). In many cases, we will see example figures depicting the objects and scenes that are under investigation in the corresponding literature to see what kinds of objects were resolved and what kinds of scenes those objects where found in.

Importantly, as we are interested in reference to visually present objects, we will take note of how other approaches represent the (relevant aspects of) the world W, how they represent the uttered referring expression U, and how they compute the function that maps from U and W to I^*, the referred object, as in 3.1, as introduced in the previous chapter:

$$I^* = f_{rr}(U, W) \tag{3.1}$$

or, the alternative stochastic approach

$$I^* = \underset{I}{\operatorname{argmax}}\, P(I|U, W) \tag{3.2}$$

We take note as to whether or not the cited approaches attempt to ground between language and the world and, when necessary, how they fit into the formal framework, if possible, and determine if they can handle definite descriptions, demonstratives, and pronouns. We will also consider if the approaches work incrementally and, of those that do, determine if they are restart- or update-incremental.

3.1 Previous Work in Reference Resolution

In this section, we will look at models and approaches to RR. We begin with approaches that do not learn a grounded meaning and do not work incrementally (denoted below as *traditional* approaches). These approaches are by definition multimodal because there must be some representation of U and W–two modalities–even though W can be represented in a similar way as a non-visual knowledge base (i.e., a database of facts). However, we will see models and approaches that incorporate additional modalities, such as the eye gaze and pointing gesture of the speaker performing the RE. That review is followed by looking at approaches that learn grounded meaning, then by looking at approaches that work incrementally. Also of interest, but not focus, is if the systems were evaluated using ASR output, as we do.

3.1.1 Traditional Approaches

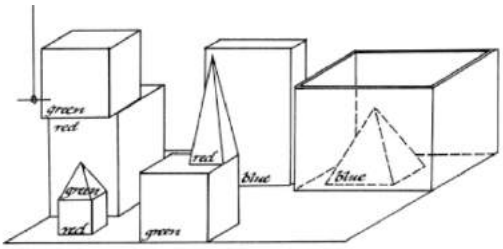


Figure 3.1: Example scene for the SHRDLU system (Winograd, 1971).

Early work included Terry Winograd (1971)'s SHRDLU system for understanding and generating REs to objects in a blocks world. The scene (Example in 3.1) was represented symbolically; i.e., W was a set of objects and properties as they are depicted in the scene without any uncertainty. An excerpt from an example interaction is given in (1). The system (S) could

perform tasks and answer questions about the state of the scene made by the human (H), such as knowing how many pyramids were outside of boxes.

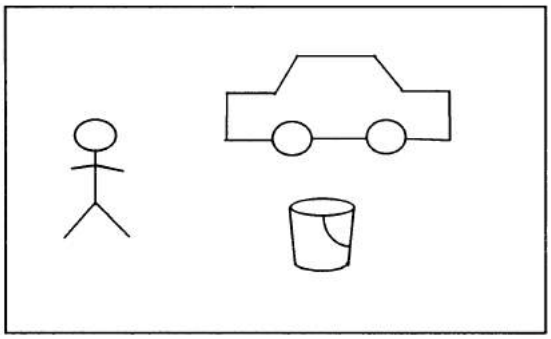
(1) a. H: Pick up a big red block.
b. S: OK.
c. H: Grasp the pyramid.
d. S: I don't understand which pyramid you mean.

In terms of Equation 3.1, U was a semantic representation (a logical form) over the words. The system could only handle specific utterances (i.e., manually typed sentences) that were parsable. A set of rules mapped W and U onto a potential I^*; i.e., given the semantic representation, the logical form could be traversed and a set of rules determined if certain logical concepts mapped to something the system could do or to an object in the scene.

Another well known early approach in resolving references was in the famous *Put That There* experiment (Bolt, 1980). In this experiment, a user was in a room with several large television screens and sensor equipment. Early forms of ASR and gesture recognition were used to determine what a user was saying (using a very limited vocabulary) and what the user was pointing at on the screen. The system could perform several actions, such as creating 2-dimensional objects, or moving those objects. Keywords were used for the actions (create, move, delete, make, or name an object). For example, a user could say *create a blue square there*, and later say *move the blue square ... there*, where the word *there* is accompanied by a clear pointing gesture with an extended arm. Resolving the reference to already-visible objects was either done by triangulating what was being pointed at, or by using a pre-determined mapping between certain colour and shape words to visible objects. The system could handle demonstratives as well as certain definite descriptions (usually accompanied by a demonstrative).

Other, later work approached the task of RR by treating objects in a scene as if they were linguistic antecedents, allowing resolution to objects without a definite description (Pineda and Garza, 2000). This is treating pronouns exophorically, as explained in the previous chapter. For example, if a scene depicts a man, a bucket, and a car, and a RE *he washed it*, it should be fairly clear as to what *he* and what *it* refers to–the man and the car, respectively (see Figure 3.2). There was not prior mention of the man, the bucket, or the car, so the pronouns, which usually refer to linguistic antecedents, now refer to objects in the scene exophorically. Using thematic information, for example, knowing that washing is a particular type of action and that cars can't perform that type of action but a man can, then *he* must refer to the man, etc. Further looked into were spacial descriptions. When considering spatial descriptions such as *city A is between city B and city C*, it was the authors' view that one can translate between scenes

and descriptions in a similar way that one can translate between two natural languages; i.e., by segmenting the scene in a syntactic way that corresponds to the description. At the time of publication, the model was not fully implemented or evaluated.



"He washed it"

Figure 3.2: Example scene for Pineda and Garza (2000). Given visual information and knowledge about the meaning of *washed*, it should be clear what *He* and *it* should refer.

Kehler (2000) took a slightly different approach to RR by focusing on the different types of REs, such as definite descriptions, demonstratives, and pronouns, and what cognitive status the speaker thinks that the listener has. E.g., a pronoun has a cognitive status of being *in focus*, whereas a demonstrative *familiar* and a description is *uniquely identifiable* (Gundel et al., 1993). Their model of RR was quite simple, in that it followed a set of rules in a hierarchical order: (1) if object is being gestured to, it is the referent (2) if the currently selected object meets semantic type constraints imposed by the RE (e.g., *the museum* requires a museum referent), it is the referent (3) if there is a visible object that is semantically compatible, it is the referent (4) a full noun phrase (such as a proper name) was used that uniquely identifier the referent. They evaluated their simple model on virtual data that the author collected: maps of a city area with identifiable landmarks, such as restaurants, museums, street names, etc. To be able to accomplish the type of RE in (1), participants could use the mouse to draw arrows on the screen. An object is selected (2) if a participant clicks on that object and there is an enlarged image of it (see Figure 3.3; "Marta Hari Grill" has been selected and enlarged). For (3), an object on the screen can be referred to by a description such as its type (e.g., restaurant or hotel), and (4) resolves names to objects. Using this approach, their model can resolve the 62 test cases correctly. However, note that the REs which made up U were mapped to the aspects of the scene W either by a human-judged resolution of what was being pointed at for (1), the current property of being selected for (2), if the RE contained the semantic type for (3), or if the full name was used, for (4). Thus no grounding takes place. To fully automate such a system in a computational component, the complete mapping between semantic types and objects would

need to be done beforehand, and the ability for the component to resolve the object to which a drawn arrow is pointing would also need to be implemented.

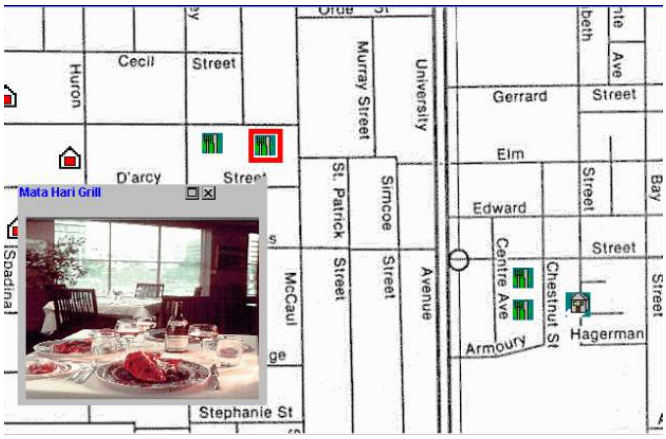


Figure 3.3: Example scene for Kehler (2000). Participants referred to visible objects on a screen.

The author notes that the nature of the task (and the participants' belief that they were talking to a system when it was in fact a human), caused them to perform more explicit REs (i.e., no real use of pronouns). It was assumed that the participants mapped certain aspects of the state of the scene (i.e., selected or not selected) to aspects of the system's cognitive state. The author also collected data where speech was the only allowed modality. It was found that participants more often used pronouns (of varying forms, e.g., *it*, *there*, *that*) when objects were in a selected state. Definite descriptions were used more often to refer to objects that were not yet in focus, e.g., *the hotel on Chestnut Street*, instead of using pointing "gestures". The conclusion is that speech-only doesn't necessarily result in more ambiguity, but in less efficient references (where efficiency here means number of words required to refer).

More recently, data-driven methods have been applied to the task of RR. This generally means that a set of example REs are collected, as well as a representation of the corresponding scenes where each RE took place. These examples can be used to train probabilistic models. How these models are trained and applied vary across approaches.

One approach used eye tracking to help with resolving references. In order to produce a RE, the speaker with the intention to refer to an object must at some point look at that object in order to pick out properties which the object has in order to form a RE based on those properties. Speakers also look at other objects in the scene (known as *distractors*) in order to determine which properties would be more useful to utter. It was shown in Prasov and Chai (2008) that incorporating gaze information improves RR performance. However, their model of RR for speech was somewhat simple. For the most part, a RE was a definite description to a

Figure 3.4: Example scene for Prasov and Chai (2008). The numbers are added for presentation. Participants answered questions about the scene; answers often included REs.

specific object in a scene (e.g., *the bed*) and could pick out that object using a simple, two-word description. See Example scene in Figure 3.4 and example utterances in (2) where the System S asked a question, and the participant produced the utterances U; REs are depicted in boldface type.

(2) a. S_1: What is your favorite piece of furniture?
b. U_1: I would say my favorite piece of furniture in the room would be **the bed**.
c. U_2: It looks like **the sheet** is made of leopard skin.
d. U_3: And I like **the cabinets** around **the bed**.
e. U_4: Yeah I think **that**'s my favorite.

Despite being data-driven in other aspects, words were simply mapped directly to the corresponding objects via a compatibility score based on pre-defined properties that could be retrieved directly from the scene representation. In some cases, pronouns were used to refer to objects (usually, also anaphorically). Certainly, a simple RR model would gain a lot from gaze information. This was extended in Prasov and Chai (2010) to 3-dimensional (still on a computer) scenes using a confusion network over the n-best list of ASR hypotheses. Similar compatibility scores were used as before, only this time with n-best ASR output which was also passed through a special kind of syntactic parser. Adding gaze helped the model improve (from a 0.619 RR accuracy speech-only baseline f-measure to 0.676 using 2204 REs).

More recently, data has been collected with specific interest in RR tasks. The REX corpora presented in Spanger et al. (2012) and Tokunaga et al. (2012) represents a collection of

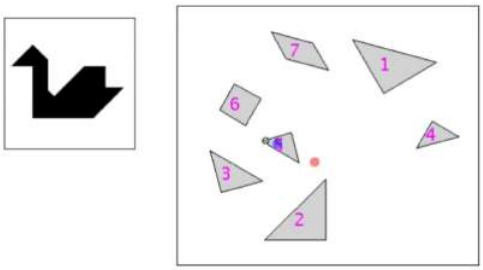


Figure 3.5: Example Tangram Board in the REX corpora; the goal shape is the swan in the top left, the shared work area is the large board on the right, the mouse cursor and OP gaze (blue dot) are on object 5, the SV gaze is the red dot (gaze points were not seen by the participants).

multimodal corpora of REs in collaborative problem solving dialogues. The problem that is collaboratively solved, usually between two human participants, is putting together a virtual puzzle to form a shape out of the pieces where one participant (the instruction giver–IG) knows what the goal shape is and the other participant can manipulate the puzzle pieces (the instruction follower–IF). Examples (3) and (4) show two example REs. The two participants must work together in order to construct the goal shape out of the pieces (example in Figure 3.5). In these corpora, tangram puzzle pieces were used. As this data will be used in an experiment in this dissertation, it will be described in greater detail in Chapter 4. In some cases, there is only speech and scene information recorded. In others, there is additionally gaze information. These data have been used in several approaches to RR which we will now discuss.

(3) a. *chicchai sankakkei*
b. small triangle

(4) a. *sono ichiban migi ni shippo ni natte iru sankakkei*
b. that most right tail becoming triangle
'that right-most triangle that is the tail'

Iida et al. (2010) approached RR with a subset of the REX data using a support vector machine (SVM) that mapped directly from U and W features to an object identifier. These features were mostly binary features, segmented into three sets; one was a set of 10 *discourse* features (e.g., a binary value indicating that an object was the most recently referred one; a binary value indicating that the time distance to the last mention of that object was more than 20 seconds), a set of 6 *action history* features (e.g., a binary value indicating that the mouse cursor was over the object at the beginning of the RE; a binary value indicating that the time distance is

less than or equal to 10 seconds since the mouse cursor was last over the object, etc.); and a set of 6 *current operation* features (e.g., a binary value indicating an object has not yet been manipulated, etc.). They further segmented the REs into a set of definite descriptions and a set of pronouns. Thus there were two models: a separate model and a combined model (which handled both pronouns and definite descriptions). Performing a 10-fold cross validation over 2,035 REs to single objects, the separate model correctly resolved 78% of the time, where the combined model correctly resolved 74% of the time. Both of these are good numbers, resolving on average 3 out of every 4 REs, but they also provided some analysis as to which features were most informative. For definite descriptions, the most informative feature was a binary value indicating that the attributes of an object are *compatible* with the attributes in the RE. This compatibility function was a set of rules that determined if an object had a property that was uttered in the RE. This is where the "meaning" of words is encoded, albeit by hand–i.e., not grounded. The most informative feature for pronouns (in this case, exophoric or demonstrative) was a binary feature indicating that the mouse was over an object. In other words, if the IF had their mouse cursor over an object, and the IG said something like *that one*, then the object with the mouse cursor over it was more likely to be the referred one.

This approach to RR was extended in Iida et al. (2011) with an addition of new data to the REX corpus where the gaze of both participants was recorded.[1] Their approach was augmented and an additional set of 14 gaze features (e.g., the frequency of fixating on an object for a certain time period before the onset of the RE). Again, they used separate models for pronouns and definite descriptions with better results for the separate model. Their analysis of the features found that gaze features were more useful for resolving pronouns, as might be expected.

The model described in Funakoshi et al. (2012), considered an extension to Iida et al. (2010), also used a portion of the REX tangram data. Their approach, however, focused on a single, unified model that could resolve definite descriptions, pronouns, as well as pointing gestures (i.e., the mouse pointer) in a single framework. They used a Bayesian network to model four variables: the observed words, the concepts denoted by the words, the reference of the RE, and the presupposed *reference domain* (the set of candidate, visually present entities). Importantly, the way U and W are mapped is via a closed set of 40 "concepts" (e.g., the word *big* would match to an object having the property of being `big`).[2] The model is also able to handle reference to single objects as well as reference to two objects (making it not directly comparable to previous work). The model can correctly resolve REs to their referent 85.6% of the time (with the best model variant).

[1] In Chapter 5 we compare the results of our model directly with theirs.

[2] The authors hint at a kind of "simulated" perceptual grounding that worked well with object groupings, but not individual objects.

Moving now to different domains and tasks, when shopping online people look at potential objects for purchase presented on their screen and clicking is a way of referring to them. Hakkani-Tür et al. (2014) presented work on using a speech-based browsing system that could verbally "click" on referred objects. They represented U as a full RE, W as a set of features (i.e., properties) of the objects (which were web links) that could be referred. The mapping between W and U was, as we've seen before, a simple check to see if a word in the RE matched the name of a property (e.g., the name of the object). Their model also makes heavy use of gaze features–people certainly look at the objects they intend to refer. Their model performed well on the data set they collected.

The model presented in Engonopoulos et al. (2013) represents W as an *observation model* (i.e., a set of features over the objects in a 3D scene from the GIVE environment (Koller et al., 2010); see Figure 3.6), and U was a *semantic* model that abstracted over the RE. The mapping between W and U was done in a similar way as other approaches we've seen so far; pre-determined rules for matching words with properties of objects.

Figure 3.6: Example scene from the GIVE environment, used in Engonopoulos et al. (2013).

Resolving references while driving was also the task in Misu et al. (2014). Their model of RR attempted to resolve REs to points of interest along the road (e.g., a business or a restaurant within the viewing area around the car). This is a challenging setting, as the car was moving and thus the domain-of-discourse changed over time; the potentially referred points of interest at the onset of one RE were different from the next RE. In order to determine the points of interest (W), they used geolocation information. They also incorporated a face tracking algorithm to determine what general direction the driver was facing when performing the RE. Most of the effort here was put into computing a distribution over candidate points of interest based on context (where the car was at the onset of a RE) and the direction the face was pointing. As with the other models we have seen so far, processing U amounted to finding words in the RE that matched pre-determined properties that a potential point of interest had (e.g., if it was the right category such was business or restaurant). They found that a big determining feature was spatial location relative to the car (the participants in their study often used *left* or *right* in the REs). Besides showing that their model had respectable reference resolution performance, they

provided additional analysis as to what features were useful under various circumstances (e.g., downtown area vs. residential area).

Discussion

As seen in this section, the task of RR can be applied in various areas, such as objects in a room, online shopping, or driving in a car. In each of the approaches in this section, there is a way of defining a function between W, U, and the intended referent I^*. In some of the cases, aspects of that function were learned from data. As the setting is situated dialogue, we have seen some work in using contextual information (e.g., where a mouse cursor is pointing, or where a speaker is looking) to resolve REs. However, none of the above cases has any notion of grounded word meaning. In each case, there is a way of incorporating the words of the REs to resolve the reference, but it is usually with a set of rules based on human judgements. Often it is just a simple check to see if a word in the RE is similar to a concept or property that an object has. The meaning of words is not necessarily learned; the contribution of words amounts, in general, to binary features that are used in a RR model (that is otherwise probabilistic).

In terms of representing the intension of words these approaches use a simple pre-defined functions that map from words to the objects that have certain properties. The functions are not learned; rather, they are specified by hand, and the set of features x is usually a set of discrete, pre-defined properties, i.e., the classes and the mechanisms for assigning objects to those classes are predetermined. In some of the above-cited work, we saw that additional modalities such as gaze improve the RR results. Certainly, this is a useful modality, but it is argued in this thesis (and as will be shown in later experiments), more can be gotten from the spoken RE if grounding occurs. In the following section, we will look at other approaches to RR that learn a grounded word meaning.

3.1.2 Approaches with Grounding

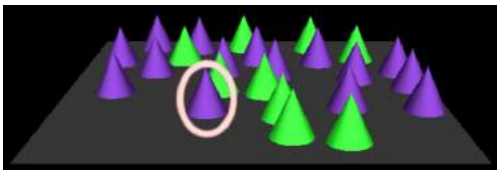


Figure 3.7: Example scene from Gorniak and Roy (2004), selection oval added for presentation.

We mentioned work by Deb Roy and colleagues in learning grounded meaning in the previous chapter. Some of their work included directly resolving references. In Gorniak and Roy

(2004) the focus was less on grounded meaning and more on learning grounded semantic composition.[3] That is, the meaning of an utterance depends on (potentially more) than the meanings of its parts, where one of those parts is the visual context. W was a (sometimes quite dense) set of 3-dimensional cones of only two colours (Example in Figure 3.7, corresponding RE in Example (5)). Participants could refer to single objects or groups of objects. They used colour terms, but more focus in this paper was put on spatial terms, as spatial descriptions tend to be more compositional (e.g., there are two REs in Example (5), each depicted in boldface type, and a relation *to the left of* between them).

(5) a. **the purple cone in the middle** to the left of **the green cones**

This was extended later in Gorniak and Roy (2005a) where the authors developed a *framework for understanding situated speech* (FUSS). One contribution of this system was to handle ambiguity and uncertainty in the ASR hypotheses. The authors pointed out that spontaneous, spoken speech is notoriously difficult to transcribe, which is the task of ASR. Many ASR systems provide an n-best list of hypotheses (in descending order from most probable to lowest), not just the single top (argmax) hypothesis, as there could be useful information in the other hypotheses. They collected interaction data (described in Gorniak and Roy (2005b)) between participants and an experimenter who controlled a computer game, but only took instructions from the participant. The game was a simple puzzle scenario, example in Figure 3.8, where the goal is to light both fire bowls at the same time. One chest contains a key that unlocks the second chest, which contains a key that unlocks one of the doors. One of the levers opens the door to the second chest, whereas the other two levers each light a fire bowl. Thus the participant played the role of one of the players, the system (the experimenter) the second role, as instructed by the participant. Thus simple RR did occur, albeit to abstract objects that didn't particularly look like their real-world counterparts. The goal of this paper was to improve what could be used from inherently noisy ASR output. Their approach used a confusion network to represent multiple ASR hypotheses as a kind of string, which was then syntactically parsed. Their approach learned a grounded meaning between confusion network segments and entities in the game. Given the difficult nature of the task, full understanding (which often included a reference to an object) only occurred in 56% of the utterances.

In Reckman et al. (2010), the authors used a virtual game scenario where specific words were learned from referred objects. Using a simple assumption that referred objects are most talked about around the times when they are involved in actions, a system component can

[3] Gorniak and Roy (2004) provides references to relevant literature in grounded word meaning for specific types of words; e.g., colours, shapes, or action verbs. They are indirectly related to the work in this thesis, but will not be directly referenced here.

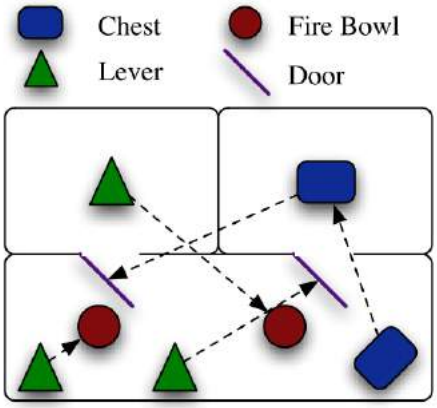


Figure 3.8: Example scene from Gorniak and Roy (2005a); example diagram of map scene. Dashed links indicate dependencies between objects.

determine if an action has occurred and then look at a window of words around that action. They learn the meaning of words using simple co-occurrence; at the moment an object is manipulated, all words within a time window are observed to co-occur with that object with the assumption that certain words that relate to that object co-occur more often than other words. An additional check provided the co-occurrences between words and objects to not be spurious; the authors only used cases where the observed co-occurrence was higher than the co-occurrence that would be expected if words and objects were distributed randomly over the game. The setting of the game was a restaurant, and a pattern did emerge: co-occurrences between words and objects seemed sensible for food and drink items as well as utensils, the menu, and the bill. Knowing what words and objects co-occur, the authors looked into what words or phrases were actually referring (albeit to food items that were not necessarily visually present during the RE, they should have existed on the menu). Their system was able to pick out referring words, such as food words, from a vocabulary of words that were uttered in all the dialogues, and those words were learned to co-occur with certain objects.

A commonly used domain that involves RR is in map navigation tasks. In a city, when a person is giving another person route directions to a specific city attraction, often those instructions have landmarks to help the directee find his way to the desired destination. For example, the direction utterance ... *and then turn left at the big sandwich restaurant*. Smaller scale directions, such as inside a building, also use landmarks; e.g., *go through the doors, then turn right at the elevator*. These landmarks are a type of visually present reference, however the listener (directee) will not see those objects until after the reference to them has been made. Such a map

setting was used in a data collection has been used in several papers on language understanding; specifically, they used the collection of map data described in Anderson et al. (1991), a collection of human/human dialogues involving cooperative path planning using maps (example scene in Figure 3.9, example instructions in (6); boldface type depicts references to landmark objects). An instruction-giver instructed an instruction-follower to navigate through the map following the indicated path (the path was only visible to the instruction-giver).

(6) a. go north
b. go past **the house**
c. you are right by **the forest**

In Levit and Roy (2007), landmarks (which are referred objects in a route instruction) were used to learn a grounded model for specific route phrases (e.g., *toward*, *away from*, *between*). Following Bugmann et al. (2004), they break down navigational instructions to *navigational information units* (NIUs; e.g., moving around objects, moving in absolute directions, turning, or verifying closeness to a landmark). In their approach, they show that most of the NIUs consisted of a type of move and a corresponding reference object (landmark). Some of the instructions use landmarks and some do not. This work was extended in Kollar et al. (2010) (albeit with different data), where spatial descriptions (e.g., *until*, *past*) were learned. The authors described a model that produced a topolographical representation based on a route instruction. Their model learned a grounded meaning for words like *past* (e.g., *walk past the door*) where a route had to have a starting point, an ending point, and some kind of relation to a landmark (e.g, *past* amounts to having the landmark on the route path between the two points without the route touching the landmark; *to* amounts to the landmark being at the end point, and *through* amounts to the route going through the landmark). In order to learn such grounded meanings, referring to landmarks (i.e., references to objects) was necessary.

The work presented in Vogel and Jurafsky (2010) also applied a novel model of grounded-language understanding to the map task. Like Levit and Roy (2007), the authors focused on spatial terms such as *above*, and *south*. Their model used a reinforcement learning approach where the task of understanding navigational instructions was accomplished due to a learned policy of following a path, given information about the map W and the instruction U from the instruction giver. In such an approach, a reward function R measures the utility (reward) of executing an action a in a particular state s: $R(s, a)$, where the state encoded information about the state of the world, the current location in the world, and the current instruction. Grounding in this case amounts to learning the reward function for all the states and utilities. After training their system, they found that when a spatial word such as *above* was used, a higher reward was given when the chosen action forged a path above an object. Thus their system, replacing the

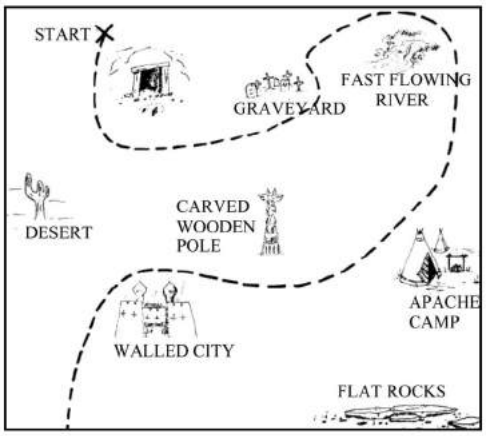


Figure 3.9: Example map task scene from Anderson et al. (1991).

instruction-follower, could arrive at the goal destination on the map via the route as depicted in the instruction-giver's map.

Following map instructions was also the task in Artzi and Zettlemoyer (2013), using data described in MacMahon et al. (2006) (example scene shown in Figure 3.10, example instruction and corresponding annotation information–scene and plan–are given in Example (7); REs represented in boldface type). For U, the authors used a semantic abstraction over the utterance in the form of *categorical combinatory grammar* (CCG). In short, this formalism provides a syntactic parse that gives rise to a logical form similar to FOL. Their approach maps from a state S (i.e., where on the map one currently is and how one is oriented) to a set of actions A (either `left`, `right`, `move`, or `null`) using a natural language instruction (i.e., an utterance). Their model learns a lexical semantics of the words, as well as the syntactic parse and logical form automatically. Their GENLEX function (Zettlemoyer and Collins, 2012) does something similar to what we spelled out earlier: learn a meaning (in this case, a rule that maps from syntax to a semantic representation). For example, introduce a category $N : \lambda x.chair(x)$ for any logical form z that contains the constant $chair$. Thus their GENLEX algorithm learns a lexicon of concepts such that they can be composed into larger logical forms.[4] Their evaluation is based on successfully predicted sequences of actions to reach the goal point on the map; resolving references was necessary in order to accomplish that goal.

[4]Their approach does not learn a perceptually-grounded meaning of individual words, rather, they claim to be learning grounded meaning of instructions (i.e., a mapping from states to a set of actions). Once a logical form is produced, they have a set of rules that map from the concepts in the logical form to properties of objects in the scene, when referring to an object is appropriate.

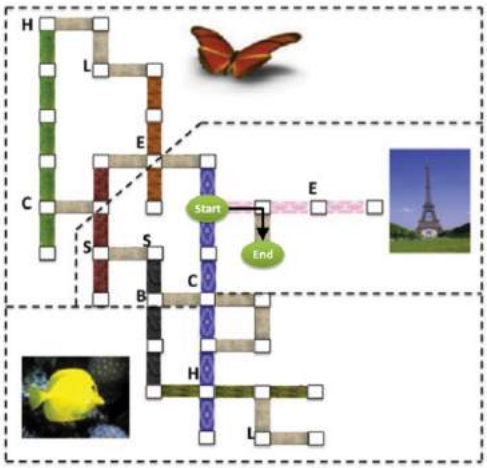


Figure 3.10: Example scene from the data map-task data collected in MacMahon et al. (2006)

(7) a. Instruction: *at the very* ***next intersection*** *take a right onto* ***the plain path*** *and follow* ***it*** *to the end of* ***the hall***

b. Landmarks: Travel(steps:1)

c. Plan: Verify(side:CONCRETE HALLWAY),
Turn(right),
Verify(back:WALL, front: CONCRETE HALLWAY, left:EASEL),
Travel(steps:1),
Verify(front:WALL)

Approaches to learning how to follow natural language instructions in a map task using the same data were also presented in Kim and Mooney (2013); Chen and Mooney (2011). As with other attempts, their approach learned a mapping between a state S and a set of actions A. In Kim and Mooney (2013), they use a *probabilistic context-free grammar* (PCFG) which provides a syntactic set of candidate parse trees over the instruction, these were then used to produce the set of instructions. The one that produces the most likely set of instructions was chosen as the parse. Their parse trees are decorated with instructions at certain nodes. The parse was then composed to form a single set of instructions.

Now moving away from map navigation and into how language grounds into sensorimotor information in a robot, the approach presented in Hsiao et al. (2008) used the task of resolving visually present objects to build a model of *object schemas* which enables a robot to encode beliefs about physical objects in the scene using processes responsible for sensorimotor inter-

action. That is, W is in part represented by the sensorimotor information from the embodied robot (e.g., an object is *graspable* if a robot's fingers can fit around it and pick it up, vs. an object that is not). The mapping function of U and W is done using a search algorithm that takes a representation of U (in this case, a parse tree) and searches W for the best fit of the RE in U.

The approach presented in Dindo and Zambuto (2010) learned a model of grounded word meaning from perception and grouped words into classes based on sensory channels (i.e., features; e.g., colour based on channels that correspond to the RGB values). The focus of this paper was learning the semantic categories (which could correspond to classes in W). The words in U associate meanings probabilistically as they are exposed to scenes and corresponding REs to objects. The scenes were between a robot and a human, where the robot knew very little about the meaning of words and acquired them through interactions with the human. An example interaction between Human H and robot R is given in Example (8).

(8) a. R: looks at object 1
b. H: *grasp the object to the left of the blue one*
c. H: points to object 1
d. R: looks at object 3
e. R: *is it the yellow rectangle?*
f. H: no

Matuszek et al. (2014) also moved away from virtual scenes and worked with scenes of tangible objects (example in Figure 3.11). They collected data from participants who described these kinds of scenes and learned a model of grounded meaning of the words. An example description is given in (9).[5]

(9) a. Three rectangles. These three. Two of them are blue. Two blue red rectangles and one red rectangle.

In such a scene, representing W is a more difficult task than reading W directly off a virtual scene where objects and corresponding properties are known. They represented the scene by producing a distribution over object properties such as colour and shape using a classifier, for each object in the scene. The mapping between U and W was a simple language model approach that treated each word as a boolean feature and the distribution over properties (representing W) as features in a logistic regression classifier. Their model also incorporated

[5]Note that in describing a scene, objects are being referred, but the goal is to describe their visual features in propositional terms, e.g., *this is blue*.

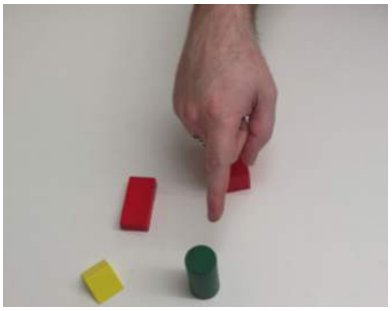


Figure 3.11: Example scene from Matuszek et al. (2014); scenes with real, tangible objects are used where participants can point directly to referred objects.

gestures (using a novel pointing recognition approach) to determine what object was being referred. Their model was able to resolve the correct object fairly well without gestures, but found some minor improvements when gestures were incorporated. Some post-analysis found that errors were mostly produced by their interpretation model (i.e., ASR and their language model), which was a fairly simple approach.

The method presented in Guadarrama et al. (2013) and extended more recently in Golland (2015) focuses on using reference in human-robot interaction tasks to learn the grounded meaning of spatial terms; e.g., *the thing in the back behind the can of spam*. They focused on using non-virtual scenes; i.e., with tangible objects on a table. Reference was made to objects using a latent ontology of spatial concepts. The mapping between the words and the concepts was done automatically (hence, grounded), but the the link between the ontology and the objects was determined by hand.

Discussion The approaches to RR in this section went beyond the traditional approaches to RR in that they all attempted to learn a meaning of language (represented as words, or as semantic or syntactic abstractions over words) that was a function of how language was used in context. In some cases, the meaning of colour words was learned; in others spatial words were learned based on how those words were used, for example, in navigating a path on a map. In most cases, a probabilistic approach was applied to the task, where training data was used to learn the function that maps the word onto the aspects of the world which that word was observed to co-occur with, such as object properties representing an object's colour or shape. As we have seen, most scenes are virtual, which is by nature symbolically represented, and hence can be directly represented; properties of objects can be read directly from the scene's representation.

This is precisely how the model presented in Chapter 5 works. Though designed to work with uncertainty in the world representation (e.g., similar to the scenes described in Matuszek et al. (2014)), it is shown in a series of experiments that the model works best when the scene is a fully-symbolic, virtual scene, as most of the scenes were in the work presented in this section.

The approaches thus far presented in this chapter have been modelled on utterance- or sentence-level input. Of necessity, some of the above approaches that produce syntactic or semantic parses over the sentences expect–indeed require–that the sentences be fully grammatical. In this thesis, we are concerned with modelling language phenomena such as RR such that it can be implemented in a component of an interactive dialogue system. In such a system, utterances are often ungrammatical. Such systems also use ASR which can produce errors in the hypothesised transcriptions. Many of the above-mentioned approaches have not been tested on ASR output, nor on ungrammatical sentences. Importantly, they are not incremental; i.e., they have not been tested with growing prefixes of sentences.

In the following section, we take look at approaches to RR that are incremental (though not in all cases grounded).

3.1.3 Incremental Approaches

Resolving references incrementally to visually present objects was looked at in Stoness et al. (2004, 2005) to provide feedback to earlier processing modules. For example, given a sentence that ambiguously referred to an object, e.g., *put the apple in the box in the corner* (which could mean the apple is in the box, or that the apple is not in the box, yet the box is in the corner) knowing something about the scene and that there is an apple in a box could provide useful feedback to a syntactic parser that provides parses with both possible attachments to disprefer the parse that does not resolve to a visually present object. The authors found that their parsers produced far fewer parse hypotheses when given feedback information. The approach to reference resolution was the approach described in Tetreault and Allen (2004) (though it mostly focused on pronoun resolution) which was not grounded; rather, it followed a rule-based approach such as those we saw above. The approach was incremental to a degree, as it generally worked with phrases (e.g., REs) before it attempted to resolve, but it worked on a finer-grained level than sentence-level. Similarly, Schuler et al. (2009) described a framework for incorporating referential semantic information from a world model directly into a language model (similar in spirit to feedback from a world model to the ASR); their approach was incremental, though the reference objects were not necessarily visually present.

The work presented in Peldszus et al. (2012) also used RR as a task to prune away unlikely syntactic parses. Their main contribution was a semantic processing module that was robust against ill-formed input (again, common in speech and ASR), respected both syntactic and

pragmatic constraints, used a principled semantic representation (*robust minimal recursion semantics* which will be described in greater detail in Chapter 5), and worked incrementally; that is, it produced underspecified semantic output monotonically at each word increment. However, as with work presented above, the extension (denotation) of the REs was determined by a set of rules which checked if a word matched the name given to a specific symbolic property. The data used in the evaluation was from the Pentomino (PENTO) puzzle piece domain (also used in, inter alia, Fernández et al. (2007), which also looked at reference as a task, but the focus of that paper was not on modelling RR). An example of this kind of scene is in Figure 3.12; example instructions given in Example (10).[6]

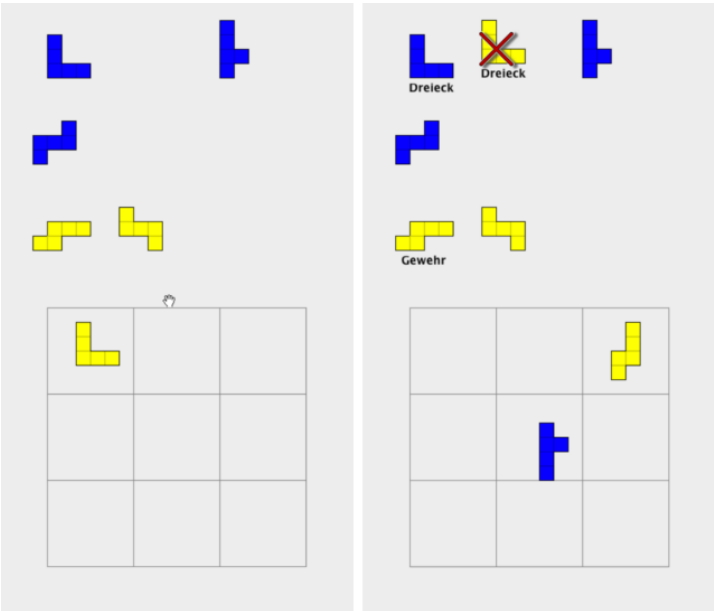


Figure 3.12: Example PENTO scene.

(10) a. move **the yellow v in the top left** to **the bottom**.
b. take **the piece in the middle on the right side**.

Siebert and Schlangen (2008) also used the PENTO domain (albeit with different data; the task was similar to the one described above) to resolve references to visually present PENTO objects. Features were extracted from the scene (e.g., object features such as size, length, shape; topological features such as groupings and distance to other objects) which made up W, and a grounded meaning of the words was learned using a set of tags that determined what

[6]We will see PENTO data more throughout the course of this thesis; several corpora will be explained in greater detail in Chapter 4.

contribution each word in an utterance makes (e.g., if it belongs to the referred object or to a nearby landmark; whether it is a colour word or a shape word, etc.) which, along with the words, comprised U. They then learn the meaning of a word by extracting all instances in the corpus where the word is found and identifying which features in the corresponding scene predict the appropriateness of that word. This was done by simple co-occurrence counting between features and words, but additional processing was done to filter out irrelevant features (for a particular word); that is, if the variance of that feature and word is above a certain threshold (determined by hand, but the authors hint that it could be determined automatically). Their model was able to resolve the referred object, given U and W, correctly 80% of the time (resolving 1 out of 7 objects; baseline of 14%).

Schlangen et al. (2009) also used the PENTO domain for application of a RR model. They used a Bayesian filtering model where the intended referent is treated as a latent variable that generates a sequence of observations. Formally,

$$P(r|w_{1:n}) = \alpha * P(w_n|r, W_{1:n-1}) * P(r|w_{1:n-1}) \tag{3.3}$$

where $P(w_n|r, W_{1:n-1})$ is the likelihood of the new observation which is modelled by referent-specific language models that approximate the joint probabilities of reference and word-sequences (n-grams in the RE; i.e., object names are part of the language model sequence, e.g., for the RE *the red circle* referring to piece X, the bigram sequence would be `the_X`, `red_X`, `circle_X`). $P(r|w_{1:n-1})$ is the prior at step n and the posterior at step $n-1$ (at the initial word, this is just a uniform distribution over the possible referents), and α is a normalising constant. Thus their model was grounded, as it learns these joint probabilities from data. Their model was able to take disfluencies into account, such as filled pauses (i.e., the speaker took extra time in producing the RE) in that the filled pauses actually provided useful information to the model and improved the belief as to which object was being referred (e.g., an object with an unusual shape would be more difficult to describe than an object that has a common shape, causing disfluencies). Importantly, their model was update-incremental; it maintained a belief state in the form of a distribution over the potentially referred objects which was updated at each word (i.e., new information was not recomputed). They report that in about 55% of the cases, their model referred to (i.e., the argmax of the distribution) was the intended object by the end of the RE. They also report incremental metrics which we will use in the evaluation of our models in Chapters 5 and 6.

Discussion Though ample work has been done in RR as a task, fewer approaches learned grounded word meanings, even fewer resolved references incrementally, and fewer still did

both. Chapters 5 and 6 present two models that do both (albeit in somewhat different ways from each other).

3.2 Natural Language Understanding and Reference

In this section we will look into some literature on *natural language understanding* (NLU). We will define NLU, see some examples of NLU in the literature, and compare it with the task of RR.

NLU (also called *spoken language understanding*–SLU–we use both interchangeably here), is defined in Hazen (2011) as "interpretation of signs (e.g., words) conveyed by a speech signal". *Interpretation* can be seen as a classification of groups of signs into classes being identified by a semantic label describing a type of semantic constituent. As put in Tur et al. (2012) for the setting of dialogue systems, NLU "aims to automatically identify the domain and intent of the user [speaker] as expressed in natural language and to extract associated arguments or slots to achieve a goal." Thus NLU goes a step beyond a meaning representation (i.e., where words are converted into logical constants, and their relations are annotated; e.g., a logical form such as FOL) and attempts to determine what the *intention* of the speaker was by uttering what she did.

The example in (11) shows an utterance in (11-a), a FOL abstraction over that utterance in (11-b) which is a meaning representation, and an NLU interpretation of that meaning in a useful format (in this case, a semantic frame (Fillmore and Baker, 2001)) in (11-c).

(11) a. Could you please hand me **that red one**?

b. $\iota xy.\ speaker(x) \wedge thing(y) \wedge red(y) \wedge give(x, y)$

c. $\begin{bmatrix} \texttt{DIALOGUE-ACT} & \text{request} \\ \texttt{REFERENT} & \text{red+thing} \\ \texttt{ACTION} & \text{give(REFERENT,speaker)} \end{bmatrix}$

Note the RE in (11-a) denoted in bold typeface, which is represented as an entity x in the FOL representation in (11-b), and as the slot value `referent` in the frame in (11-c). This frame gives more practical information for a later component (such as the dialogue manager, as explained in Chapter 2); if such a component controlled a robotic arm, this information would be more useful than the logical form in (11-b) to tell it what the speaker intended; namely, a request to hand over a particular object. The component would need to be able to interpret the constants in (11-c) in order to perform the action to move the robotic arm.

We can see from this example that NLU is a slightly broader task than RR. Like RR, NLU is concerned with going beyond a meaning representation and interpreting an speaker's intention

by performing an utterance–though RR assumes that the intention is referring to an object.

Note, however, (following Heintze et al. (2010)) that the RE represented abstractly in the REFERENT slot in (11-c) has not yet been resolved. The whole NLU frame breaks down the utterance, determines the overall goal of the utterance (i.e., the dialogue act–the first slot), the referent, and the action to be performed (slot 3); namely, giving the referent to the speaker. Resolving the speaker is done to the person who last spoke, that is, made the request, is the recipient of the give action. But what object does the system give to the speaker? The NLU abstraction only picked out the words that belong to the RE for the object to be given, but it didn't actually resolve which object it was. This is an additional step that is sometimes a sub-component of NLU, but in any case it is the component that we are concerned with: a component for resolving REs. A frame with a resolved referent, would be more like the example shown in (12) where the identity of the object has been resolved, uniquely from the other visually present objects (e.g., object with ID 3 out of a potential scene with 12 objects, where the component assigns each perceived object a unique ID).

(12) a.
$$\begin{bmatrix} \texttt{DIALOGUE-ACT} & \text{request} \\ \texttt{REFERENT} & o_3 \\ \texttt{ACTION} & \text{give(REFERENT,speaker)} \end{bmatrix}$$

In non-situated dialogue systems, the NLU component doesn't necessarily need to perform this kind of RR. For example, we mentioned in the previous chapter the *Air Travel Information System* (ATIS) (Dahl et al., 1994; Hemphill et al., 1990) which is a commonly used data set for NLU research. An example utterance and corresponding (annotated) NLU frame are shown in Example (13). The corpus has 17 different intents (i.e., dialogue acts; e.g., *flight* means the user wishes to book a flight; this makes up the `goal` slot in (13-b)). In order to fill the slots in the frame, it can be treated as a tagging task known as *concept tagging*, where the slot values are the tags and the values are the corresponding words in the utterance.

(13) a. What flights are there arriving in Chicago after 11pm?

b.
$$\begin{bmatrix} \text{GOAL} & \text{flight} \\ \text{TOLOC.CITY_NAME} & \text{Chicago} \\ \text{ARRIVE_TIME.TIME_RELATIVE} & \text{after} \\ \text{ARRIVE_TIME.TIME} & \text{11pm} \end{bmatrix}$$

Once each word is tagged, a database query can be created (by a set of rules) and executed which returns the desired information. This kind of NLU task is quite useful for speech-based database information lookup tasks. It has been applied to hotel booking in the French MEDIA

corpus (Bonneau-Maynard et al., 2006) as well as transportation information in the Polish LUNA corpus (Marciniak et al., 2010) and the Let's Go! corpus (Raux et al., 2005) which are all collected telephone conversations, and are hence non-situated (in the way that situated has been defined in Chapter 1).

Most approaches to NLU using concept tagging have applied and compared various machine learning methods, where features are generally words (and, in some cases, part of speech tags) within a certain context. Meurs et al. (2009b, 2008b) applied *dynamic Bayesian networks* (DBN), a graphical model approach, to NLU in the MEDIA corpus, where words and phrases were related to each other in the DBN structure. This work was extended in Lefevre (2007); Meurs et al. (2009a), which used multiple-levels of DBNs to produce input for a *conditional random field* (CRF) to predict the slots. Markov logic networks (MLN), another graphical model approach, were also applied to the MEDIA data in Meurs et al. (2008a). Meza-Ruiz et al. (2008) applied MLNs to ATIS, yielding respectable results when considering data across the entire utterance. Hahn et al. (2011) provided a comparison of various machine learning methods and applied them to several tasks (including MEDIA and LUNA). Dinarelli et al. (2012) attempted to perform NLU on several tasks by adding long distance dependencies by re-ranking a typical model of NLU using features from dependency information and a SVM classifier.

Chinaei et al. (2009) applied a more involved method of NLU using unsupervised *Hidden Topic Markov Models* (HTMM) for recovering the user intention. More recently, Tur et al. (2012) applied deep convex networks for semantic utterance classification, a task similar to NLU (where the utterance domain is determined, rather than the intent). Another approach to discriminative classification was applied to ATIS in Mairesse et al. (2009), where *semantic tuple classifiers* were used.

Another, recent and novel approach also deserves mention. Henderson et al. (2014) used "deep learning" recurrent neural networks, not to the task of NLU directly; rather, they attempted to treat the semantic frame as a latent variable and directly predict a dialogue decision. Such an approach is feasible in a non-situated, minimally-interactive task such as this where the words of the entire utterance provide the observed variable which is used to predict the dialogue decision.[7]

A task that required understanding of fairly complicated speech was presented in Liang et al. (2013). The data task was to retrieve facts about United States geography (facts stored

[7]Despite the original definition that NLU provides an interpretation of the intention, it doesn't provide or make use of individual word meanings, which is something we are interested in. For example, in (13), we see that the word *Chicago* is tagged as TOLOC.CITY_NAME. This tells us that *Chicago* is a city's name, but not which city, nor does it tell us what a city name actually is. The city name is a pre-defined constant (e.g., a database table column) that the system can use. Meaning is by no means represented anywhere, though this kind of NLU approach works for practical database lookup systems.

in a database called the GEO corpus), using a hand-typed utterance (not speech, in this case), e.g., *state with the largest area* should return Alaska.[8] To get the proper result, their approach applied a semantic abstraction over the typed utterance the authors called *dependency-based compositional semantics* (DCS). More formally, a probabilistic model learns the mapping from a question x to a latent logical form z, which is then evaluated with respect to world w, in this case a database of facts, which produces an answer, y; see Figure 3.13. The resulting DCS tree can then be traversed to generate queries.[9]

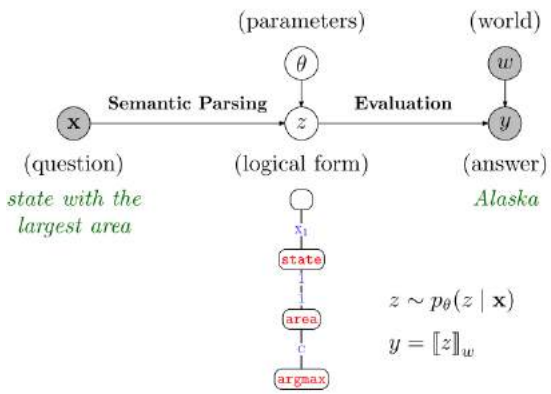


Figure 3.13: Example DCS parse for the utterance *state with the largest area*, taken from Liang et al. (2013)

The answer is an entity such as a city, state, or naturally occurring entity such as a river, which in some ways is resolving a referring expression (albeit in an attributive way); the intention of the speaker is always a question, and the desired result is always an answer in the form of the name of an entity. This task is somewhat different from the task presented here, however, where we are interested in directing someone's attention to a visually present object.

We already looked into some situated approaches to NLU in Section 3.1 that, of necessity, had some kind of component (or part of the NLU model) that resolved referring expressions (or, the entire model of NLU was in fact a model of RR), such as the map direction task approaches. In some cases, these were not grounded approaches; a direct mapping from U to W was done via rules; in other cases, a grounded word meaning was learned from data.

Efforts have been made to model NLU incrementally. This amounts to filling a frame (ei-

[8]No, it's not Texas.

[9]In fact, when traversing a DCS tree to produce a query, nodes in the tree map directly to database operations (e.g., join, aggregate, execute, etc.). This kind of semantic parser is quite effective, provided the world is represented in such a pre-defined database, and that questions and answers can be provided. This, along with the concept-tagging approaches to NLU seem to show that approaches to NLU depend heavily on how the data they work with is represented (an insight credited to Jana Götze).

ther pre-defined or one that is dynamically being built) as the utterance unfolds, word for word. Traum et al. (2012) presented a fully-functional dialogue system that handled incremental understanding and feedback in a situated, multi-party situation. Their NLU component produced semantic representations and predictions of final-utterance meaning when given (partial) ASR output. Their approach was restart-incremental as information that was processed in a previous increment was re-processed (see Chapter 2 for a definition of types of incremental systems). Though entities in the discourse and world needed to be resolved to concepts in the NLU, this generally amounted to pronouns (e.g., *I*, *we*, etc.) and not to visually present objects.

The approach to NLU presented in Kennington and Schlangen (2012, 2014) was also (restart) incremental, and for part of the NLU frame, reference was made to visually present objects. Reference could be made via definite descriptions or exophoric pronouns to PENTO objects on a virtual screen (similar to 3.12), example utterances (in German; this corpus of data will be explained in greater detail in the next chapter) and corresponding NLU frames in (14) and (15), REs depicted in bold typeface. The utterance in (15) follows directly after the utterance in (14), both the definite description and pronoun refer to the same object in the scene, identified as o_4.

(14) a. *drehe die Schlange nach rechts*
b. rotate **the snake** to the right
c. $\begin{bmatrix} \text{ACTION} & \text{rotate} \\ \text{OBJECT} & o_4 \\ \text{RESULT} & \text{clockwise} \end{bmatrix}$

(15) a. *drehe sie nochmal*
b. rotate **it** again
c. $\begin{bmatrix} \text{ACTION} & \text{rotate} \\ \text{OBJECT} & o_4 \\ \text{RESULT} & \text{clockwise} \end{bmatrix}$

This is closer in spirit to the work done in this thesis, albeit the model (applied *Markov Logic Networks*) were slower than could be realised in a real-time dialogue system. It is useful to note, and we will explore this further in a later chapter, that in reality all three slots (a pre-defined set) were filled with separate models; each model was essentially performing a RR task, albeit the OBJECT slot was to visually present objects.

3.3 Generation of Referring Expressions

Though not the main concern of this thesis, in this section, we will look at generation of REs (henceforth *referring expression generation*–REG). That is, in a situated dialogue setting, what in particular the speaker is doing in order to produce her RE (until now we have only considered what the listener is doing during the comprehension of a RE). We take this slight detour for several reasons. First, generation and comprehension of REs is presumably done using similar methods (to a certain point!). Second, the model of RR presented in Chapter 5 is explicitly modelled generatively; indeed, the story behind the model is from the viewpoint of the speaker, turned on its head to be used for comprehension. A review of the most relevant literature, therefore, in REG will provide useful background for modelling RR.

Krahmer and van Deemter (2012) provides a fairly recent survey of computational research on REG. As RR is a sub-task of NLU, REG is a subtask of natural language generation (NLG), the latter of which is concerned with the process of automatically converting non-linguistic information (such as a scene representation) into natural language text. Within NLG, REG has received a good deal of attention. Of necessity, all NLG must contain some kind of REG component or sub-model of some kind (as is the case with RR within NLU). The authors cite several examples, such as when providing flight information, if someone asks for the cheapest flight, then that particular flight needs to be found and the NLG response must contain a reference to that particular flight. Depending on the context, a REG component must determine whether to use a pronoun or a definite description.[10] For systems that can produce definite descriptions, an important consideration is: what should be said? I.e., which set of properties distinguishes the referent from the others, and how should that selection of properties in turn be used to produce a natural language RE?

The authors note several early systems, including SHRDLU (Winograd, 1971), which we looked at above. Also mentioned was Dale and Reiter (1995) which was concerned with the link between Gricean maxims and REG. Consider the two examples in (16):

(16) a. sit by **the table**
b. sit by **the brown wooden table**

In a situation where there was only one table which happened to be brown and wooden, a listener would successfully resolve either RE. However, the additional information in (16-b) mentions information that isn't useful to the listener, which could trigger the listener into questioning, for example, if there is another table somewhere that he did not see (i.e., the RE in

[10] See Gundel et al. (1993) which looks at how cognitive states affect the choice of RE, either pronoun, demonstrative, or descriptive.

(16-b) violates Grice's Maxim of Quantity: do not make your contribution more informative than is required).[11] Thus a good REG component would take this maxim into account when generating REs.

For more discussion and references to other relevant work in REG (including generating references to more than one object), the reader is referred to Krahmer and van Deemter (2012). For the rest of this section, we will put additional focus on two papers. The first because it presents a situation of referring to visually present objects in a situated dialogue setting, and their model takes gaze of the listener into account when producing the REs; the second because it is similar to the RR model presented Chapter 5.

We will first look at Koller et al. (2012). The setting where the REs are produced is the GIVE scenario, as in Figure 3.6. Using the REG model described in Garoufi and Koller (2011), instructions are produced and played to a human participant who controls where in the GIVE scene she is. An eye tracker that tracks the gaze of that participant gives up-to-date information about what she is looking at. The authors extended the REG model to incorporate information from the participant's gaze; for example, if the participant recently heard a generated RE and subsequently looks at the wrong object on the screen, a new RE could be produced (e.g., *no, not that button, the other blue one*) to direct their attention to the originally intended object. They found that incorporating gaze resulted in more task success, and faster success times. Thus incorporating multimodal information, such as from gaze, can help in a situated dialogue scenario. We show in both of our models that gaze can be incorporated to aid in referential success.

The REG model presented in Mast et al. (2014) is similar in spirit to the model presented in Chapter 5. At the heart of the model are properties that an intended referent has. The authors note a criticism of earlier work in REG (e.g., Garoufi and Koller (2011) described above) that properties must be fully discretised before hand, making no use of gradedness (e.g., if two objects are in a room and neither of them are prototypically red–thus neither have the property of being red–but one has a slightly redder hue than the other, an informative RE would be to refer to *the redder one*). The example given in Mast et al. (2014) is size; an object can be considered *big* if it's bigger than other visually present objects, even if it's not a globally-accepted concept of big (e.g., a big mouse is still smaller than a small elephant). Thus their model can take gradedness into account; the properties are not represented as existing or not existing, rather, each object has each property to a certain degree. Though the authors don't explicitly state that their model is grounded, i.e., that the scores of gradedness is learned from data, there is no reason that such information could not be learned.

Their model's aim is to identify the description D that maximises the probability of the

[11] Though see van Deemter et al. (2012); Fernández (2013) which discuss how humans do this regularly.

listener to identify the referent x, with a conditional probability $P(D|x)$. However, this would produce descriptions that are too descriptive (i.e., violation of Grice's maxim of quantity, see above). This must be counterweighted with some kind of measure of description acceptability, which in this case is $P(x|D)$, i.e., the probability that a given description would be accepted–by a human–given the object, which leads to their desired model (weighted with α values):

$$D^* = argmax(1-\alpha)P(x|D) + \alpha P(D|x) \tag{3.4}$$

Where D^* is the chosen description for x from the candidate set. The graded properties are encoded in $P(x|D)$ using Bayes' Rule:

$$P(x|D) = \frac{P(D|x)P(x)}{P(D)} \tag{3.5}$$

where D is a set of tuples that relate the graded features (colours, sizes, locations) to respective feature values for an object in a scene. Each feature of an object is treated as independent, thus the individual descriptions that would be produced by a given feature (e.g., $P(D = red|x = o_1)$ = 0.65) are multiplied together. $P(x)$ is a uniform prior, but the authors hint that it could have some kind of distribution over x using a contextual saliency model. $P(D)$ is the probability that the description D suits a particular object; if the probability D correctly identifying a unique object is low (i.e., D describes more than one object), then $P(D)$ here should be low. We show in Chapter 5 that the model presented there is similar to this one; it can incorporate a prior saliency and properties can be graded.

It is not within the scope of this thesis to determine what constitutes an incremental REG or NLG component (for discussion, see Buschmeier et al. (2012)), but all of the approaches cited in this section attempt to formulate an entire RE before it is presented as speech.

3.4 Reference in other Research

The task of resolving references has been used in a number of research fields. In this section, we look at a brief survey of some of these.

Work has been done in psychological experiments, for example, where it was found that humans negotiate references differently depending on if they "novices" (e.g., someone not familiar with the layout of a city) and experts (e.g., a local resident of a city) (Isaacs and Clark, 1987). It was found that the participants generally assessed expertise (e.g., *do you know*

where X is?), then an expert would supply expertise (e.g., by using specific landmarks that experts know well, such as a reference to a known building), which allows the novice to learn expertise. Tanenhaus and Spivey-Knowlton (1995) showed how the visual context can affect the mental processes of resolving referring expressions. They had subjects follow instructions to manipulate real objects that they could see and touch. By tracking the participants' eye gaze patterns, they found that the visual context influenced spoken word recognition (e.g., if someone says *gr*, but the rest of the word was difficult to hear, it is more likely to have been a prefix for the word *green* rather than *grey* if there were green objects, but no grey objects).

The task of RR has been used in human-machine interaction studies as a way to help robots learn about establishing common ground (see Clark (1996)). Joyce Chai and colleagues have worked in this area. For example, in Lui et al. (2012), they observed that a robot's representation of W is impoverished when compared to a human's representation of W. They modelled a way to overcome these perception mismatches by mediating those differences, e.g., by asking questions about the scene to the human, allowing the robot to improve its representation. This was extended in Chai et al. (2014) with improved methods of representing W. In both cases, mapping W and U to I was done by representing both W and U as relational graphs and merging them; the referent graph which was most matched to the graph for U was the identified referent. Cantrell et al. (2010) also used the task of RR in human-robot interaction with some degree of incrementality and grounding.

We mentioned the work of Luc Steels in the previous chapter, some of which we will now look into further. Though not the overt focus of this thesis, an important part of learning grounded meaning of words is how the world is represented; i.e., what features from the world should be considered (e.g., if colour is an important feature, how should it be represented?). Steels (1996) focused on automatically determining features that should be added to a repertoire of features that are used to distinguish objects from each other. More attuned to the focus of this thesis is the work presented in Spranger and Steels (2012) where resolving reference to visually present objects was used for the acquisition of syntactic structure. To this end, they placed two robots in a scene with coloured blocks and the robots were to communicate about their shared space and the objects in it. As they communicated with each other about their surroundings, they referred to objects (using spatial language) which gave rise to a usable grammar.

The task of RR was also used in Salvi et al. (2012), where the focus was learning grounded meaning of words (e.g., object descriptions, but also actions such as *grasp*, *touch*, etc.) based on audio, visual, and contact sensor information for a robot. The world W is a set of actions, object properties, and effects. The robot had some actions that it could perform (e.g., grasping and moving an object). In order to perform any task, it needed to understand an instruction. Of

interest to this work is how the visual information was represented. The robot could visually perceive a scene using a camera, the feed from the camera was processed using computer vision techniques to extract discrete features about each detected object, such as colour, shape, and size. These were mapped to co-occurring instructions (e.g., *pick up the ball*) using a simple maximum likelihood estimation. A Bayesian Network was used as the model which provided the links between the perceived (and computed) properties and the words. Their model learned what actions to perform when given certain action words, and what objects to perform those actions upon based on object properties (e.g., their model learned that when the word *red* was in the utterance, a red object is more likely to be manipulated). The scenes were quite simple with only a few objects, but their simple approach was able to ground word meanings (albeit from a bag-of-words approach) to visual and contact sensor information. Their model did not work incrementally.

Discussion We mentioned above, but have seen again in the course of this chapter, that referring to visually present objects is a very common task that humans perform, and it is an essential sub-task of any NLU or NLG component, particularly in a dialogue system. Using RR as a task is also a natural task where grounding between language and perception can be learned.

3.5 Related Work on Meaning Representation

In this short section, we look at other approaches to representing meanings of words and how they relate to our task and the background set forth in the previous chapter.

Kelleher et al. (2005) approached RR using perceptually-grounded models, focusing on saliency and discourse context. In Gorniak and Roy (2004), descriptions of objects were used to learn a perceptually-grounded meaning with focus on spatial terms such as *on the left*. Steels and Belpaeme (2005) used neural networks to connect language with colour terms by interacting with humans.

Recent efforts in multimodal distributional semantics have also looked at modelling word meaning based on visual context. Originally, vector space distributional semantics focused words in the context of other words (Turney and Pantel, 2010); recent multimodal approaches also consider low-level features from images. Bruni et al. (2012) and Bruni et al. (2014) for example model word meaning by word and visual context; each modality is represented by a vector, fused by concatenation. Socher et al. (2014) and Kiros et al. (2014) present approaches where words/phrases and images are mapped into the same high-dimensional space. While these approaches similarly provide a link between words and images, they are typically tailored

towards a different setting (the words being descriptions of the whole image, and not utterance intended to perform a function within a visual situation).

Distributional approaches don't necessarily attempt to determine classes and the mechanisms that assign objects to classes, as we set forth to do in the models presented in Chapters 5, and particularly in Chapter 6. Larsson (2015) is closest in spirit to what we are attempting here; he provides a detailed formal semantics for similarly descriptive terms, where parts of the semantics are modelled by a perceptual classifier (as we do in the WAC model in Chapter 6). These approaches had limited lexicons (where we attempt to model all words in our corpus), and do not process incrementally, which we do here.

3.6 Chapter Summary

In this chapter, we reviewed literature relevant to modelling the resolution of referring expressions. We looked at "traditional" (i.e., non-grounded, non-incremental) approaches. We looked at methods that learned grounded word (or other aspects of language) meanings which were learned from data. We also saw some approaches that were incremental; both restart- and update-incremental.

For completeness, we also briefly looked at relevant work in natural language understanding and natural language generation, and how each of those fields make use of reference resolution in order to work properly, with a brief look at how others have approached meaning representation.

4
Data

In the previous chapter, we saw various data and scenes that were used in the explained approaches to reference resolution. In this chapter, we look more closely at the data that we use to evaluate the models described in Chapters 5 and 6, namely the PENTO data in Section 4.1, REX (tangram) in Section 4.2, and ATIS data in Section 4.3. In the final section, we take a closer look at the data and provide some comparisons that will be useful for evaluations of the two models in Chapters 5 and 6.

4.1 Pentomino Puzzle Tiles

The most common data we will see in the experiments uses geometric pentomino puzzle tiles as the visually present objects that are referred. Pentomino tiles are made up of 5 squares in all possible configurations, totaling 12 different shapes.[1] These shapes are depicted in Figure 4.1. Every object resembles (in some cases, with a little bit of imagination) certain alphabet characters, also depicted in the figure. We use these characters to refer to the shapes of the objects in this and future chapters (e.g., T, X, etc.). Note that the tiles depicted in Figure 4.1 can be oriented in different ways in the data (e.g., the T can be turned on its side) and of course

[1]Pentomino tiles are a step up from Tetris shapes, which are made up of 4 squares, for a total of 7 possible shapes.

tile colours and sizes can vary.

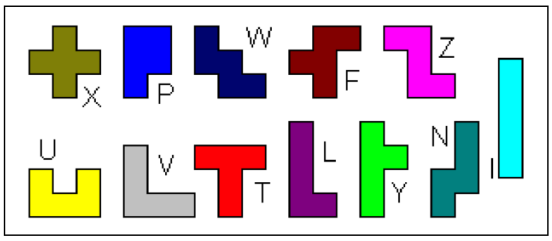


Figure 4.1: All 12 of the possible pentomino tile shapes and corresponding alphabet characters.

Several corpora exist that make use of pentomino tiles. In this thesis, we describe and use three of those. The first we call ACTION where there was a dialogue task to manipulate tiles (e.g., rotate or move them) on a virtual game board and reference to visually present objects was necessary in order to resolve the intended instruction. The second corpus, denoted the TAKE corpus, is a corpus of referring expressions with corresponding scenes (also virtual game boards), as well as recorded and aligned gaze and pointing gesture data. The third corpus, TAKE-CV, uses real pentomino tiles in a non-virtual scene; participants referred to objects only using speech. We will now look at these in greater detail.

4.1.1 ACTION

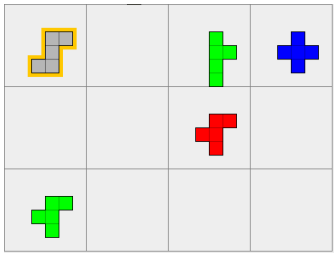


Figure 4.2: Example ACTION Board

The data described in this section has been used for evaluating a model of RR (Heintze et al., 2010), NLU (Kennington and Schlangen, 2012), and an incremental parser (Peldszus et al., 2012). This corpus was collected in a Wizard of Oz study, where the user goal was to instruct the computer to pick up, delete, rotate or mirror puzzle tiles on a rectangular board (as in Figure 4.2), and place them onto another one.[2] For each utterance, the corpus records the state

[2]Wizard of Oz studies consist of a human participant that interacts with an automatic "system". The participant is told that he is interacting with an automatic system. However, unknown to the participant, the system is actually controlled by a human confederate called the "Wizard".

of the game board before the utterance (i.e., each object is assigned an identifier, e.g., o_1 and each object's properties such as their colour, shape, row, and column are recorded), the immediately preceding system action, and the intended interpretation of the utterance, as understood by the Wizard, in the form of a semantic frame specifying action-type and arguments, where those arguments are objects occurring in the description of the state of the board. The language of the corpus is German. The corresponding utterance and frame to the scene in Figure 4.2 are shown in Example (1). We have hand-transcriptions (vocabulary size of 237, avg. of 5.4 words per utterance) as well as ASR transcriptions (using Sphinx with leave-one-out language models; vocabulary size of 261, avg. of 5.1 words per utterance).

(1) a. *drehe die Schlange nach rechts*
b. rotate the snake to the right
c. $\begin{bmatrix} \text{ACTION} & \text{rotate} \\ \text{OBJECT} & o_4 \\ \text{RESULT} & \text{clockwise} \end{bmatrix}$

(2) a. *drehe sie nochmal*
b. rotate it again
c. $\begin{bmatrix} \text{ACTION} & \text{rotate} \\ \text{OBJECT} & o_4 \\ \text{RESULT} & \text{clockwise} \end{bmatrix}$

(2) is an example of an utterance that follows directly after (1). The pronoun *sie* refers to the object (*die Schlange* / *the snake*) in the previous utterance. This is an example of an anaphoric pronoun, however, in this case it is strictly exophoric reference in that the pronoun resolves to an object, not to a linguistic antecedent; about 35% of the REs were of this type, the rest were definite descriptions.

Importantly, the identifier of the referent is recorded in the frame as the OBJECT. A RR model would need to recover the correct object identifier that belongs to the object intended by the speaker's RE, given information about the state of the board and the corresponding RE.

4.1.2 TAKE

The TAKE corpus was first described in Kousidis et al. (2013). In this Wizard of Oz study, the participant was confronted with a PENTO game board containing 15 pieces in random colours, shapes, and positions, where the pieces were grouped in the four corners of a screen, as seen in the example in Figure 4.3.

The participants were seated at a table in front of the screen. Their gaze was then calibrated

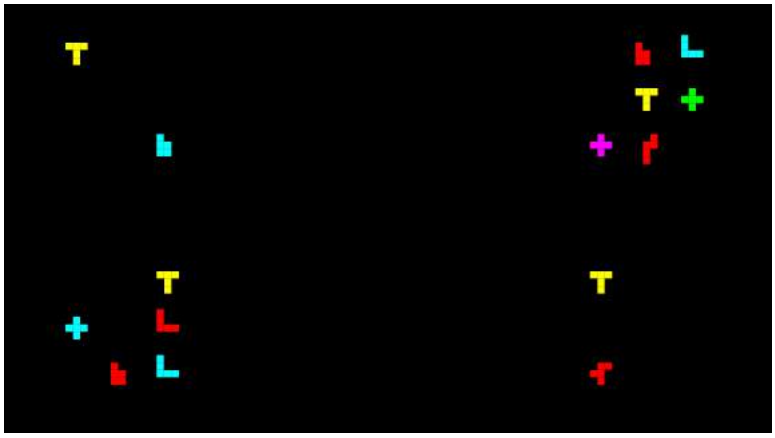


Figure 4.3: Example PENTO board for gaze and deixis experiment; the yellow T in the top-right quadrant is the referred object.

with an eye tracker (*Seeingmachines FaceLab*) placed above the screen and their arm movements (captured by a Microsoft Kinect, also above the screen) were calibrated by pointing to each corner of the screen, then the middle of the screen. See Figure 4.4 to see the environment of the experiment.

They were then given task instructions: (silently) choose a PENTO tile on the screen and then instruct the computer game system to select this piece by describing and pointing to it. When a piece was selected (by the wizard; depicted on the screen as a yellow outline around a tile), the participant had to utter a confirmation (or give negative feedback) and a new board was generated and the process repeated. We denote each of these instances as an *episode*. The utterances, board states, arm movements, and gaze information were recorded in a similar fashion as described in Kousidis et al. (2012). The wizard was instructed to elicit pointing gestures by waiting to select the participant-referred piece by several seconds, unless a pointing action by the participant had already occurred. When the wizard misunderstood, or a technical problem arose, the wizard had an option to flag the episode. In total, 1214 episodes were recorded from 8 participants, all university students. All but one were native speakers; the non-native spoke proficient German. Example (3) shows an episode with original German, English gloss, and referred object identifier, corresponding to Figure 4.3. We have hand-transcriptions (vocabulary size of 383, avg. of 13.3 words per utterance) and ASR output from Google ASR (vocabulary size of 1049, avg. of 6.8 words per utterance).

(3) a. *dann nehmen wir noch das zw- also das zweite t das oben rechts ist ... aus dieser*

gruppe da da möchte ich gern das gelbe t haben ... ja

b. then we take now the se- so the second t that is on the top right ... out of this group there I would like to have the yellow t ... yes

c. $\begin{bmatrix} \text{REFERENT} & o_3 \end{bmatrix}$

Figure 4.4: Depiction of TAKE task; the participant is seated at a table in front of a large-screen which has randomly placed pentomino objects.

The scenes in this corpus are virtual, and hence we can query the scene representation directly about objects and their properties. However, in some experiments, we are also interested in extracting properties of objects automatically using computer vision processing techniques. To get closer to conditions as they would have when working with camera images (e.g., variations of colour due to variations in lighting, distortion of shapes due to camera angles, etc.), we pre-processed these images by shifting the colour spectrum as follows: the hue channel by a random number between -15 and 15 and the saturation and value channels by a random number between -50 and 50. For the object shapes, we apply affine transformations defined by two randomly generated triangles and warp the image using that transform. This generates more complex shapes that retain some notion of their original form. Figure 4.5 shows a game board that has been distorted from its original (Figure 4.3).

Thus each episode has two sets of scene depictions, the original set of scenes (which the participants saw in order to produce their REs) where there is no uncertainty in the objects and their properties (e.g., the fact that o_1 is red is read directly from the scene representation) and a distorted set, where there is uncertainty in the objects and their properties. We describe how the objects and their properties were extracted from the distorted scenes in Chapters 5 and 6 (each model required different information from the distorted scenes).

Figure 4.5: Distorted scene (original in Figure 4.3)

4.1.3 TAKE-CV

This corpus of data was originally described in Kennington and Schlangen (2015). In this Wizard of Oz setting, participants were seated in front of a table with 36 Pentomino puzzle pieces that were randomly placed with some space between them, as shown in Figure 4.6. Above the table was a camera that recorded a video feed of the objects, processed using OpenCV (Pulli et al., 2012) to segment the objects; of those, one (or one pair) was chosen randomly by the experiment software. The video image was presented to the participant on a display placed behind the table, but with the randomly selected piece (or pair of pieces) indicated by an overlay. Figure 4.7 shows a depiction of the setup.

The task of the participant was to refer to that object using only speech, as if identifying it for a friend sitting next to the participant. The wizard (experimenter) had an identical screen depicting the scene but not the selected object. The wizard listened to the participant's RE and clicked on the object she thought was being referred to on her screen. If it was the target object, a tone sounded and a new object was randomly chosen. This constituted a single *episode*. If a wrong object was clicked, a different tone sounded, the episode was flagged, and a new episode began. At varied intervals, the participant was instructed to "shuffle" the board between episodes by moving around the pieces.

The first half of the allotted time constituted phase-1. After phase-1 was complete, instructions for phase-2 were explained: the screen showed the target and also a landmark object, outlined in blue, near the target (again, see Figure 4.6). The participant was to refer to the target using the landmark. (In the instructions, the concepts of landmark and target were explained

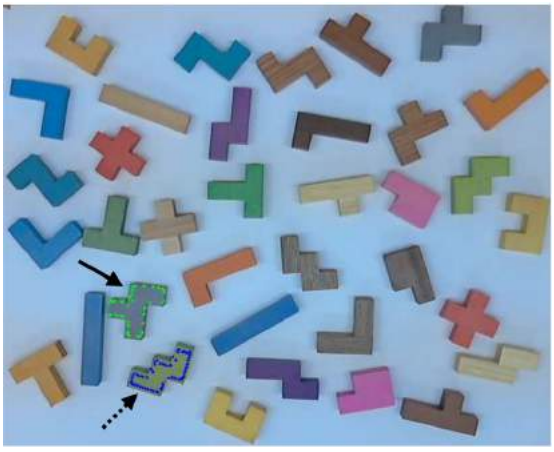


Figure 4.6: Example episode for *phase-2* where the target is outlined in green (solid arrow added here for presentation), the landmark outlined in blue (dashed arrow).

in general terms.) All other instructions remained the same as phase-1. The target's identifier, which was always known beforehand, was always recorded. For phase-2, the landmark's identifier was also recorded.

Nine participants (6 female, 3 male; avg. age of 22) took part in the study; the language of the corpus is German. Phase-1 for one participant and phase-2 for another participant were not used due to misunderstanding and a technical difficulty. This produced a corpus of 870 non-flagged episodes in total. Even though each episode had 36 objects in the scene, all objects were not always recognised by the computer vision processing. On average, 32 objects were recognized.

To obtain transcriptions, we used Google Web Speech (with a word error rate of 0.65, as determined by comparing to a hand transcribed sample) This resulted in 1587 distinct words, with 15.53 words on average per episode. The objects were not manipulated in any way during an episode, so the episode was guaranteed to remain static during a RE and a single image is sufficient to represent the layout of one episode's scene. Each scene was processed using computer vision techniques to obtain low-level features for each (detected) object in the scene which were used for the word classifiers.

We annotated each episode's RE with a simple tagging scheme that segmented the RE into words that directly referred to the target, words that directly referred to the landmark (or multiple landmarks, in some cases) and the relation words. For certain word types, additional information about the word was included in the tag if it described colour, shape, or spatial

Figure 4.7: Depiction of the TAKE-CV task; the participant sat at a table with 36 PENTO objects. A camera (not pictured, though the feed the camera produced can be seen in the screen behind the objects) fed the image information to the computer-vision software which randomly picked which object to select as the referent, and (if applicable) as the landmark. Those selections were displayed on the screen behind the objects.

placement. The *direction* of certain relation words was normalised (e.g., *left-of* should always denote a landmark-target relation). This represents a minimal amount of "syntactic" information needed for the application of the models and the composition of the phrase meanings. An example RE in the original German (as recognised by the ASR), English gloss, and tags for each word is given in (4).

(4) a. *grauer stein über dem grünen m unten links*
b. grey block above the green m bottom left
c. tc ts r l lc ls tf tf

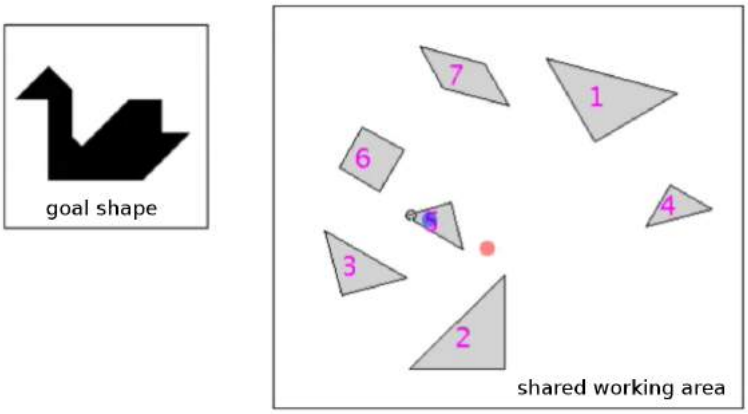


Figure 4.8: Example Tangram Board; the goal shape is the swan in the top left, the shared work area is the large board on the right, the mouse cursor and OP gaze (blue dot) are on object 5, the SV gaze is the red dot. Labels and object identifiers added for presentation.

4.2 The REX Corpora of Tangram Puzzle Dialogues

The REX data consists of 7 multimodal corpora of REs in collaborative problem solving dialogues (Tokunaga et al., 2012). Of particular interest to us is a corpus that used Tangram puzzle tiles consisting of 2 large triangles, 1 medium triangle, 2 small triangles, a small square and 1 small parallelogram; see Figure 4.8. Two human participants worked together to form a goal shape (the corpus had 4 possible shapes, depicted in Figure 4.10) out of the 7 tiles. As shown in Figure 4.9, each participant had a computer screen in front of them and each could see the shared workspace, as labelled in Figure 4.8. The participants each had a defined role: the participant who could see the goal shape on her screen, denoted *solver* (SV), instructed the other participant, denoted *operator* (OP), to manipulate the tiles (i.e., move or rotate) on his screen using the mouse (i.e., the SV was unable to manipulate the tiles, and the OP was unable to see the goal shape. Figure 4.9 represents what the SV could see; the OP could only see the shared working area. The participants were able to communicate freely with each other.

The data were transcribed and the REs were annotated with the ELAN multimodal annotation tool.[3] Also annotated was the object identifier (or set of object identifiers) to which an RE referred. See Figure 4.8; each object is labelled (for presentation; during the task there were no such labels) with an integer representing that object's unique identifier throughout the duration of the dialogue. The mouse pointer (which both participants could see, but only the OP could

[3] http://www.lat-mpi.eu/tools/elan

Figure 4.9: Example of Tangram task experiment setting.

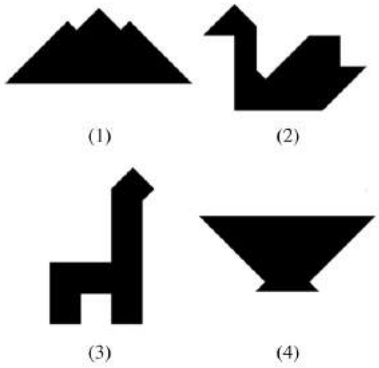


Figure 4.10: Possible goal shapes for the Tangram task.

control) actions were also annotated; the annotations provide information as to which object the mouse cursor is over at any given time, and which object is being moved by mouse actions. The spatial layout of the board was also recorded each time an object was manipulated.

The specific corpus we use is named T2009-11 and was originally described in Iida et al. (2011) (as an extension to Spanger et al. (2012)) where, in addition to the above-mentioned annotations, the gaze of the SV and the gaze of the OP were also recorded and annotations as to which tile OP and which SV was looking at.

This task environment provided frequent use of REs that aimed to distinguish puzzle tiles (and tile groups) from each other. The following are some example REs from the REX corpus:

(5) a. *chicchai sankakkei*
b. small triangle

(6) a. *sono ichiban migi ni shippo ni natte iru sankakkei*
b. that most right tail becoming triangle
'that right-most triangle that is the tail'

Example (5) is a typical example of an RE as found in the corpus. Note that this at the same time constitutes the whole utterance, which hence can be classified as a non-sentential utterance (Schlangen, 2004), which was eluded to in Chapter 2 as something that a model of RR would need to handle. Its transliteration consists of 8 Japanese characters, which could be tokenized into two words. The more difficult RE shown in Example (6) requires the model to learn how spatial placements map to certain descriptions. Because this was a highly interactive

setting, many exophoric pronouns were used, e.g., *sore* and *sono*, both meaning *that*.[4] Pronoun references like this made up around 32% of the utterances.

Importantly, this corpus provides REs that represent the three kinds of REs which we are interested in: descriptions, demonstratives, and pronouns.

4.3 Airline Travel Information System

We also use the *airline travel information system* (ATIS) data (Dahl et al., 1994; Hemphill et al., 1990) as preprocessed by Meza-Ruiz et al. (2008) and He and Young (2005). The data represent telephone conversations, annotated with a semantic frame that acts as an abstraction over an utterance which can then be used for a database query to retrieve information. For each utterance, we have a semantic frame and tag sequence that aligns semantic concepts, example in (7). The only modality for this corpus is speech (i.e., transcribed utterances). There is no visually present object that is to be referred.

We use the ATIS training set, which consists of 4481 utterances between 1 and 46 words in length (avg 11.46; sd 4.34), with a vocabulary of 897 distinct words. There are 3159 distinct frames, 2594 (58%) which occur only once. An additional slot, `goal` represents the overall goal of the utterance.

(7) a. What flights are there arriving in Chicago after 11pm?

b. $$\begin{bmatrix} \text{GOAL} & \text{flight} \\ \text{TOLOC.CITY_NAME} & \text{Chicago} \\ \text{ARRIVE_TIME.TIME_RELATIVE} & \text{after} \\ \text{ARRIVE_TIME.TIME} & \text{11pm} \end{bmatrix}$$

The ATIS data has been used in various NLU papers, some of which were referenced in the previous chapter. It continues to be used in research (though see Tur et al. (2010) for discussion on what else can really be learned from ATIS), and we use it here to demonstrate that at least one of the models described in this thesis can compete with state-of-the-art NLU approaches; which shows that the task of NLU can often be approached with reference resolution techniques.

4.4 A Closer Look at the Data

In this section we take a closer look at the data. We will see how the world W can be represented, some phenomena that occur in the REs, compare some of the data sets, and fine-tune

[4] To be precise, *sono* is a demonstrative adjective.

one important assumption made in Chapter 2 about reference domains.

4.4.1 Objects, Properties, and Features

Before a scene is even presented to a human participant, it has a visual layout. Objects are distinguishable from each other and, in the case of virtual scenes, each object has a pre-defined set of properties that is accessible to the RR component. For example, each PENTO object has a colour and shape. Depending on the task, it can also have a row and a column, or a quadrant (as in the case of TAKE corpus). Other properties could specify whether an object is flipped, or to what degree it is rotated. For each of the tangram objects, there is a shape (colour is the same for all objects, so it isn't particularly useful for our purposes), a size, and a discretised spatial placement where each object is either on the `top`, `center`, or `bottom`, and each object is either on the `left`, `middle`, or `right`. Thus each object has a set of properties that "belong to" it, distinguishing it from other objects. Figure 4.11 shows two objects, their integer object identifiers, and their corresponding properties.

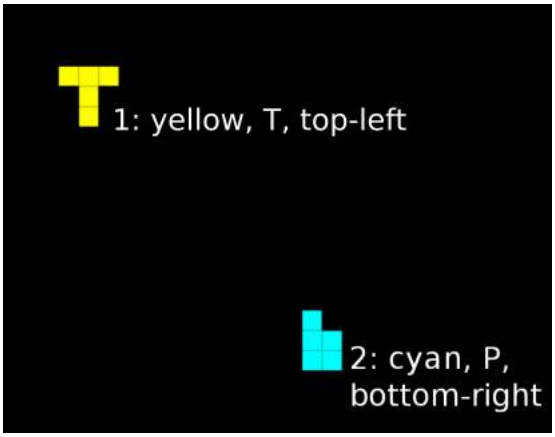


Figure 4.11: Example PENTO board with corresponding properties.

Some properties remain static across an entire dialogue, such as colour or shape, whereas others can change over time. For example, in the tangram data, if an object is moved from the right to the left, it no longer has the `right` property and gains the `left` property as long as it remains on the left. When incorporating additional modalities, such as gaze and pointing gestures, objects can receive properties that correspond to those modalities, such as `gazed-at` or `pointed-at` for the duration of that gaze fixation or pointing gesture, respectively.

We show in Chapter 5 that properties play a key role in the model. Indeed, they *are* the representation of the world W that are used to ground language. The set of properties can be seen as hierarchical (e.g., colours are typed as colour and shapes are typed as shape) or a flat

ontology where all properties are treated equally. We will motivate the latter use in Chapter 5.

The TAKE-CV data is not represented virtually as the other PENTO and tangram data sets; the set of objects and properties is not known beforehand (to the system) but must be computed using computer vision techniques. We will sketch this process briefly: first, the objects are segmented, then each object's low-level features are extracted. We distinguish here between low-level *features* and properties. Properties are as we have explained them: they are pre-defined and symbolic. The features are the low-level computer-vision results, such as an object's RGB value, it's x,y coordinates (treating the scene as a 2-dimensional plane), among others. These features can be used to learn a model that produces a distribution over properties, as would be necessary for the model presented in Chapter 5, or they can be used directly, as done using the model presented in Chapter 6.

Whether we use properties or features to represent objects in a scene, each object does need to be segmented and distinguished from the others. This is necessary for the task of RR. Grounding amounts to finding the function that maps between language and properties, or between language and features.

4.4.2 Comparison of ACTION and REX

The prominent language that we will use to evaluate the two models will have REs in German, but we will also use data that has REs in Japanese. More specifically, the languages and tasks of the PENTO ACTION data and that of REX beg a comparison. German is a head-initial language whereas Japanese is a head-final language, making the REs somewhat different. The task for the PENTO data was also human-machine interaction–i.e., despite being a human "behind the curtain", the human participant thought he was talking to a system. The REX data is a human-human interaction task, where REs are shorter and often ungrammatical and highly dependent on context. Table 4.1 summarises the differences between the PENTO ACTION data and the REX data. We will look at some of these differences in greater detail in Chapter 5. A good model of RR should be able to handle both kinds, given appropriate training data, without changing the framework.

4.4.3 Gaze and Pointing Gestures in TAKE

Kousidis et al. (2013) provided a description of how the gaze and deixis were recorded and analysed for the TAKE corpus. In this section, we explain some of these analyses as this data is used in several experiments in Chapters 5 and 6.

	PENTO	REX
language	**German**	**Japanese**
language type	SVO	SOV
phrase type	head-initial	head-final
avg utt length	7-8	4-5
number of objects	15	7
interactivity	human-wizard	human-human
recorded gaze	SV (speaker)	SV, OP
% of pronoun utts	0%	32%

Table 4.1: Summary of differences between PENTO and REX tasks.

Gaze

As explained above, the TAKE corpus also includes information captured from a Seeingmachines Facelab eye tracker the eye movements on the large screen in front of the participant. More specifically, treating the screen as a 2-dimensional plane, the x and y coordinates were recorded. Note that, given the task, this gaze information comes from the speaker and not the listener, which is commonly what is collected (see, for example, Koller et al. (2012)). We used the I-DT algorithm described in Nyström and Holmqvist (2010) to detect eye fixations. The algorithm can use velocity or a "dispersion" area to determine of individual samples of gaze coordinates belong to a longer fixation. This kind of fixation algorithm is necessary due to limitations of the eye tracker's accuracy. The algorithm yields a duration of the fixation and a centroid coordinate of the area where the fixation took place. The centroid then becomes the point where the fixation is computed to be focused. This process is depicted in Figure 4.12 as superimposed on a TAKE scene; the points represent individual gaze samples, the four large circles represent detected fixations.

Using these computed fixations, we were then able to determine which TAKE objects were looked at, when they were looked at, and for how long. Using 1051 episodes, we found that, after an initial scan of the screen, participants looked at their intended object about 1.5 seconds before the onset of the RE and for about 1 second after the onset of the RE. During the rest of the RE, there appears to be random noise in the fixations as participants look around the screen in rapid succession, presumably taking note of potential distractors. With this information, we can conclude that gaze information can be useful, particularly around the onset of the RE.

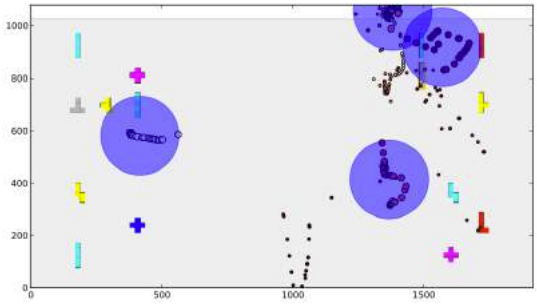


Figure 4.12: Fixation detection using the I-DT algorithm; circles show the dispersion radius threshold.

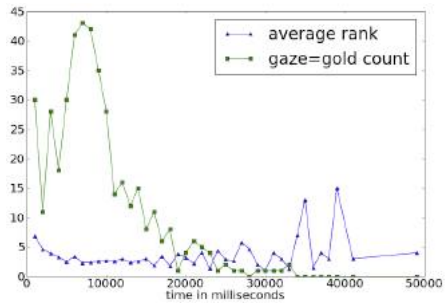


Figure 4.13: Average rank and counts over time for all episodes.

Pointing Gestures

Given the nature of the task, a specific type of pointing gesture was used: the participants that pointed at objects on the screen generally had outstretched arms and used their index fingers to point. This kind of outstretched arm was accurately detected using the Microsoft Kinect. We used the recorded joint information to determine what area of the screen the participants were pointing at.

Following Sumi et al. (2010), we calculated vector information between the hand and the head, as well as between the hand and the shoulder. When the hand is raised above a threshold, the vector is counted as a pointing gesture. An application of this algorithm is shown in Figure 4.14 where the red line (at 0.4) is the threshold. The left hand was used for pointing in the first segment, then the right hand in the second segment. After a couple of seconds, the left hand was again used. Extending these vectors to the screen (i.e., at a measured distance from the joints), treating the screen as a 2-dimensional plane, we were able to determine the coordinates of where the participant was pointing. This coordinate could then be used in experiments and analysis.

We were able to use this information for analysis of how pointing gestures were used. Using 868 episodes (those that had computable pointing gestures; overall participants pointed in 60% of all episodes), we looked at how distractors might play a role in the choice to use pointing gestures. Referred objects in a particular scene could share colour, shape, or spatial area (dividing the scene into four sections) with other distractor objects. The left plot in Figure 4.15 shows that participants did not point more than normal when only one property was shared regardless of how many other distractor objects were present, but pointing increased when two

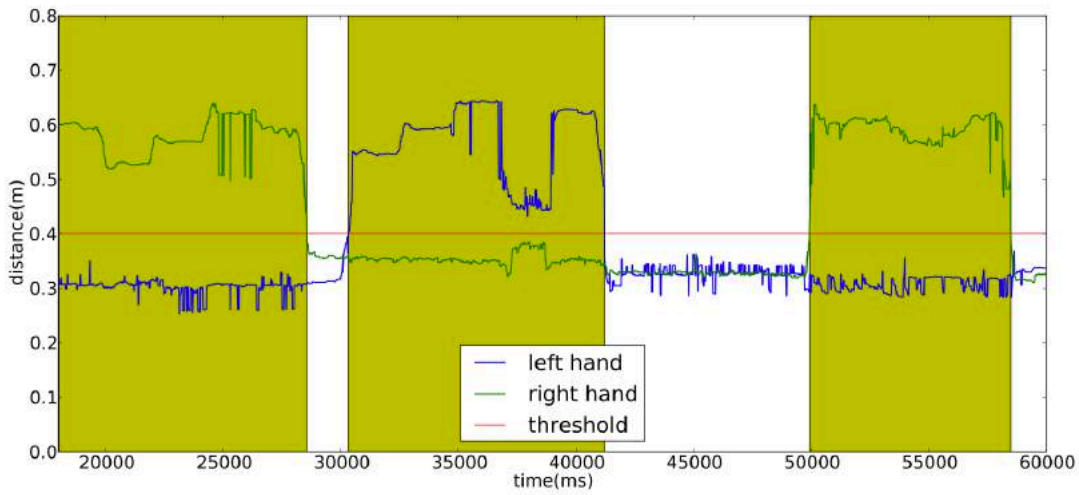


Figure 4.14: Detection of pointing thresholds by distance of left (blue) and right (green) hand from body.

or more properties were shared by the referent and the distractors. The right plot in the figure shows that participants point more when the number of same coloured objects increases, regardless of their position or shape, marking colour as an important visual aspect of the objects. Shape was never considered a distractor property.

4.4.4 Spatial Language

We now move to the kind of language that is used to refer to objects. In this section we put focus on definite descriptions of a specific type: those that use *spatial language*. Spatial language is used when intrinsic features or properties that an object has is not enough to distinguish that object from another one. For example, consider the scene depicted in (8):

(8) a. ●●■
b. the red circle

The corresponding RE in (8-a) is not descriptive enough to distinguish which of the two red circles is being referred. It would take an additional bit of information to make that distinction. This is often where spatial language is used. There are two kinds of spatial language that occur in our selected data, namely *global* spatial language and *relative* spatial language. Corresponding examples of these, with relation to the scene in (8) can be found in (9):

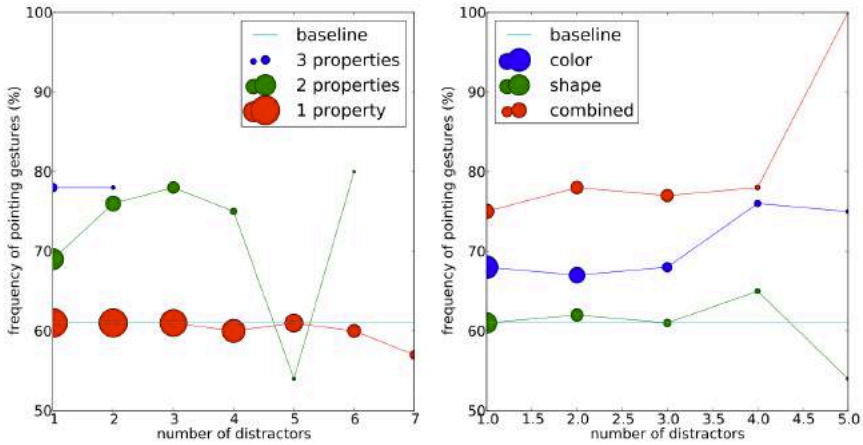


Figure 4.15: Frequency of pointing gestures as a function of the presence of distractors. Dot size denotes the confidence of each point, based on sample size.

(9) a. the red circle on the left
b. the red circle next to the green square

The RE in (9-a) is an example of a RE that uses global spatial language. That is, given a reference frame about a scene, one can distinguish objects from each other by using global information, such as *left*, *right*, *top*, *bottom*, or *middle*. Other global terms might include *near* or *far*, but we won't run into those terms very often in our data. The RE in (9-b) is an example of a RE that uses relative spatial language because it describes one object, the target, as relative to another object, the landmark, and is really made up of two REs, though presumably one cannot be resolved without the other. This is where prepositions such as *left of*, *above*, *below*, *next to*, etc., are used. Relative spatial language is somewhat difficult because one has to resolve first where a landmark is, then how the target object relates to that landmark. This is where REs can become arbitrarily complex, e.g., *the red circle next to the green square above the blue triangle below the*, etc. Spatial language in dialogue has received a fair amount of recent attention. The reader is referred to Coventry et al. (2009) for a concise overview.

4.4.5 Reference Domains

We now expand upon an assumption we made in Chapter 2. Following Funakoshi et al. (2012), we assume a reference domain (Salmon-Alt and Romary, 2001) at each RE. Reference domains are a set of referents implicitly presupposed at each use of a RE. Reference domains were

originally defined to be mental objects having properties such as type, focus, or saliency, and that were equipped with internally structured partitions. A partition is a subset of a reference domain that can be referred in the context of a given RE. For example, given a scene, and a RE prefix *on the top*, a partition of the scene would then only consider objects that fit the description as being on the top. While we don't model the handling of partitions explicitly here, we can treat the objects that are visually present directly before the onset, and during the execution of a RE, as a reference domain.[5]

Each of the Figures in this chapter depicting examples of scenes we use to evaluate our models are each an individual reference domain. For most of our data, the reference domain is unique in that the objects aren't always the same and they fit various arrangements. For example, in the TAKE data, there are always 15 objects and they are always grouped in the four quadrants, but the shapes, colours, and their placement is chosen at random and displayed. This constitutes a single reference domain; the participant can then choose an object and refer to it. In the TAKE-CV data, often the reference domain remains static across multiple episodes, though the object that is being referred is randomly chosen and rarely is the same object chosen more than once in a given scene configuration.

To make the assumption we are trying to convey more concrete, we are assuming that there is a set of objects that are visually present before the onset of a RE, and that particular RE in fact refers to one of those objects. Other things that could be referred, such as a linguistic antecedent or some other abstract concept, are not considered as being referred in our framework.

4.5 Chapter Summary

In this chapter, we looked at the data that is used in subsequent chapters to evaluate the models of reference resolution that are presented. We looked at three sets of PENTO data, REX data, as well as the ATIS corpus. We saw how each of these sets of data were collected and annotated. We also took a closer look at how the scenes are represented as properties or low-level features. We also considered spatial language and the assumption that we use a single reference domain for each referring expression.

[5] One could argue that our approaches to RR are in fact partitioning the reference space incrementally, but it is done probabilistically, thus never excluding any object from being in any partition that might result from part of a RE.

5

The Simple Incremental Update Model: A Generative Model of Incremental Reference Resolution

Simplicity is the final achievement. After one has played a vast quantity of notes and more notes, it is simplicity that emerges as the crowning reward of art.

- Frédéric Chopin

This chapter presents the *Simple Incremental Update Model* (SIUM), a generative model of reference resolution. The model is formulated and implemented to work in an update-incremental fashion, without re-computing previous increments (as explained in Chapter 2). At each increment, a distribution over the candidate visually present objects is computed, where the probability of each object in I represents the degree of belief that it is the referred one.

In the following section, we show how the model is inspired from a IU-network architecture. We then explain the formulation of the model with a toy example of how it works. Central to the model are *properties*, which act as a mediator between the visually present objects and the referring expression. That section is followed by an explanation of a series of experiments

which rigorously evaluate the model. This chapter then concludes with a discussion of the model's strengths and a discussion of its shortcomings.

5.1 Discovering SIUM in the IU-network

In Chapter 2, we saw an explanation of the IU-model of incremental dialogue processing. In this section, we instantiate a theoretical dialogue system built on the IU-model and see how processing a RE can give rise to a specific formulation of how an object is to be resolved. As an overview, in order to resolve a RE, the dialogue system must be able to hear and somehow represent the audio signal of the speaker. This is done with an ASR module that takes in audio via microphone and transcribes it into text. The dialogue system must also be able to perceive the objects that could be referred. This "perception" of a scene is, of course, not the same kind of perception as humans; rather, for this model, it is a symbolic representation of the scene (this is explained in greater detail below). There can also be some kind of semantic processing of the ASR output. Such processing provides an abstraction over the utterance such that the relations between words can be represented. Some examples of this were given in Chapter 2, which we repeat here:

(1) a. *the small red circle*
b. $\iota x(small(x) \wedge red(x) \wedge circle(x))$
c. *the red circle on the left*
d. $\iota x(rex(x) \wedge circle(x) \wedge on_left(x))$

For each of the above examples, there is an entity x that exists in the scene. The role of the RR module is to link that x with that visually present object.[1] In order to do that, there also needs to be some kind of module that represents the scene in a way that is accessible to the RR module (much like how ASR made the acoustic signal accessible by representing it as a series of words–i.e., some kind of transduction). We call this the *visual* module. The visual module represents the scene symbolically in that we assume that objects are pre-segmented and that their properties are known without uncertainty. For example, the scene in (2) below has three objects. Each can be identified (reading from left to right) as o_1 with properties `red`, `circle`, `left`; o_2 has the properties `red`, `circle`, `middle`; and o_3 has the properties `green`, `square`, `right`. This information is what is passed on to the RR module as perceptual information. The task of the RR module is to determine the identifier (o_1, o_2, or o_3) of the

[1] Though note that in the experiments below, in many cases we do not apply any kind of semantic abstraction; rather, given our assumptions on uniqueness and existence, that there is an entity x that needs to be resolved to an object and the words or ngrams can be used.

object that is being referred to by the RE.

(2) a. ●●■

The dialogue system as explained is depicted in Figure 5.1 using a slightly more complicated scene. We can begin to see that, during the resolution process, the RR module depends on the output from the semantics and visual modules, the semantic module depends on the output from the ASR module, and both the ASR module and the visual module are what "perceive" the RE and the scene, respectively. Following the IU-module architecture more closely, we assume that IUs produced by some module A are the result of IUs produced by modules that fed into module A. This is where the *grounded in* (GRIN) and *same-level links* (SLL) form a IU-network (as opposed to the network of IU-modules as seen in Figure 5.1). Figure 5.2 shows an example of this for the RE *the red cross on the top left* for a scene of 5 objects depicted in the figure.

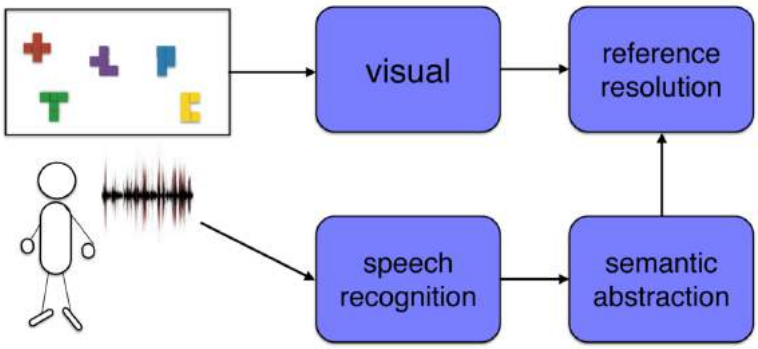


Figure 5.1: A speaker produces a RE that is intended to refer to an object in the scene. The RE is transcribed by the speech recogniser module and that output is passed to a semantic abstraction module. The visual module represents the scene and both the visual and semantic module pass their representations onto the reference resolution module.

From Figure 5.2, it is shown that, when resolving a reference, aspects of the scene need to be linked with aspects of the RE. For the example RE and scene depicted in Figure 5.2, we can see that resolving the RE to o_2 depends on the words (via the semantics), and that words "pick out" properties that the intended object has; i.e., the word *red* picks out the property of o_2's redness, the word *cross* picks out its cross-ness, etc. Those words are specifically chosen to distinguish that object from the others based on its visual features, resulting in the resolution of the entity x as o_2. Thus following the arrows from the RR level, we see that words in the SEM and ASR levels are based on the features that are in the *visual* level. With this, a generative story begins to emerge: when a speaker intends to refer to an object, she ranges over the properties which she perceives that object to have, and she then utters words in a RE that pick out those

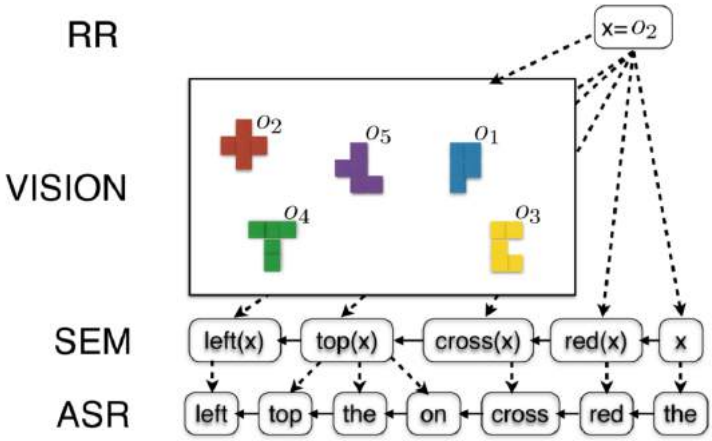


Figure 5.2: Example IU-network for the RE *the red cross on the top left*; the RR module grounds-into visual and semantic IUs, the semantic IUs ground into words from ASR. Dashed arrows represent GRIN relations and solid arrows represent SLL relations.

chosen properties.

This stance on resolving RE comes from the side of the speaker, even though we are ultimately interested in resolving REs on the side of the listener. Such a stance assumes that a speaker would utter a RE such that it is resolvable by a listener. More specifically, the speaker would observe properties that the object has and utter the words that correspond to those properties. This is somewhat reversed from the general function of RR presented in Chapter 2 (repeated below in 5.1 for convenience):

$$I^* = \underset{I}{\operatorname{argmax}}\, P(I|U, W) \tag{5.1}$$

Which, given a RE U and the state of the world W (i.e., the scene), 5.1 yields the identifier I of an object in W. The generative model that we describe in the next section is directly derived from this function, but factored and made incremental in order to match the intuitions from the IU-network as explained in this section.

5.2 Model Definition

In this section, we explain the generative model of RR, how it is derived generally and then how it can be made to work incrementally. We then give some toy examples of it in use, and give some open questions about the model–what we expect it to be able to do, and what we do not expect it to be able to do.

5.2.1 General Derivation

To make (5.1) generative, we apply Bayes' Rule:

$$P(I|U,W) = \frac{P(U|I,W)P(I|W)}{P(U|W)} \tag{5.2}$$

In order to arrive at an indented object using the formulation in (5.2), one must maintain a language model for all possible intentions and all possible world configurations that could exist. This clearly isn't feasible. To simplify the problem, we can make some assumptions. First, we can assume that the words in U are uttered precisely to identify the intended object, even if U doesn't name the object directly. Therefore, here we insert R, a mediating variable between U and I with the assumption that R will represent more directly what is uttered in U while also maintaining a direct connection to the intended object:

$$P(I|U,W) = \sum_{r \in R} \frac{P(U|R=r)P(R=r|I,W)P(I|W)}{P(U|W)} \tag{5.3}$$

In fact, R represents *properties* that objects are defined to have which generally map to words in REs. We can compute $P(R|I,W)$ easily by reading off properties of the objects in W. We assume that the joint distributions $P(I|W)$ and $P(U|W)$ are identical to the single distributions $P(I)$ and $P(U)$ respectively, by assuming that W influences neither I nor U. With these simplifications, we can rewrite (5.3) as follows:

$$P(I|U,W) = \frac{1}{P(U)} P(I) \sum_{r \in R} P_w(U|R=r)P(R=r|I) \tag{5.4}$$

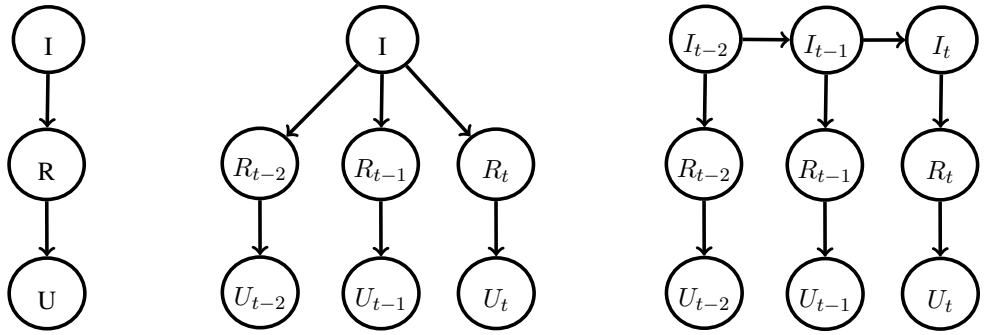


Figure 5.3: Graphical representations of model: general model (left) shows how what is uttered (U) is dependent on properties (R) that an object which is intended to be described (I) has. The centre version shows how this unfolds in a restart-incremental fashion; the intention is constant. The update-incremental version (right) does not need to recompute what has already been computed in previous steps.

(We can move $P(U)$ and $P(I)$ out of the summation as they do not depend on R.) That is, we assume that R is only conditioned on I, and U is only conditioned on R. The generative story follows from the IU-network as described above: a speaker intends to refer to an object I, then selects properties R that she perceives the intended object to have, and utters the RE U to "pick out" those properties such that a listener can identify I. The leftmost figure in Figure 5.3 represents this graphically.

5.2.2 Deriving an Incremental Model

What the leftmost figure in Figure 5.3 shows is a model based on an entire RE, but what we want is a model that works incrementally, word for word. That is, given an intended object I that is constant across an entire RE U, we want to model the contribution of each word and, as mentioned above, we assume that the properties in R correspond more directly to words in a RE. This is represented graphically in the centre figure of Figure 5.3 which spans 3 words.

This of necessity alters the formulation in (5.4), where I is conditioned on a joint distribution of all other variables with a corresponding U_k and R_k for each word in a RE. Clearly, this isn't quite what we want as it would require a different formulation for each length of RE. On a more practical level, as a reference resolution component implemented to reflect such a model processes word by word, it would of necessity compute the entire RE until that point, i.e., it would recompute parts that have already been computed in a previous increment. This is the restart-incremental variant of incremental processing as explained in Chapter 2. An update-

incremental model has the pleasing result of saving processing time during resolution, but we show presently that it allows for a simpler derivation (given some additional assumptions) than the restart-incremental version would have. We show this for a two-word RE but it is trivial to show that it works for REs of all lengths. In the update-incremental model, we treat I as different variables at each increment, where I in the current step is dependent on all other variables in the current step and the previous step (i.e., word increment):

$$P(I_2|I_1, U_1, U_2, R_1, R_2) = \frac{P(I_1, I_2, U_1, U_2, R_1, R_2)}{P(I_1, U_1, U_2, R_1, R_2)} \tag{5.5}$$

Which can be factored in a similar way as (5.4), marginalizing out R_1 and R_2:

$$P(I_2|I_1, U_1, U_2) = P(I_2|I_1)P(I_1) \sum_{r_2 \in R_2} \frac{P(U_2|R_2)P(R_2|I_2)}{P(U_2)} \sum_{r_1 \in R_1} \frac{P(U_1|R_1)P(R_1|I_1)}{P(U_1)} \tag{5.6}$$

With this, we can further apply a trick to marginalise over I_1, but explicitly define $P(I_2|I_1)$ as an enforcement on object identity; i.e., we define it as a function that is explicitly set to zero when I_1 does not equal I_2. This is similar to a simple Bayesian update, making the need to range over all possible combinations unnecessary, which brings us the complexity savings (and hence computation savings) that we are looking for. Moreover, the rightmost summation in (5.6) over R_1 is precisely the computation that occurred in the previous incremental step (i.e., the first word of this example two-word RE), which is a distribution over I. We therefore treat $P(I_1)$ as that distribution, effectively making it a prior probability that is set to the posterior of the previous step. We further drop $P(U_k)$ by assuming that all words are equally likely to be uttered. By applying these simplifications, we arrive at the model we are looking for, applied at each word increment, which fits the intuitions of the theoretical IU-network described above:

$$P(I|U) = P(I) \sum_{r \in R} P(U|R = r)P(R = r|I) \tag{5.7}$$

It can be trivially shown that (5.6), equivalently (5.7), can be applied to REs of all lengths; where each R_k is marginalised as well as all I_k, forcing identity at each step. This results in the graphical model portrayed on the right of Figure 5.3, unrolled over 3 words. Though I_k is not the same across increments, forcing identity effectively makes it remain constant across the RE with $P(I)$ keeping track of what has already been computed at earlier increments.

The model is composed of several sub-models, $P(I)$, $P(U|R)$, and $P(R|I)$. When implemented in a component that resolves references, each sub-model performs a specific task and, in some instances, the models are learned from data. We explain these sub-models in greater detail in the remainder of this section. That is then followed by some toy examples. First, however, we explain further what we mean by the properties in R, which play an important role in our model.

Properties in R

Properties in our model can be visual properties such as colour (e.g., red or green), shape (e.g., cross or v-shaped), or spatial placement (`left-of`, `below`, etc.). The purpose of the properties is to ground objects with language in a more fine-grained way than with the object itself. This is an intuitive observation, in that the words people use to refer to objects (particularly visible objects) are visual properties such as shape, spatial arrangement, etc. As is shown in the experiments, the choice of properties is crucial to the success of the model. One can perhaps see the set of properties as a flat (or at least very shallow) ontology; properties of all types are treated as equals.

The properties can also be where additional modalities are incorporated into the model. For example, in a scene where a speaker is pointing at an object and saying the word *that*, the object being pointed at could have a `pointed-at` property; the model would then learn that pointed-at grounds to demonstrative words, such as *that*. We explore this in Experiment 3 below.

Linking Objects and Properties: $\mathbf{P(R|I)}$

The sub-model $P(R|I)$ provides the link between objects and the properties that those objects have. Here we follow, to our knowledge, a novel approach, by deriving this distribution directly from the scene representation. We assume that with equal probability one of the properties that the intended object actually has is picked to be verbalised, leaving zero probability for the ones that it does not have.[2] This in a way is a rationality assumption that we eluded to above: a rational speaker will, if at all, mention properties that are realised and not others (at least in non-negative contexts).

As eluded to above, there may be cases where the property to be uttered isn't quite clear. For this reason, $P(R|I)$ can also have uncertainty in its representation by maintaining a distribution over properties, where the probability that a property has represents the degree of belief that

[2]Certainly, this is a rather naive assumption as certain properties could be more salient, or allow the object to be easier uniquely identified, but this formulation works well in practice.

it belongs to that particular object. It isn't completely intuitive for a generative model such as this to consider uncertainty in the scene, given the way it has been formulated. Uncertainty in colours, for example, could mean two things: 1) that there is an actual problem on the side of the speaker with the perception of the properties, or 2) the speaker recognises that the properties to be uttered are not completely prototypical to what might be understood by the listener, so uncertainty is implicit in the distribution over colours in how the utterance is expressed (e.g., *the red one* vs. *the reddish one*). Option 2 makes more sense here; making a RE that signals to the listener that there might be some uncertainty in the colours allows them to be more accommodating in how the distribution is spread away from the prototypical colour that was uttered. In either case, the rationality assumption of the speaker should hold, i.e., that certain properties are picked out that the object does have, but if the speaker recognises a problem in perception either on her side or the side of the listener, accommodations can be made in the way the RE is produced.

Besides uncertainty, $P(R|I)$ could also encode saliency information in the distribution over properties, giving some objects a higher probability of being the referred one (in which case, $P(R)$ in the derivation would not be uniform and would be better left in the model). Whether $P(R|I)$ encodes uncertainty about a scene or saliency in properties is a design decision. In Experiment 4, we examine how the model performs when varying uncertainty.

Linking Language and Properties: $\mathbf{P(U|R)}$

The sub-model $P(R|U)$ represents the grounded mapping between properties and language; i.e., aspects of REs that can be used to pick out those properties. More semantically, this model can be seen as a function from a linguistic element, such as a word, to a semantic concept (e.g., the word *red* maps to the concept of *redness* represented in this model by a corresponding property) where the set of properties represents the set of semantic concepts that words can map to.

This is one point where our model departs from most previous work (as presented in Chapter 3) in that the mapping between words and concepts is not pre-defined by rules. Rather, $P(U|R)$ can be learned directly from data by (smoothed) Maximum Likelihood estimation. For training, we assume that the property R that is picked out for verbalisation is actually observable. In our data, we know which properties the referent actually has, and so we can simply count how often a word (or its derived semantic representation) co-occurred with a given property, out of all cases where that property was present.

For the experiments described below, we make a technical modification to the model by applying Bayes' Rule to $P(U|R)$:

$$P(I|U) = P(I) \sum_{r \in R} P(R = r|U) P(R = r|I) \tag{5.8}$$

which cancels $P(U)$ (before it was dropped from (5.7)) and introduces $P(R)$ into the summation, but $P(R)$ can be dropped since (in this work) it can be approximated with a uniform distribution. This is motivated by the assumption that $P(R|U)$ is easier to learn using standard available classifiers (with R as class labels; another approach, which we do not explore here, would be to train $P(U|R)$ as a family of language models). The formulation in (5.8) represents the model that we use in the experiments below.

Contextual Prior: $\mathbf{P(I)}$

The sub-model $P(I)$ acts as a prior in our model and provides a way of keeping track of the distribution over I as the RE incrementally unfolds. At the beginning of the computation for an incoming RE, we set the prior $P(I)$ to a uniform distribution (or, it can be used to encode initial expectations about intentions; i.e., prior gaze information). For later words, it is set to the *posteriori* of the previous step, and so this constitutes a Bayesian updating of belief (as explained above, with a trivial, constant transition model that equates $P(I_{t-1})$ and $P(I_t)$).[3]

5.2.3 Examples

Example without Uncertainty in $P(R|I)$

This example task is reference resolution in a shared visual context: there is an intention to refer to a visible object. An object reference is made by applying the data in Table 5.1 to our model as formulated in Equation 5.8, taking the highest-ranked object in the resulting distribution as the referent. For this example, assume that there are two objects `obj1` and `obj2`, and four properties to describe those objects, `red`, `round`, `square` and `green`. `obj1` happens to be a red ball, with properties `red` and `round` (●); `obj2` is a red box, with the properties `red` and `square` (■). The utterance for which we want to track a distribution over possible referents, going word-by-word, is *the red ball*.

We now need the models $P(R|U)$ and $P(R|I)$. We assume the former is learned from data, and for the four properties and three words gives us results as shown in Table 5.1-A (that is, $P(U = \textit{the}|R = \texttt{red}) = 0.03$). The model $P(R|I)$ can be read off the representation of the scene: if you intend to refer to object `obj1` ($I = \texttt{obj1}$), you can either pick the property `red`

[3] In that sense, our incremental understanding could be called "intra-sentential belief tracking," in analogy to the current effort to track system belief about user intentions across turns (Williams, 2010; Ma et al., 2012).

A: $P(R|I)$

word	**red**	**round**	**square**	**green**
the	0.03	0.02	0.02	0.02
red	0.82	0.009	0.09	0.01
ball	0.02	0.9	0.02	0.07

B: $P(R|U)$

int.	**red**	**round**	**square**	**green**
`obj1`	0.5	0.5	0	0
`obj2`	0.5	0	0.5	0

C: $\mathbf{P(I)} \sum_{\mathbf{r} \in \mathbf{R}} \mathbf{P(R = r|U)P(R = r|I)}$

I	U	**red**	**round**	**square**	Σ	$P(I\|U)$
`obj1`	*the*	.015	.01	0	.025	.05
`obj2`		.015	0	.01	.025	.05
`obj1`	*red*	.41	.0045	0	.41	.47
`obj2`		.41	0	.045	.46	.54
`obj1`	*ball*	.01	.45	0	.46	.96
`obj2`		.01	0	.01	.02	.04

Table 5.1: Application of utterance *the red ball* using the model, where `obj1` is the referent. I represents a distribution over the two objects, R is the set of properties which those objects have.

or the property `round`, so both get a probability of 0.5 and all others 0; similarly for `obj2` and `red` and `square`.

Table 5.1-C also shows an application of the full model to our example utterance. The cells in the columns labelled with properties show $P(R|U)P(R|I)$ for the appropriate properties and intentions (objects), the column Σ shows results after marginalizing over R. The final column then factors in $P(I)$ with a uniform prior for the first word, and the respective previous distribution for all others, and normalises.

As these numbers show, the model behaves as expected: up until *red*, the utterance does not give enough information to decide for either object as the two probabilities are roughly equal, but once *ball* is uttered `obj1` is the clear winner. This example illustrates how the model works in principle and showed that it yields the expected results in a simple toy domain.

Example with Uncertainty in $P(R|I)$

The above example represents a scene where the properties have no uncertainty. However, uncertainty about the scene can be represented in $P(R|I)$ by producing a distribution over properties for each I. Compare, for example, Table 5.1-B with Table 5.2 where there is an example distribution over a set of properties (i.e., all properties would sum to 1). The resolution

$P(U|R)$ **with uncertainty**

int.	**red**	**green**	**square**	**round**
`obj1`	0.4	0.04	0.05	0.41
`obj2`	0.4	0.1	0.35	0.15

Table 5.2: $P(R|I)$ when using properties with uncertainty, normalised for type (e.g., colour and shape).

procedure using SIUM would proceed in the same way as it did in the above example. However, being a generative model and, given our generative story, an intention (object) in I is selected, properties in R that belong to that object are then identified, and the corresponding utterance U which conveys those properties in an acceptable, grammatical utterance is then produced (barring the fact that certain properties might be chosen over others to be more distinguishing from other objects).

5.2.4 Open Questions about the Model

Here we set forth several questions about SIUM as a model of grounded incremental RR. In each case, an experiment is cited which addresses that particular question.

- *How can grounding occur?* Experiment 1 shows that the mapping between U and R can be learned by various means; we explore using co-occurrence counting, a Naive Bayes classifier, and a maximum entropy classifier.
- *How cam the* RE *be represented?* Experiment 1 shows that U can be represented by just single words, ngrams, or by semantic abstractions over the unfolding referring expressions.
- *Can the model accommodate uncertainty in the perception of the world?* As explained in the previous section, $P(R|I)$ can represent scenes with or without uncertainty in the properties. This is shown practically in Experiment 4.
- *Can the model take contextual saliency into account?* Experiment 3 shows that an initial distribution over the objects that is based on contextual saliency (which can also be learned using SIUM) can be used as an initial contextual prior.
- *Is the model robust to noise in the representation of the* RE*?* As is shown in Experiments 1, 2, and 4, SIUM is robust to ASR transcriptions as well as hand-transcribed referring expressions.

- *Is the model update-incremental in that it processes word-by-word without re-computing previous steps?* It is shown in all experiments that SIUM updates its distribution at each word increment where the information that was processed in a previous steps is not re-processed in the current step, as explained in Chapter 2.

- *Can the model take gaze and deixis (i.e., pointing gestures) into account?* SIUM can incorporate additional modalities (here, gaze and deixis) in two ways:

  - Experiment 2 shows that additional modalities can be incorporated into SIUM by treating them as individual models of reference resolution that produce distributions over the objects; a final interpolated model is then the final fusion of the three modalities.

  - It was shown in Kennington et al. (2013) that an object that is being pointed at can receive a property that represents that fact, e.g., a `pointed-at` property for the duration of that pointing gesture. We further show in Experiment 3 that mouse-pointing "gestures" can be represented in the same way. The same was also done for gaze properties; e.g., the closest object with a fixation received the `gazed-at` property for the duration of the fixation, thus words are grounded to these properties as they represent modalities in real-time.

- *Can the model handle more than one kind of* RE*?* Similar to the claims made in Funakoshi et al. (2012), SIUM is a "unified model" in that it can handle definite descriptions, handle deictic gestures, and handle exophoric pronouns without changing the model formulation, in fulfilment of some of the goals of this thesis.

- *Can the model work with my language?* The model has been tested on German (Experiments 1 and 2), Japanese (Experiment 3), and English (Experiment 5) data, giving us reason to believe that (given data) it can be made to work on a variety of languages.

The model is not designed to handle referring expressions with relations, e.g., *the red one next to the green book* (there were only several cases where this occurred in our experiment data, explained below), nor can it handle negations as it is formulated. The scope of the referring expression is assumed to be processed by another component. We leave addressing these issues to future work.

We explain several experiments in the sections that follow. Each explains the part of SIUM that is being varied or tested, the data that is used, the metrics for evaluation, and the results. Each experiment has remarks about that particular experiment; general discussion is left until the Discussion in Section 5.8.

5.3 Experiment 1: Varying Representation and Grounding of Referring Expressions

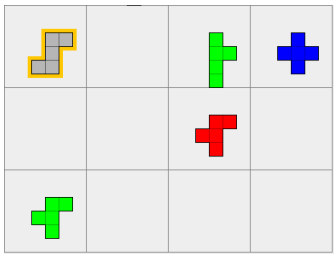


Figure 5.4: Example Pentomino Board

Following Kennington et al. (2013, 2014c), The goal of this experiment is to establish that the model works under basic conditions: there is a representation of U and of W and the model learns a mapping between them. We vary how U is represented (by the words of the RE and by a semantic abstraction over the RE) and how the sub-model $P(U|R)$ is computed. $P(R|I)$ will not have uncertainty as the objects are from virtual data.

5.3.1 Data

This section uses the ACTION data as explained in Chapter 4. An example of two semantic frames and and corresponding utterances are repeated below, corresponding to the scene in Figure 5.4.

(3) a. *drehe die Schlange nach rechts*
b. rotate the snake to the right
c. $\begin{bmatrix} \text{ACTION} & \text{rotate} \\ \text{OBJECT} & \text{object-4} \\ \text{RESULT} & \text{clockwise} \end{bmatrix}$

(4) a. *drehe sie nochmal*
b. rotate it again
c. $\begin{bmatrix} \text{ACTION} & \text{rotate} \\ \text{OBJECT} & \text{object-4} \\ \text{RESULT} & \text{clockwise} \end{bmatrix}$

(4) is an example of an utterance that follows directly after (3). The pronoun *sie* refers to the object (*die Schlange* / *the snake*) in the previous utterance. This is an example of anaphora

resolution (however, in this case it is strictly exophoric reference in that the pronoun resolves to an object, not to a linguistic antecedent; about 35% of the REs were of this type, the rest were definite descriptions); the model needs to accommodate this kind of discourse context. How this is done will be explained below.

5.3.2 Task & Procedure

The task is RR. At each increment, given a representation of the utterance U and a representation about the state of the world W, SIUM returns a distribution over all objects; the probability for each object represents the strength of the belief that it is the referred one. The argmax of the distribution is chosen as the hypothesised referent. In this experiment, we are only interested in the OBJECT slot, the other two slots don't refer to visually present objects (however, as was shown in Kennington et al. (2013), each slot can be treated as a RR task).

All results were obtained by averaging the results of a 10-fold validation on 1500 Pento boards (i.e., utterances+context, as in Kennington and Schlangen (2012)). We used a separate set of 168 boards for small-scale, held-out experiments. We compare three model variants using hand-transcribed utterances and ASR utterances as well as under varied conditions of U, namely when U simply represents the current word in the RE and $P(R|U)$ is computed by simple maximum likelihood (i.e., co-occurrence counts as in the example above). This model variant will be referred to as the *unigram* variant. Another variant will use *ngrams* (unigrams, bigrams, and trigrams) for U and $P(R|U)$ will be computed with a maximum entropy classifier. The final variant will also compute $P(R|U)$ with a maximum entropy classifier, but it use a semantic representation for U with output from a parser that produces a *Robust Minimal Recursion Semantics* (RMRS) semantic representation (Copestake, 2007). This will be referred to as the RMRS variant. Such a representation provides our model with a structured way to abstract over the surface forms. We will first give a brief explanation of the RMRS framework, then describe how it is different from using ngrams for U.

5.3.3 Abstracting over U: Robust Minimal Recursion Semantics

RMRS is a framework for representing semantics that factors a logical form into *elementary predicates* (EP). For example in Table 5.3, the first row represents the first word of an utterance, *take*, and the corresponding RMRS representation; the EPs *take* and *addressee* are produced. The EPs in this example have *anchor* variables and in most cases, an EP has an argument *entity*. Relations between EPs can be expressed via *argument relations*, e.g., for *take* in the table, there is an `ARG1` relation, denoting *addressee* as the first argument of the predicate *take*. Other relations include `ARG2` and `BV` (relating determiners to the words they modify). A full example

of an utterance and corresponding RMRS representation can be found in Table 5.3, where each row in the word column makes up the words of the example utterance.

In this experiment we are interested in processing utterances incrementally. As argued in Peldszus et al. (2012), RMRS is amenable to incremental processing by allowing for *under-specification* in how relations are represented. Table 5.3 has an example of an underspecified relation: when the second word *the* is uttered, the RMRS segment predicts that the entity represented by `x14` will be the `ARG2` relation of the EP for *take*, but the actual word that produces the EP that has `x14` as an argument has not yet been uttered. Each row in the table represents what we would want an RMRS parser to produce for our model at each word increment.

word	**RMRS segment**
take	$a7 : addressee(x8), a1 : take(e2), ARG1(a1, x8)$
the	$a13 : def(), ARG2(a1, x14), BV(a13, x14)$
red	$a33 : red(e34), ARG1(a33, x14)$
cross	$a19 : cross(x14)$
next to	$a49 : next(e50), ARG1(a49, x14), ARG2(a49, x53)$
the	$a52 : def(), BV(a52, x53)$
blue	$a72 : blue(e73), ARG1(a72, x53)$
piece	$a58 : piece(x53)$

Table 5.3: Example RMRS representation for the utterance *take the red cross next to the blue piece*. Each row represents an increment of the utterance.

A more detailed explanation of RMRS can be found in Copestake (2007). There are two key differences between the RMRS model and that of the ngram and unigram models: first, as stated above, U is represented by a RMRS semantic abstraction and the sub-model $P(U|R)$ learns the mapping between RMRS and the world. For example, in Figure 5.3, U can either be represented by words (i.e., the **word** column), or by the RMRS representation (**RMRS segment** column). Second, instead of applying the model at every word, the RMRS model can tell us *when* to apply the model; RMRS produces entities, and entities of a certain type (specifically for RMRS, type x) are to resolve to objects in the real world. We only need to apply the model for the words that are within the scope of that entity. For example, by the time *red* is uttered in Figure 5.3, the processing for entities `x8`, `e2`, and `e12` is complete, but the processing for `x14` is under way, and active as long as `x14` is referenced as an entity in the RMRS increment.

The following properties in R were used: colour (e.g., `red`, `blue`, etc.), shape (all 12 shapes can be represented by a similar-looking character, e.g., X, T, etc., as explained in Chapter 4), row, column, and an additional *selected* property. Objects on the game board that had a visual outline around them (e.g., the object in the top-left in Figure 5.4) had this property. Only one object on the board at a time could have this property. When selected, participants

often used pronouns to refer to this object. Using this property, the model can learn to ground pronouns to objects without needing to describe them again (matching the $recently_referred$ predicate in Chapter 2).

5.3.4 Metrics

To give a picture of the overall performance of the model, we report **accuracy** (how often was the argmax the gold, referred target) and **mean reciprocal rank** (MRR) of the gold target in the distribution over all the objects (like accuracy, higher MRR values are better; values range between 0 and 1). The use of MRR is motivated by the assumption that in general, a good rank for the correct object is desirable, even if it doesn't reach the first position, as when integrated in a dialogue system, this information might still be useful to formulate clarification questions.

To test how well our model worked incrementally, we followed previously-used metrics for evaluation (Schlangen et al., 2009; Kennington et al., 2013):
first correct: how deep into the RE does the model predict the referent for the first time?
first final: how deep into the RE does the model predict the correct referent and keep that decision until the end?
edit overhead: how often did the model unnecessarily change its prediction (the only *necessary* prediction happens when it first makes a correct prediction)?

5.3.5 Results

Utterance-level Results

The results in Figures 5.5 and 5.6 show what would be expected: a simple maximum likelihood model that uses the current word provides a good baseline, but is improved upon when considering ngrams and using maxent to compute $P(R|U)$. Things further improve when U is represented by RMRS, which provides additional structure. These are welcome results, the model is already robust under simple assumptions, but can be improved upon when resources such as a maxent classifier or a RMRS parser exist that can be used on the data. We assume from these results that the mapping between U and R was learned well, by both co-occurrence and by the maximum entropy classifier. For example, the German word *rot* (red) was learned to map well to the property `red`, and pronoun words such as *sie* or *es* mapped to the `selected` property. Indeed, many of the REs that contained pronouns were correctly resolved.

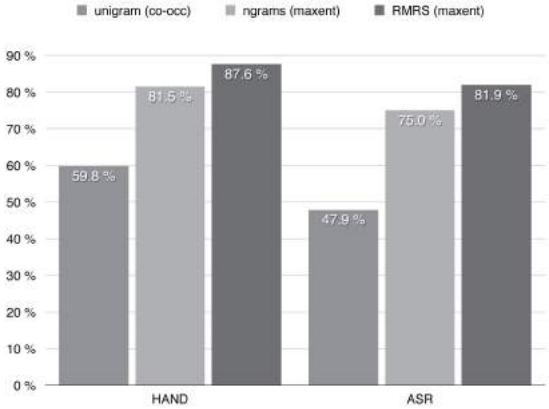


Figure 5.5: Accuracies for Experiment 1.

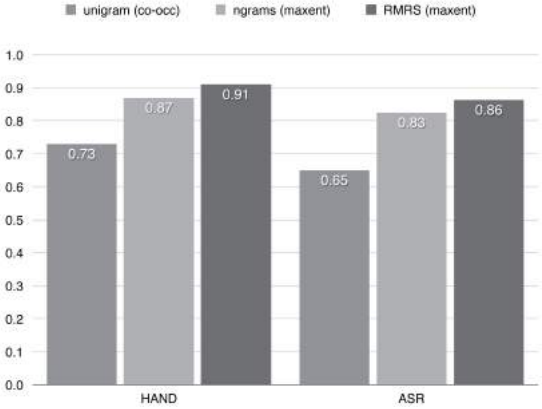


Figure 5.6: MRR for Experiment 1.

Incremental Results

Figure 5.7 shows the incremental results, comparing the NGRAM and RMRS models. Ideally, an incremental model would make an early correct decision and not change its mind. The figure shows that both models are respectable; both have relatively early first correct on average, and a fairly early first final with little edit overhead (12.4% for NGRAM and 8.1% for RMRS). However, the overall winner is RMRS as it makes an early decision with very little edit overhead before making a final decision, on average, earlier than the NGRAM model.

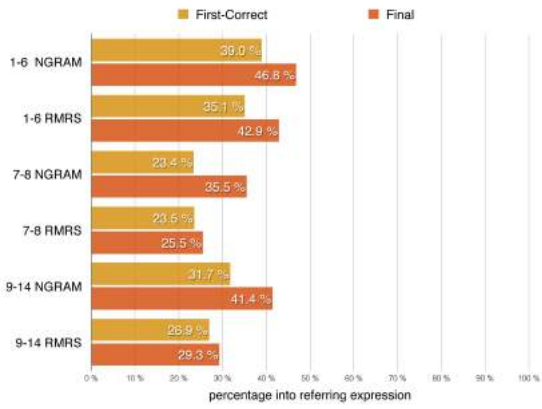


Figure 5.7: Incremental results for Experiment 1, earlier denotes better results.

Figure 5.8 illustrates incremental performance by showing the distribution over the pieces using the NGRAM model (though all three models are similar; lighter means higher probability)

for the utterance *das graue Teil in der ersten Reihe nehmen* (*the gray piece in the first row take* / take the gray piece in the first row) for each word in the utterance. When the first word, *das* is uttered, it already assigns probabilities to the pieces with some degree of confidence (note that in German, *das* (the) denotes the neuter gender, and the piece on the right with the lowest probability is often referred to by a noun (Treppe) other than neuter). Once *graue* (gray) is uttered, the distribution is now more even upon the three gray pieces, which remains largely the same when *Teil* (piece) is uttered. The next two words, *in der* (in the) give more probability to the left gray piece, but once *ersten Reihe* (first row) is uttered, the most probable piece becomes the correct one, the gray piece on the top.

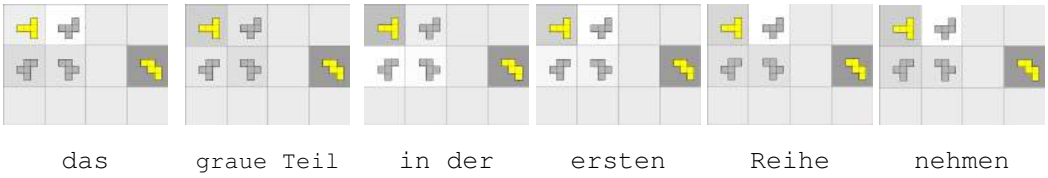


Figure 5.8: Example of reference resolution for the utterance: *das graue Teil in der ersten Reihe nehmen / the gray piece in the first row take*; lighter cell background means higher probability assigned to piece.

5.3.6 Remarks

This experiment sets the table for the SIUM model. Under varied conditions of U (unigram, ngram, and a semantic abstraction) and $P(R|U)$ (simple maximum likelihood co-occurrence or maximum entropy), the model is robust in both hand- and ASR-transcribed data. Using a semantic abstraction helps, but as it is sometimes difficult to obtain a semantic parser that works incrementally in a particular language (importantly, which can also work with speech data), it is useful to know that using ngrams provides respectable results. In the following experiments, we will continue to only use ngrams to represent U.

5.4 Experiment 2: Fusion with Gaze & Deixis

In this experiment, we show that SIUM can incorporate additional modalities such as gaze and deixis by treating the other modalities as models of RR of their own and interpolating the results.

5.4.1 Data

This experiment makes use of the TAKE data as explained in Chapter 4. An example episode is shown below, corresponding to the scene in Figure 5.9.

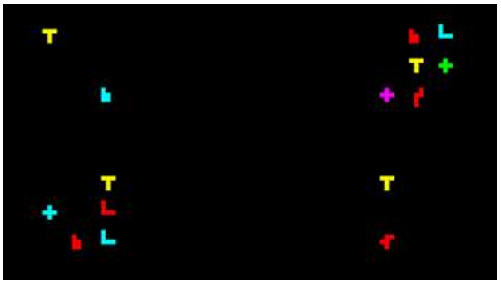


Figure 5.9: Example Pento board for gaze and deixis experiment; the yellow T in the top-right quadrant is the referent.

(5) a. *dann nehmen wir noch das zw- also das zweite t das oben rechts ist ... aus dieser gruppe da da möchte ich gern das gelbe t haben ... ja*

b. then we take now the se- so the second t that is on the top right ... out of this group there I would like to have the yellow t ... yes

c. $\begin{bmatrix} \text{REFERENT} & \text{object-3} \end{bmatrix}$

5.4.2 Fusing SIUM with Gaze & Deixis

The full model combines the evidence from linguistic information with evidence from other information sources such as the speaker's gaze and pointing gestures. For each, we calculate a reference point (R) on the scene: for gaze, the fixated point as provided by an eye tracker; for deixis, the point on the scene that was pointed at based on a vector calculated from the shoulder to the hand (as described in Kousidis et al. (2013), using the Microsoft Kinect). The centroids of all the objects (I) can then be compared to that reference point to yield a probability of that object being 'referred' by that modality (i.e., gazed at or pointed at) by introducing a Gaussian window over the location of the point:

$$p_{distance}(R_i, I_j; \sigma) = \exp -\frac{(x_i - x_j)^2}{2 * \sigma^2} * \exp -\frac{(y_i - y_j)^2}{2 * \sigma^2} \tag{5.9}$$

where the mean is R and σ is set by calculating the standard deviation of all the object centroids and the reference point. This can then be normalised over all the $p_{distance}$ scores to produce a distribution over I for each modality where the closer the object is to the reference point, the higher its probability. (We implicitly make the somewhat naive assumption here that

the referent will be looked at by the speaker most of the time during and around the RE. This is in general not true (Griffin and Bock, 2000), but works out here.)

Our final model of RR fuses the the three described modalities of speech, gaze, and deixis using a linear interpolation, where the α parameters are learned from held-out data by ranging over values such that the α values sum to one, and computing the average rank (metric explained below), retaining the α values that produced the best score for that set:

$$P(I|S) = P(I|S_1)\alpha_1 + P(I|S_2)\alpha_2 + P(I|S_3)(1 - \alpha_1 - \alpha_2) \tag{5.10}$$

5.4.3 Task & Procedure

The task is the same as Experiment 1. Beyond Experiment 1, we also want to incorporate deixis and gaze as separate models. Results were obtained by averaging the results of a 10-fold validation on 1000 Pento boards (i.e., utterances+context+gaze+deixis). The properties used for each object were colour, shape, and quadrant (there is no need to use the `selected` property here, as episodes were presented in isolation; i.e., there was no discourse context here as there was in Experiment 1).

5.4.4 Metrics

The metrics in this experiment are the same as in Experiment 1.

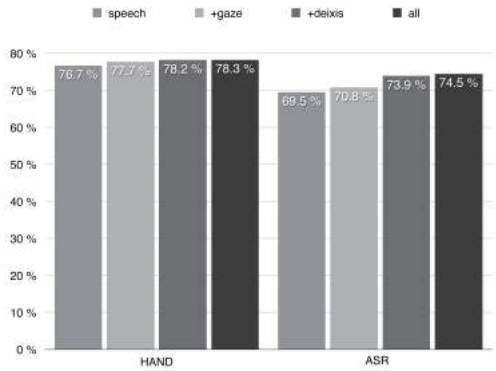


Figure 5.10: Comparison of model accuracies.

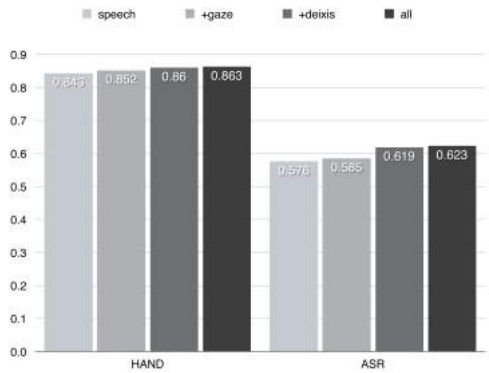


Figure 5.11: Comparison of MRR for the models.

5.4.5 Results

The results are shown in Figures 5.10 and 5.11. In both cases, there is improvement when incorporating gaze and deixis individually and together (the differences in all models is significant, except between HAND+deixis and HAND+all). The improvements are somewhat greater for ASR, as one might expect, as the model can benefit from the other modalities. Gaze doesn't generally help as much as deixis; we assume because of our naive assumption that the target object will be looked at most of the time during an episode.

5.5 Experiment 3: Fusing Deixis and Gaze as Properties under High Interactivity

Following Kennington et al. (2015b), In this experiment, we seek to evaluate SIUM's ability to work in a less constrained environment and on a very different language (Japanese) from what it has been evaluated on in the previous two experiments (German). The differences are explained in Chapter 4. In this experiment, gaze and a kind of deixis are incorporated into the model as in the previous experiment, however they are incorporated in a different way: instead of being interpolated as separate models, properties are extracted from gaze and deixis and included in R, directly used as part of SIUM. This is explained in further detail below. We also explore how setting the initial prior $P(I)$ to a distribution over the objects using contextual saliency information can improve the task performance.

Data

The corpus used in this experiment is the REX data described in Chapter 5, briefly repeated here. The REX corpora are a collection of human/human interaction data where the participants collaboratively solved Tangram puzzles. In the setting for this experiment, anaphoric references (i.e., pronoun references to entities in an earlier utterance, e.g., "move *it* to the left") and exophoric references via definite descriptions (i.e., references to real-world objects, e.g., "*that* one" or "the big triangle") are common (note that both refer in different ways to objects that are physically present). The corpus also records an added modality: the gaze of the puzzle *solver* (SV) who gives the instructions and that of the *operator* (OP), who moves the tangram pieces. The mouse pointer controlled by the OP could also be considered a modality, used as a kind of pointing gesture that both participants can observe. The goal of the task was to arrange puzzle pieces on a board into a specified shape (example in Figure 5.12), which was only known to SV and hidden from OP.

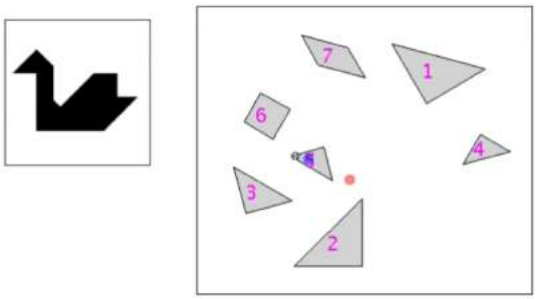


Figure 5.12: Example Tangram Board; the goal shape is the swan in the top left, the shared work area is the large board on the right, the mouse cursor and OP gaze (blue dot) are on object 5, the SV gaze is the red dot.

This environment provided frequent use of REs that aimed to distinguish puzzle pieces (and piece groups) from each other. The following are some example REs from the REX corpus:

(6) a. *chicchai sankakkei*
b. small triangle

(7) a. *sono ichiban migi ni shippo ni natte iru sankakkei*
b. that most right tail becoming triangle
'that right-most triangle that is the tail'

In Experiment 2, SIUM was applied to two datasets from the Pentomino domain in the two previous experiments, where the speaker's goal was to identify one out of a set of puzzle pieces. However, in these datasets, the references were "one-shot" and not embedded in longer dialogues, as is the case in the REX corpus.

Task & Procedure

The task is the same as Experiments 1 and 2. An important difference to Experiment 2, however, is that the gaze and pointing (via mouse cursor) modalities are represented in the set of properties R, and not as separate models, as explained in further detail below.

The procedure for this experiment is as follows. In order to compare our results directly with those of Iida et al. (2011), we provide our model with the same training and evaluation data, in a 10-fold cross-validation of the REs from 27 dialogues (the T2009-11 corpus in Tokunaga et al. (2012)). For development, we used a separate part of the REX corpus (N2009-11) that was structured similarly to this one.

As noted above, $P(R|I)$ models the likelihood of selecting a property of a candidate object for verbalisation; this likelihood is assumed to be uniform for all the properties that the candidate object has. We derive these properties from a representation of the scene; similar to how Iida et al. (2011) computed features to present to their classifier: namely **Ling** (linguistic features), **TaskSp** (task specific features), and **Gaze** (from SV only). Some features were binary, others such as shape and size had more values. Table 5.13 shows all the properties that were used here. Each will now be explained.

Ling	**TaskSp**
`tri/squ/pgram`	`most_recent_move`
`small/med/big`	`mouse_pointed`
`left/mid/right`	
`prev_referred`	**Gaze**
`top/cen/bottom`	`most_gazed_at`
`referred_5`	`gazed_at_in_utt`
`referred_10`	`longest_gazed_at`
`referred_20`	`recent_fixation`

Figure 5.13: List of properties used for each source of information.

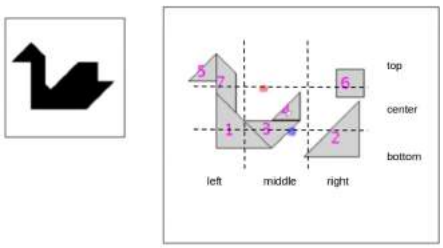


Figure 5.14: Tangram later in the dialogue; the notion of *right-ness* and other spatial concepts changes throughout the dialogue (compare to Figure 5.12), the grids are added to show which objects receive which horizontal and which vertical properties.

Ling Each object had a shape, size, and relative position to the other pieces. We determined by hand the shape and size properties which remained static through each dialogue. The position properties were derived from the corpus logs. For each object, the centroid of each object was computed. Then, the vertical and horizontal range for all of the objects was calculated, then split into three even sections in each dimension (see Figure 5.14). An object with a centroid in the left-most section of the horizontal range received a `left` property, similarly `middle` and `right` properties were calculated for corresponding objects. For vertical placement, `top`, `center` and `bottom` properties were given to objects in the respective vertical segments. Figure 5.14 shows an example segmentation. Each object had a vertical and a horizontal property at all times, however, moving an object could result in a change of one of these spatial properties as the dialogue progressed. As an example, compare Figure 5.12, which is a snapshot of the interaction towards the beginning, and Figure 5.14, which shows a later stage of the game board; spatial layout changes throughout the dialogue.

These properties differ somewhat from the features for the Ling model presented in Iida

et al. (2011). Three features that we did use as properties had to do with reference recency: the most recently referred object received the `referred_X` properties, if an object was referred to in the past 5, 10, or 20 seconds.

TaskSp Iida et al. (2011) used 14 task-specific features, three of which they found to be the most informative in their model. Here, we will only use the two most informative features as properties (the third one, whether or not an object was being manipulated at the beginning of the RE, did not improve results in a held-out test): the object that was most recently moved received the `most_recent_move` property and objects that have the mouse cursor over them received the `mouse_pointed` property (i.e., like a pointing gestures–see Figure 5.14; object 4 would receive both of these properties). Each of these properties can be extracted directly from the corpus annotations.

Gaze Similar to Iida et al. (2011), we consider gaze during a window of 1500ms before the onset of the RE. The object that was gazed at the longest during that time received a `longest_gazed_at` property, the object which was fixated upon most recently during that interval before the RE received a `recent_fixation` property, and the object which had the most fixations received the `most_gazed_at` property. During a RE, an object received the `gazed_at_in_utt` property if it is gazed at during the RE up until that point. These properties can be extracted directly from the corpus annotations. Other gaze features are not really accessible to an incremental model such as this, as gaze features extracted from gaze activity over the RE can only be computed when it is complete. Our Gaze properties are made up of these 4 properties, as opposed to the 14 features in Iida et al. (2011).

Properties over Time In the previous experiments it was certainly the case that properties could change over time, but in this experiment the change in properties can happen very quickly for the simple fact that eye gaze fixations can changed from moment to moment and objects are being manipulated in real-time. Figure 5.15 shows how properties for objects can change over time. At t_1, the `prev-referred` and `recent-fixation` properties belonged to o_5, but at t_2 those properties were assigned to the recently-manipulated o_4, a change which could take place quickly between (or during) REs. Similarly, o_4 had a `middle` property (as could other objects) at t_1, but by t_2 it had been moved, thus losing that property and gaining the `left` property.

P(R|U) For $P(R|U)$ we trained a Naive Bayes classifier. On the language side (the variable U in the model), we used n-grams over Japanese characters (we attempted tokenisation of the

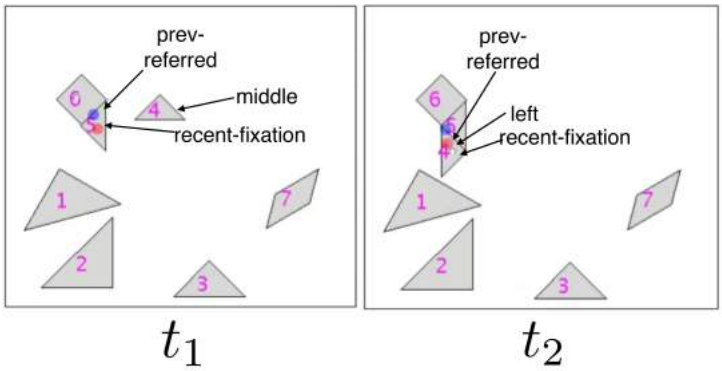


Figure 5.15: Illustration of how properties change over time: at t_1, o_5 had the `prev-referred` and `recent-fixation` properties, but later at t_2, both of those properties were assigned to o_4. Also, at t_1, o_4 had a `middle` property, but by t_2 it had lost that property and gained the `left` property.

REs into words, but found that using characters worked just as well in the held-out set).

P(I) The prior $P(I)$ is handled differently in this experiment than in previous experiments. Recall that $P(I)$ is the posterior of the previously computed increment. In the first increment, it can simply be set to a uniform distribution as was done in. Here, we apply a more informative prior based on saliency. We learn a *context model* which is queried when the first word begins, taking information about the context immediately before the beginning of the RE into account, producing a distribution over objects, which becomes $P(I)$ of the first increment in the RE. The context model itself is a simple application of the SIUM, where instead of being a word, U is a token that represents saliency. The context model thus learns what properties are important to the pre-RE context and provides an up-to-date distribution over the objects as a RE begins.

5.5.1 Metrics

In order to compare with previous work, we will show only model accuracies. We will also show incremental metrics for our model as explained in Experiment 1.

5.5.2 Results

Utterance-level Results

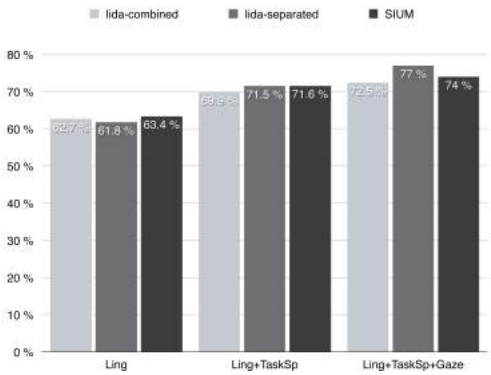


Figure 5.16: Comparison of model accuracies.

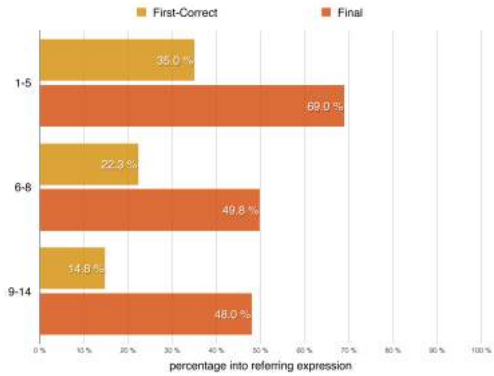


Figure 5.17: Incremental results for SIUM, numbers represent % into RE.

Results of our evaluation are shown in Figure 5.16. The SIUM model performs better than the combined approach of Iida et al. (2011), and performs better than their separated model—when not including gaze. This is a welcome result, as it shows that our very simple incremental model that uses a basic classifier is comparable to a non-incremental approach that uses a more complicated classifier. It further shows that the SIUM model is robust to using TaskSp and Gaze features as properties, as long as those features are available immediately before the RE begins, or during the RE.

The best-performing approach is the Iida2011-separated model with gaze. This is the case for several reasons: first, their models use features that are not available to our incremental model (e.g., their model uses 14 gaze features, some of which were based on the entire RE, ours only uses 4 properties). Second, and more importantly, separated models means less feature confusion: in Iida et al. (2011) (Section 5.2), the authors give a comparison of the most informative features for each model; task and gaze features were prominent for the pronoun model, whereas gaze and language features were prominent for the non-pronoun model. We also tested SIUM under separated conditions to better compare with the approaches presented here. The separated models, however, did not improve. This, we suppose, is because the model grounds *language* with properties (see Discussion below). An interactive dialogue system might not have the luxury of choosing between two models at runtime. We assume that a model that can sufficiently handle both types of utterances is to be preferred to one that doesn't.

An important thing to note is the difference between the Ling and the Ling+TaskSp+Gaze for the SIUM model. This can be compared to the improvements in the previous experiment

when incorporating deixis and gaze; here, two (similar) modalities were also incorporated, but this time as properties. It would appear here that improvements are potentially better when treating modalities as properties instead of as separate models, as in the previous experiment.

Incremental Results

Table 5.17 shows how our model fares using the incremental metrics described earlier. (As this has not been done in Iida et al. (2011), direct comparison is not possible.) For the evaluation, REs are binned into short, normal, and long (1-5, 6-8, 9-14 characters, respectively, based on what the average numbers of words in REs in this corpus is), to make relative statements ("% into the utterance") comparable.

As noted before, an ideal system would make the first correct decision as early as possible without changing that decision. The results in the table show a respectable incremental model; on average it picks the right object early, with some edit overhead (less than 1%, making unnecessary changes in its prediction), finally fixing on a final decision before the end of the RE with low edit overhead, meaning it rarely changes its mind once it has made a decision. These incremental results are consistent with previous work for the SIUM; overall, the model is stable across the RE.

5.5.3 Analysis

Incorporating saliency information via a context model proved to be quite useful for this data. In this experiment, we computed the initial $P(I)$ using a context model instantiated by SIUM. By considering only this saliency information, the context model can predict the referred object in 41% of the cases. It also learned which properties were important for saliency (that is, these are the properties that the model would most likely select): `recently_fixated`, `most_gaze_at`, `longest_gazed_at`, `prev_ref`, as might be expected. In less than 2% of the cases, the context model referred to the correct object, but was wrongly "overruled" when processing the corresponding RE.

There were shortcomings, however. In previous work, it was shown that SIUM performed well when REs contained pronouns (see Experiment 1 above and Kennington et al. (2013)). However, in the current experiment we observed that REs with pronouns were more difficult for the model to resolve than the model presented in Iida et al. (2011). We surmise that SIUM had a difficult time grounding certain properties, as the Japanese pronoun *sore* can be used anaphorically or demonstratively in this kind of context (i.e., sometimes *sore* refers to previously-manipulated objects, or objects that are newly identified with a mouse pointer over them); the model presented in Iida et al. (2011) made more use of contextual information when

pronouns were used, particularly in the combined model which incorporated gaze information, as shown above.

5.5.4 Remarks

This and the previous experiment were multimodal; both used gaze information from the speaker, as well as pointing information (computed deictic gestures in the previous experiment, and "pointing" gestures with the mouse cursor in this experiment), yet the two modalities were fused with SIUM in very different ways. In the previous experiment, each modality was treated as an individual model of RR, then interpolated. In this experiment, the modalities were treated as properties of objects (e.g., an object could have the `recent_fixation` property). We point out here that the model can make use of both approaches. Treating additional modalities as properties is possibly the more principled approach as certain words, e.g., *sore* (*that*) can ground with certain properties (e.g., `mouse_pointed`).

It was also shown in this experiment that the initial $P(I)$ can encode contextual information to give a more informed initial distribution over the objects. When such information is not available, the initial $P(I)$ can be treated as uniform, as done in the previous two experiments.

5.6 Experiment 4: Uncertainty in the Perception of the World

In this experiment, we focus on $P(R|I)$. Until now, we have claimed that R would be picked out to be uttered by a speaker based on properties that an object has; however, here we insert uncertainty into R, as is explained below. How this is manifested in SIUM is shown in Table 5.1-B which shows that properties in R for each object in I contain no uncertainty and Table 5.2, which shows that an object has a distribution over property types.

5.6.1 Data

The data used in this experiment is the same data as Experiment 2, with some additional derivations of the scenes, which will now be described.

5.6.2 Scene Processing

Following Kennington et al. (2015a), we want our model to work with images of real objects as input, even though for our particular data the scenes are represented symbolically (that is, we know without uncertainty each piece's shape, colour, and position). Using the images that were generated from these symbolic descriptions by Kousidis et al. (2013) and performing computer vision on them does not introduce much uncertainty, as there is no variation in colour

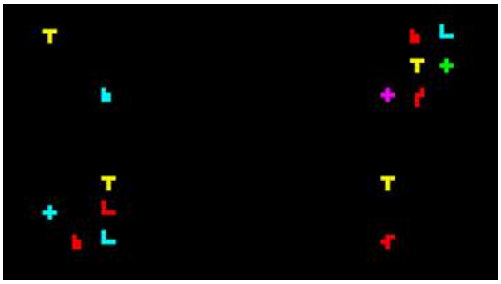


Figure 5.18: Example Pento board for gaze and deixis experiment; the yellow T in the top-right quadrant is the referent.

Figure 5.19: Example Pento Board that has been distorted from its original form (Figure 5.18).

or appearance of individual shapes, and so the data cannot serve to form generalisations. To get closer to conditions as they would hold when working with camera images (e.g., variations of colour due to variations in lighting, distortion of shapes due to camera angles, etc.), we pre-processed these images: We shifted the colour spectrum as follows: the hue channel by a random number between -15 and 15 and the saturation and value channels by a random number between -50 and 50. For the object shapes, we apply affine transformations defined by two randomly generated triangles and warp the image using that transform. This generates more complex shapes that retain some notion of their original form. Figure 5.19 shows a game board that has been distorted from its original (Figure 5.18).

Using these distorted images, we processed each image using the Canny Edge Detector (Canny, 1986) and used mathematical morphology to find closed contours of the objects, thereby segmenting the objects from each other. We acquired the boundary of the objects (always 15 of them), following the inner contours as identified by the border tracing algorithm (Suzuki and Abe, 1985). For each individual object we then extract the number of edges, RGB (red, green, blue) values, HSV (hue saturation value), and from the object's *moments*: its centroid, horizontal and vertical skewness (third order moments measuring the distortion in symmetry around the x and y axis), and the orientation value representing the direction of the principal axis (combination of second order moments).

From these features, we compute a distribution over the set of colours and shapes using a SVM classifier for each. Position properties are computed using a set of rules; objects above a certain y value receive the `top` property, if below then `bottom`, if to the left of a certain x threshold, the `left` property, if to the right, the `right` property (other position properties have 0 probability). With these properties, each object now has all colours and shapes, albeit

with differing probabilities. This represents the uncertainty in the properties for each object.

5.6.3 Task & Procedure

The task is RR, as described earlier. At each increment, the model returns a distribution over all objects; the probability for each object represents the strength of the belief that it is the referred one. The argmax of the distribution is chosen as the hypothesised referent.

Using 1000 episodes, we evaluate our model across 10 folds, where 900 episodes (utterances+scenes) were used for training, and the remaining 100 were used to test the model. Our baseline model is *random selection* (which gets an accuracy of 7%).

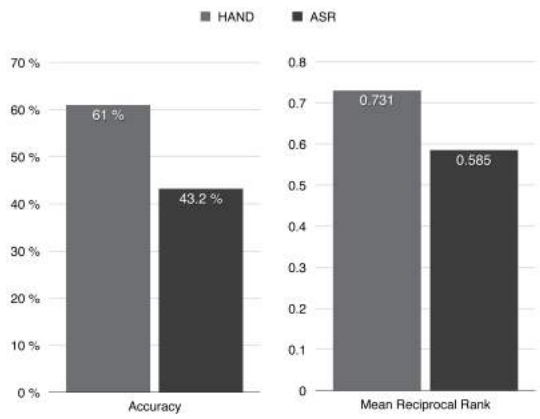


Figure 5.20: Results of our model in accuracies; higher numbers are better results for accuracies, lower numbers denote better results for average rank.

5.6.4 Metrics

The metrics in this experiment are the same as in Experiment 1.

5.6.5 Results

Utterance-level Results

Compared to the speech-only results for Experiment 2 in Figure 5.10, when inserting uncertainty into the model, there is a fairly dramatic decrease (from 76.7% to 61%) in RR accuracy. The model still picks out the top referent out of 15 more than 61% of the time, but there is a clear performance hit. We will provide further discussion below.

Incremental Results

Compared to Experiment 2 (speech model only), the incremental results shown in Figure 5.21 have first-correct and first-final values that are much later. On average, when uncertainty exists in the representation of W, SIUM makes a final decision before the end of the RE, but is later than before in coming to that decision after an increased amount of edit overhead. Overall, the model remains fairly robust even when there is uncertainty in how the scene is represented.

% edit overhead	
1-6	3.8
7-8	17.2
9-14	27.5
% never correct	
all lengths	32.0

Table 5.4: % edit overhead and never correct

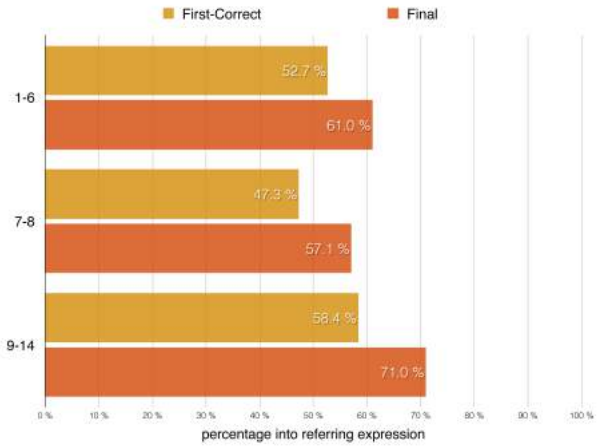


Figure 5.21: Incremental Performance

5.6.6 Systematic Insertion of Uncertainty

In this section we explore how well SIUM performs when varied levels of uncertainty are inserted into R, specifically the colours. Using the data from Experiment 2 and 3, each object could have one of 7 colours (`red`, `blue`, `green`, `yellow`, `pink`, `gray`, `cyan`). In a scene without uncertainty, each object has only one colour that receives all of the probability mass. Increasing uncertainty amounts to removing probability mass from that colour and distributing it across the other colours, where colours that are closer in the colour spectrum (e.g., `red` is closer to `pink` than it is to `green`) get more mass than those that are farther away. As the amount of probability mass removed from the original colour increases, so does the entropy of the distribution over the colours. All other non-colour properties were their original values (i.e., they were fully observed).

Using a single fold of the data for training, we evaluated 100 boards 100 times, each time increasing the entropy over the colours for each object on each board. Figure 5.22 shows the accuracy of the model at each point of average entropy over the object colours. As expected, as the average entropy over the colours increases so does the uncertainty, which leads to decreased

accuracy. A major drop off occurs around 1.7, where the probability mass over the colours became more uniform. To compare, the average entropy over the colours of the distorted images used in this experiment was 0.047, which is fairly low and should be stable when compared with Figure 5.22, but since there is also uncertainty in the shapes (average entropy of 0.032) which is a more realistic setting, the results were considerably worse.

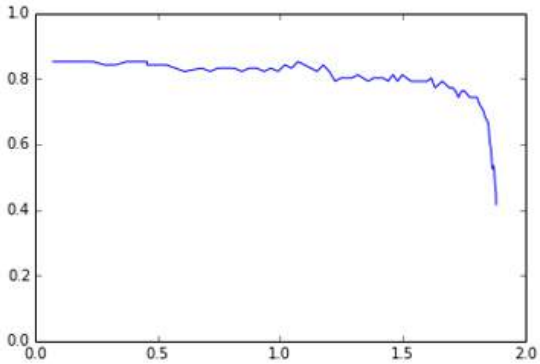


Figure 5.22: Accuracy of model (y-axis) decreases as average entropy over colours increases (x-axis).

5.7 Experiment 5: Using SIUM for Natural Language Understanding

We have shown that SIUM is a robust model of RR in various settings and conditions. In this experiment, we want to explore if the model could be applied to a standard task of *natural language understanding* (NLU) where the model must fill in an unspecified number of slots in a semantic frame, rather than referring to an object in a scene. To do so, we treat the task of NLU as a RR task.

5.7.1 Data

For this experiment, we apply SIUM to the well-known *airline travel information system* (ATIS) data (Dahl et al., 1994; Hemphill et al., 1990) as preprocessed by Meza-Ruiz et al. (2008) and He and Young (2005). For each utterance, we have a semantic frame and tag sequence that aligns semantic concepts, example in (8). Note here (following Heintze et al. (2010)) a big difference between this task and the previous experiments is that in the ATIS representations, the slot values are simply the words in the utterance. The word itself is its own semantic representation and no additional semantic abstraction is performed. In the previous experiments, the slot is known beforehand (i.e., a slot for the referent) and filling it requires interpreting the

utterance to determine the referent. Note that this data is in English, which has not been applied to SIUM in the previous experiments.

We use the ATIS training set, which consists of 4481 utterances between 1 and 46 words in length (avg 11.46; sd 4.34), with a vocabulary of 897 distinct words. There are 3159 distinct frames, 2594 (58%) which occur only once. An additional slot, `goal` represents the overall goal of the utterance.

(8) a. What flights are there arriving in Chicago after 11pm?

b. $$\begin{bmatrix} \text{GOAL} & \text{flight} \\ \text{TOLOC.CITY_NAME} & \text{Chicago} \\ \text{ARRIVE_TIME.TIME_RELATIVE} & \text{after} \\ \text{ARRIVE_TIME.TIME} & \text{11pm} \end{bmatrix}$$

5.7.2 Task & Procedure

The task for SIUM is to determine the slots and fill the `goal` slot. Filling the `goal` slot is similar to the task of RR in previous experiments, but instead of assuming the domain of discourse is a set of visible objects, it is the set of possible values that can fill the `goal` slot (e..g, `flight`).

A single model is trained for the slots and an additional model is trained for the `goal` (i.e., two instantiations of SIUM). Each slot is predicted incrementally using the current trigram, the part-of-speech tag, and the previous slot tag (i.e., these features make up U). For the `goal`, each word is processed incrementally, the argmax prediction at the end of the utterance is used as the final guess. We modelled $P(U|R)$ using a maxent classifier for the slots and a naive Bayes classifier for the `goal` (determined using a development set).

For the slots, I is the set of possible slots (e.g., `toloc.city_name`, `arrive_time.time`, etc.). For `goal`, I is the set of possible goals. As these are non-visual entities, it isn't clear what constitutes the properties R for I. In this case, we can treat the entity as its own property, and that is the only property that it has (e.g., $P(I = toloc.city_name|R = toloc.city_name) = 1.0$, all other values for R are 0).[4]

5.7.3 Metrics

In order to directly compare our results with previous work, we calculate an f-score of the full frame as predicted by the end of the utterance.

[4]This is an abstract kind of property as noted earlier, e.g., `new-york` is a property of *New York*.

5.7.4 Results

The results in Figure 5.23 show that SIUM does not produce state-of-the-art results, as expected. As it processes incrementally, it only uses left-context information for U, whereas other approaches use the global context of the entire utterance, yielding better results. We aren't concerned here that it does not produce the best results, the goal of this experiment was to show that SIUM, though designed to be a robust approach to incremental RR, can also be applied to standard NLU tasks with respectable results. This is useful knowledge when designing a dialogue system, it would potentially be better to use a single framework for several tasks such as NLU or RR.

MAIRESSE ET AL. (2009)	94.5
HE AND YOUNG (2005)	90.3
ZETTLEMOYER AND COLLINS (2007)	95.9
MEZA-RUIZ ET AL. (2008)	91.56
SIUM	88.2

Figure 5.23: Results for ATIS evaluation.

5.8 Discussion

We have tested the SIUM in several experiments. We have shown that the model works respectably in a NLU and and several RR tasks. It can make use of a semantic representation (Experiment 1), but fares respectably when using just ngrams. It can incorporate additional modalities, such as gaze and pointing gestures by treating them as separate models and interpolating them (Experiment 2), but can also directly make use of the modalities by treating them as properties (Experiment 3). The model can handle definite references, as well as pronouns (of an exophoric nature, by adding properties to recently referred objects) and deictic exophoric references (Experiments 1 and 2). The model has been tested on German (Experiments 1, 2, and 4), Japanese (Experiment 3), and English (Experiment 5) data.

We can also substantiate from Experiments 1 and 3 that this model is (at least partially) a unified model of RR similar to Funakoshi et al. (2012), in that it can resolve definite descriptions, pronouns, and deictic gestures in a single framework. Indeed, words in the REs are grounded to properties that represent aspects of the world that are definite (such as colour, shape, and position), discourse context, i.e., pronouns (such as the `selected` property in Experiment 1, or the `most_recent_moved` property in Experiment 3), as well as deictic

gestures (such as the `mouse_pointed` property). Note, however, that all of these types of reference do indeed refer to visually present objects and the pronouns are not resolved to linguistic antecedents.

In terms of noise and uncertainty, the model seems to handle speech recognition output respectably (Experiments 1 and 2), and as Experiment 4 showed, it is fairly robust to uncertainty in the scene representation. Additional analysis of SIUM is provided in Chapter 7.

Importantly, SIUM is quite simple to implement and works incrementally. SIUM has been implemented as an incremental processing module in the INPROTK toolkit (Baumann and Schlangen, 2012).[5]

With this evidence, we can easily make the claim that SIUM fulfils the goals set forth in the introductory chapter of this thesis:

- The model can resolve referring expressions incrementally (i.e., word by word).
- The model learns, given data, a mapping between visually present objects (represented either directly or by some mediating variable) and words in referring expressions.
- Given novel REs and novel scenes, the model can generalise and resolve under new circumstances.
- The model can handle definite descriptions and demonstrative references (i.e., REs with pointing gestures) in a single framework.
- The model can be implemented as a component in a spoken dialogue system.
- The model can be evaluated to show that it resolves references despite possible issues (e.g., noisy conditions) with the representation of the referring expressions or the objects.

The other two goals (formulating the model to account for word meaning and fitting it into a formal framework), however, are not addressed. That these are not addressed are examined in the following section.

5.9 Intension, Extension, and the Simple Incremental Update Model

In this section we connect SIUM with the grounded FOL framework set forth in Chapter 2:

$$[\![w]\!]_{obj} = \lambda x.\phi_w(x) \tag{5.11}$$

[5] `https://bitbucket.org/inpro/inprotk`

Recall that ϕ_w is some kind of concept or class for w (i.e., a word), and x is the set of features of an object that allow some kind of mechanism to assign that object to that particular class. For SIUM, the set of classes are in fact the properties that make up R. For example, the property of being a cross-shape would be X, or in the framework $\lambda x.X(x)$ and the property for red is `red`, or $\lambda x.red(x)$ in the framework. Assigning objects to those properties, for the most part, is done via rules (the $P(R|I)$ portion of the model) if the scene data are already represented symbolically as in the virtual scenes of ACTION and REX. The model does learn, however, when those concepts should be used by learning the mapping between words and properties in $P(U|R)$; or, for most of our experiments, $P(R|U)$, which produces a distribution over R given some kind of representation of U (i.e., either ngrams or a semantic abstraction). All properties are instantiated as their respective versions of formula 5.11 (albeit with weights).

For SIUM, the *extension* for a given RE is the distribution over the candidate objects. The *intension* is a little less clear; the meaning of a word, for example, is the probability distribution over R which that word produces in a given context (i.e., given an ngram sequence or a semantic abstraction), thus every object belongs to every class to different degrees, where the degree of each class falls within a normalised distribution over the classes. This doesn't quite get at the meaning of words, but it is a step closer than simply mapping directly from words to concepts by hand.

In short, SIUM is a step closer to our goal of being able to assign objects to classes, but we have to predefine the set of classes. In the next chapter, we will look at another model of RR that does not require a pre-determined set of classes; i.e., it can map directly from I to U without a mediating variable like R.

5.10 Chapter Summary

In this chapter, we motivated the SIUM by how a model of RR would be conceptually implemented in a IU-module network. We derived a generative model from the joint probability of intended objects I, properties that those objects might have R, and the aspects of REs U. We explained the SIUM formulation, properties, and gave a toy example. We then looked at several experiments and cited work that rigorously evaluated the model when varying variables such as U and R, as well as the sub-models, such as $P(U|R)$ and $P(R|I)$. We found that the model works well when there is uncertainty on both U (via noisy ASR) and R (via a distribution over properties).

6

The Words-as-Classifiers Model: A Discriminative Model of Incremental Reference Resolution

All our knowledge has its origins in our perceptions.

- Leonardo da Vinci

This chapter presents the *Words-as-Classifiers* (WAC) model, a discriminative model of reference resolution. Like SIUM in the previous chapter, WAC is formulated and implemented to work in an update-incremental fashion, without re-computing previous increments. It differs from SIUM in some key ways that improve over SIUM, the principle difference being that the model doesn't require the mediating properties variable to map from visual aspects to words– that is done directly. It also doesn't require a symbolic representation of the world W, rather it can use low-level features to represent objects. Also, being discriminative, it models directly the mapping between aspects of the world and aspects of the referring expressions.

In the following section, we define and explain the model. That is followed by two experiments each with error analyses, and further discussion on how this model fits into the formal

framework set forth in Chapter 2. One of the principle motivations for this model is that it directly models word meanings, something which the SIUM model was not fully able to do.

6.1 Model Definition

The basis of this model is a function from perceptual features of a given object to a judgement about how well those features "fit" together with a word in a RE. Such a fit, we claim, models the word meaning. Two different types of word meaning are modelled:

1. *simple reference*: Those that "pick out" the properties of single objects. The properties referred to here are the same properties as have been described thus far, such as colour (e.g., the word *green* in the RE *the green square*), shapes, global-spatial descriptions (e.g., *top*, *left*), etc; following Kennington et al. (2015a).

2. *relational reference*: Those that pick out relations between two objects, such as *next-to*, *above*, etc. (e.g., *the red circle next to the green square*); following Kennington and Schlangen (2015).

Importantly, these word meanings are learned from instances of language use, i.e., by observing REs and some kind of information as to the object that was actually referred.

Once these word meanings are learned, the second component then is the *application* of these word meanings in the context of a novel RE. This application gives the desired result of a probability distribution over candidate objects, where the probability expresses the strength of belief in the object being the referent. These applications are strictly compositional in the sense that the meanings of the more complex constructions are a function of those of their parts.

6.1.1 Word Meanings

Word Meanings in Simple References

The first type of word meaning we model, that of a simple reference, picks out a single object via its visual properties.[1] To model this, we train for each word w from our corpus of REs a binary logistic regression classifier that takes a representation of a candidate object via visual features ($\mathbf{x}$) and returns a probability p_w for it being a good "fit" to the word (where $\mathbf{w}$ is the weight vector that is learned and σ is the logistic function):

$$p_w(\mathbf{x}) = \sigma(\mathbf{w}^\intercal \mathbf{x} + b) \tag{6.1}$$

[1] Visual features are the types of features that we use here, but in principle any kind of feature can be used.

Formalising the correspondence mentioned above, the meaning of a word can in this approach then be seen as the classifier itself, a function from a representation of an object to a probability, which matches the secondary goal of this thesis to model word meanings (though perhaps a loose definition of "meaning" as it pertains to visual perception; here, w represents the variable *word*):

$$[\![w]\!]_{obj} = \lambda \mathbf{x}.p_w(\mathbf{x}) \tag{6.2}$$

Where $[\![w]\!]$ denotes the meaning of w, and $\mathbf{x}$ is of the type of feature given by f_{obj}, the function computing a feature representation for a given object. This is depicted graphically in Figure 6.1

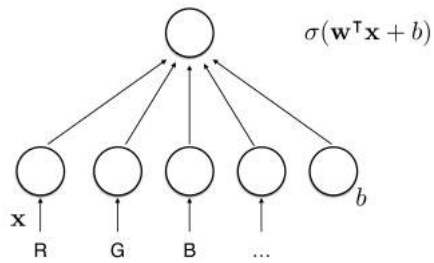


Figure 6.1: Representation as 1-layer NN.

We train these classifiers using a corpus of REs coupled with representations of the scenes in which they were used, and an annotation of the referent of that scene. In the set of training samples, logistic regression classifiers needs access to *positive* and *negative* training samples. To get positive training examples, we pair each word of a RE with the features of the referred object. To get negative training examples, we pair the word with features of (randomly picked) other objects present in the same scene, but *not* referred to by the RE. This selection of negative examples makes the assumption that the words from the RE apply only to the referent. This is wrong as a strict rule, as other objects could have similar visual features as the referent (e.g., there could be two red circles in a scene where one of them is the referent–hence the positive example–and the other was randomly chosen as the negative example); for this to work, however, this has to be the case only more often than it is not.

Word Meanings in Relational References

The second type of word that we model, that of a relational reference, expresses a relation between objects. Its meaning is trained in a similar fashion as simple reference, except that it is presented a vector of features of a *pair* of objects, such as their euclidean distance, vertical and horizontal differences, and binary features denoting higher than/lower than and left/right relationships. Thus the only difference between *relational* word and *simple* word classifier are the features that are presented to them.

6.1.2 Application and Composition

In this section we explain how the model is applied in a RR task.

Simple References

The model just described gives us a prediction for a word paired with an object (or pair of objects). What we wanted, however, is a distribution over all candidate objects in a given utterance situation, and not only for individual words, but for incrementally growing REs. We begin by looking at the simple references—which roughly corresponds to simple NPs—that refer only by mentioning properties of the referent (e.g. *the red cross on the left*). To get a distribution for a single word, we apply the word classifier to all candidate objects and normalise; ($\mathbf{x}_i$ is the feature vector for object i, $normalize()$ vectorized normalisation, and I a random variable ranging over the candidates; again, here w is a word):

$$\begin{aligned} [\![w]\!]^W_{obj} &= normalize(([\![w]\!]_{obj}(\mathbf{x}_1), \ldots, [\![w]\!]_{obj}(\mathbf{x}_k))) \\ &= normalize((p_w(\mathbf{x}_1), \ldots, p_w(\mathbf{x}_k))) \\ &= P(I|w) \end{aligned} \tag{6.3}$$

In effect, this combines the individual classifiers into something like a multi-class logistic regression (i.e., maximum entropy) model—but, importantly, only for application. The training regime did not need to make any assumptions about the number of objects present, as it trained classifiers for a 2-class problem (how well does this given object fit to the word?). The multi-class nature is also indicated in Figure 6.2, which shows multiple applications of the logistic regression network for a word, and a normalisation layer on top.

To compose the evidence from individual words $w_1, \ldots, w_k$ into a prediction for a 'simple' RE $[_{sr} w_1, \ldots, w_k]$ (where the bracketing indicates the structural assumption that the words belong to one, possibly incomplete, 'simple reference'), we compute a weighted interpolation

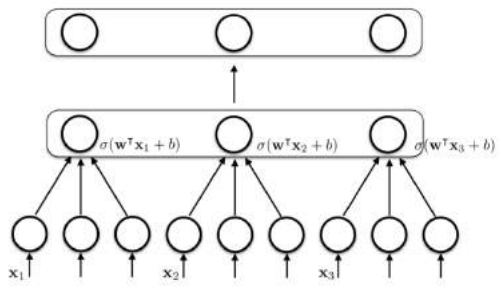


Figure 6.2: Representation as network with normalisation layer.

of each classifier's contribution to the RE. In it's simplest form, this is an *average* of the contributions of its constituent words, which we do here. The averaging function $avg()$ over distributions then is the contribution of the construction 'simple reference (phrase)', sr, and the meaning of the whole phrase is the application of the meaning of the construction to the meaning of the words:

$$\begin{aligned} [\![[_{sr} w_1, \ldots, w_k]]\!]^W &= [\![sr]\!]^W [\![w_1, \ldots, w_k]\!]^W \\ &= avg([\![w_1]\!]^W, \ldots, [\![w_k]\!]^W) \end{aligned} \tag{6.4}$$

where $avg()$ is defined as

$$\begin{aligned} avg([\![w_1]\!]^W, [\![w_2]\!]^W) &= P_{avg}(I|w_1, w_2) \text{ with } P_{avg}(I = i|w_1, w_2) \\ &= \frac{1}{2}(P(I = i|w_1) + P(I = i|w_2)) \text{ for } i \in I \end{aligned} \tag{6.5}$$

The averaging function is inherently incremental, in the sense that $avg(a, b, c) = avg(avg(a, b), c)$ and hence it can be extended "on the right". This represents an update-incremental model, resulting in an intersective way of composing the meaning of the phrase. This cannot account for all constructions (such as negation or general quantification), of course; we leave exploring other constructions that could occur even in our 'simple references' to future work.

Relational References

Relational references have a more complex structure, being a relation between a two simple references: reference to a *landmark* and a reference to a *target*. This structure is indicated abstractly in the following 'parse': $[_{rel}[_{sr}w_1, \ldots, w_k][_r r_1, \ldots, r_n][_{sr}w'_1, \ldots, w'_m]]$, where the w are the target words, r the relational expression words, and w' the landmark words.

As mentioned above, the relational expression similarly is treated as a classifier (in fact, technically we contract expressions such as "to the left of" into a single token and learn one classifier for it), but expressing a judgement for pairs of objects. It can be applied to a specific scene with a set of candidate objects (and hence, candidate pairs) in a similar way by applying the classifier to all pairs and normalising, resulting in a distribution over pairs:

$$[\![r]\!]^W = P(R_1, R_2|r) \tag{6.6}$$

We expect the meaning of the phrase to be a function of the meaning of the constituent parts (the simple references, the relation expression, and the construction), that is:

$$\begin{aligned}[\![[_{rel}[_{sr}w_1, \ldots, w_k][_r r][_{sr}w'_1, \ldots, w'_m]]]\!] = &[\![rel]\!]([\![sr]\!][\![w_1 \ldots w_k]\!], [\![r]\!], \\ &[\![sr]\!][\![w'_1 \ldots w'_m]\!])\end{aligned} \tag{6.7}$$

(dropping the indicator for concrete application, ^{W} on $[\![\,]\!]$, for reasons of readability).

What is the contribution of the relational construction, $[\![rel]\!]$? Intuitively, what we want to express here is that the belief in an object being the intended referent should combine the evidence from the simple reference to the landmark object, and that for the relation between them (e.g., *next to*). Instead of averaging (that is, combining additively), as for sr, we combine this evidence multiplicatively: If the target constituent contributes $P(I_t|w_1, \ldots, w_k)$, the landmark constituent $P(I_l|w'_1, \ldots, w'_m)$, and the relation expression $P(R_1, R_2|r)$, with I_l, I_t, R_1 and R_2 all having the same domain, the set of all candidate objects, then the combination is

$$\begin{aligned}P(R_1|w_1, \ldots, w_k, r, w'_1, \ldots, w'_m) = &\sum_{R_2}\sum_{I_l}\sum_{I_t} P(R_1, R_2|r) * P(I_l|w'_1, \ldots, w'_m) * \\ &P(I_t|w_1, \ldots, w_k) * P(R_1|I_t) * P(R_2|I_l)\end{aligned} \tag{6.8}$$

As for SIUM, The last two factors force identity on the elements of the pair and target and landmark, respectively (they are not learnt, but rather set to be 0 unless the values of R and

I are equal), and so effectively reduce the summations so that all pairs need to be evaluated only once. The contribution of the construction then is this multiplication of the contributions of the parts, together with the factors enforcing that the pairs being evaluated by the relation expression consist of the objects evaluated by target and landmark expression, respectively.

6.1.3 Open Questions about WAC

Similar to the previous chapter, we identify here several open questions about the WAC model as explained. Each question is explained and most are substantiated in the experiments described below.

- *How can grounding occur?* Grounding occurs directly; indeed the "meanings" as described above are the classifiers themselves that model the fit between low-level features and words.
- *Can the model accommodate uncertainty in the perception of the world?* In a very different way than SIUM, WAC takes uncertainty into account as the classifiers learn what a prototypical object looks like which fits that particular word (e.g., prototypical red objects), but those that are not still fit to a certain degree.
- *How cam the* RE *be represented?* By the nature of the model, we focus on the words of the REs, but this could potentially be applied to a semantic abstraction (e.g., the elementary predicates of RMRS explained in the last chapter could replace words).
- *Is the model robust to noise in the representation of the* RE*?* Each word learns its own mapping from examples of usage; as long as a word is represented enough times as used in a specific type of context, even words that are misrecognised by ASR could learn a meaning.
- *Is the model incremental in that it processes word-by-word without re-computing previous steps?* The two experiments below show that this is indeed the case.
- *Can the model take gaze and deixis into account?* We show in Experiment 1 that models of gaze and deixis can be fused with WAC by interpolating the distributions. Incorporating modalities directly into the classifiers is left to future work.
- *Can the model handle more than one kind of* RE*?* At the moment, WAC can only handle definite descriptions directly, but Experiment 1 shows that it can interpolate deixis.
- *Can the trained classifiers be applied to other tasks?* We show in the experiments how the classifiers can be learned and applied in reference tasks. In the analysis section,

we can already see that isolated classifiers learn the grounded semantics that one would assume they should learn and can already be applied directly to determine if, for example, an individual object fits to an individual classifier. How these classifiers could be used, for example, in generation tasks is left for future work.

In the following sections, we explain two experiments to evaluate the classifiers and models. The first makes use of virtual scenes and simple references, the second makes use of scenes with real, tangible objects where simple and relational references are made. Each experiment explains the data, task, procedure, and metrics for evaluation along with results. General discussion and analysis is left to the end.

6.2 Experiment 1: Resolving References made to Virtual Objects

The experiment presented in this section follows from Kennington et al. (2015a).

6.2.1 Data

The data used in this experiment is the same data as Experiment 4 as explained in the previous chapter (the TAKE data from Chapter 4) which used raw features of objects extracted from distorted images. Example images of pre- and post-processed (distorted) scenes are given in Figures 6.3 and 6.4, respectively. In Experiment 4 in the previous chapter, the raw features were used as features to a classifier that produced a distribution over a pre-defined set of discrete properties. Here, we use the features directly as features to the word classifiers as explained above.

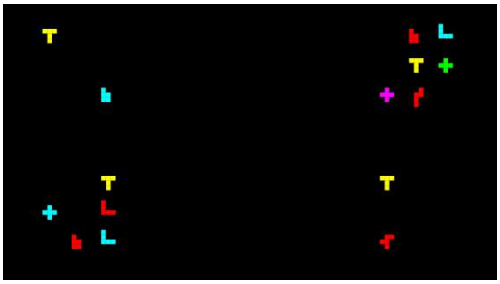


Figure 6.3: Example Pentomino board for gaze and deixis experiment; the yellow T in the top-right quadrant is the referred object.

Figure 6.4: Pentomino Board that has been distorted from its original form (Figure 6.3); all objects have distorted shapes and colours.

6.2.2 Evidence from Gaze and Deixis

In this experiment, similar to Experiment 4 in the previous chapter, the full model of RR combines the evidence from linguistic information with evidence from other information sources such as the speaker's gaze and pointing gestures. For each, we calculate a reference point (R) on the scene: for gaze, the fixated point as provided by an eye tracker; for deixis, the point on the scene that was pointed at based on a vector calculated from the shoulder to the hand (as described in Kousidis et al. (2013), using the Microsoft Kinect). The centroids of all the objects (I) can then be compared to that reference point to yield a probability of that object being 'referred' by that modality (i.e., gazed at or pointed at) by introducing a Gaussian window over the location of the point:

$$p_{distance}(R_i, I_j; \sigma) = \exp -\frac{(x_i - x_j)^2}{2 * \sigma^2} * \exp -\frac{(y_i - y_j)^2}{2 * \sigma^2} \tag{6.9}$$

where the mean is R and σ is set by calculating the standard deviation of all the object centroids and the reference point. This can then be normalised over all the $p_{distance}$ scores to produce a distribution over I for each modality where the closer the object is to the reference point, the higher its probability.

Our final model of RR fuses the the three described modalities of speech, gaze, and deixis using a linear interpolation, where the α parameters are learned from held-out data by ranging over values such that the α values sum to one, and computing the average rank (metric explained below), retaining the α values that produced the best score for that set:

$$P(I|S) = P(I|S_1)\alpha_1 + P(I|S_2)\alpha_2 + P(I|S_3)(1 - \alpha_1 - \alpha_2) \tag{6.10}$$

6.2.3 Task & Procedure

The task is RR. At each increment, the model returns a distribution over all objects; the probability for each object represents the strength of the belief that it is the referred one. The argmax of the distribution is chosen as the hypothesised referent.

Using 1000 episodes, we evaluate our model across 10 folds (900 episodes for training, 100 for evaluation). Our baseline model is a generative model of RR that will be described below (random baseline is 7%). We also incorporate gaze and deixis by treating them as individual

RR models and interpolating their distributions with the distribution given by the model. We ran the experiments twice, once with hand-transcribed utterances as basis for U, and once with ASR output. The α weights (Equation 6.10) for hand-transcribed data were for speech, deixis and gaze: 0.72, 0.16, and 0.12 respectively, and for ASR 0.53, 0.23, 0.24, respectively (note that for ASR, more weight was given to the non-speech models).

For this experiment, we will compare the results of the WAC with the results of SIUM, as explained in Experiment 4 in the previous Chapter (i.e., a distribution over properties was obtained from a classifier that used the raw object features).

6.2.4 Metrics

The metrics for this evaluation are the same as used for SIUM in the previous chapter, repeated here for convenience. We report *accuracy* (how often was the argmax the intended referent) after the full referring expression has been processed.

We also look into how the model performs incrementally; explanation repeated here for convenience. For this, we followed previously used metrics (Schlangen et al., 2009; Kennington et al., 2013), where the predicted referent is compared to the gold referent at each increment:

• **first correct**: how deep into the RE (%) does the model predict the referent for the first time?

• **first final**: if the final prediction is correct, how deep into the RE was it reached and not changed?

• **edit overhead**: how often did the model unnecessarily change its prediction (the only *necessary* prediction change happens when it first makes a correct prediction)?

6.2.5 Results

As Figures 6.5 shows, our model performs well above the SIUM baseline, for all settings. (The *random selection baseline* sits at 7%.) We assume that it performs better not only because as a discriminative model (i.e., it does not need to model the full joint distribution, as was necessary for SIUM), but also because it directly learns a connection between words and visual features and does not need to go through a set of pre-determined features, which could be considered as a "lossy" compression of information. The Figures also show that using ASR output does have an impact on the performance, as expected. The speech+deixis models tend to work better than speech+gaze models in terms of accuracy; we speculate that this is due to the (naive) assumption implicit in our setup that participants gaze at the referred object most of the time, where in fact they often look at distractors, etc., making gaze a noisier model of predicting the referent. Overall, there is about a 6% increase when both modalities are included when using

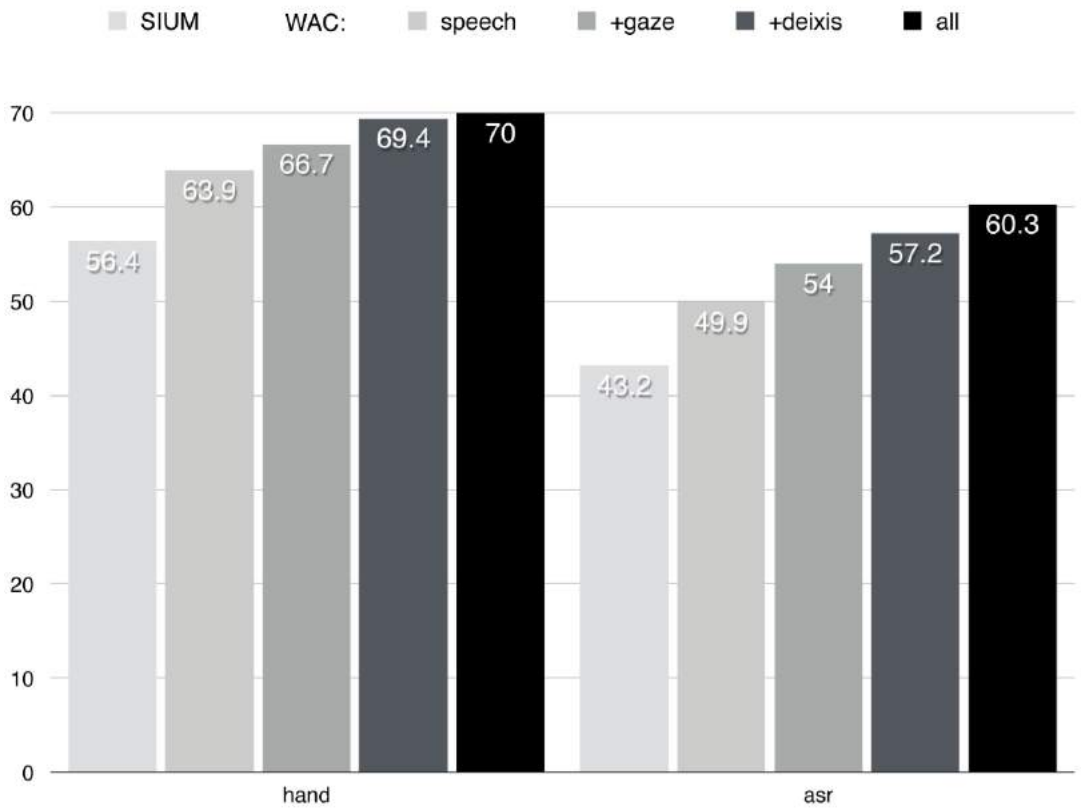


Figure 6.5: Results of our model in accuracies; higher numbers denote better results.

hand-transcribed data. The increase when including both modalities is slightly larger (10%) with ASR. This nicely shows that when given noisier linguistic information, the model can partially recover by taking more benefit from interpolating with other information sources.

6.2.6 Further Analysis

Analysis of Selected Words

We analysed several individual word classifiers to determine how well their predictions match assumptions about their lexical semantics. For example, the classifier for the word *links* (*left*) should yield a high probability when given an object representation where the x-coordinate values are small (i.e., on the left of the screen), and lower probabilities for x values that are high. This was indeed the case, as shown in Figure 6.6. This is a nice feature of the model, as

objects that are in the middle of the scene can still be described as *on the left*, albeit with a lower probability, representing allowance in deviation from the prototypical "left" which fulfils a goal of this thesis. We also tested how well classifiers were learned for colour words. In Figure 6.7 we show how changing the H S V features (representing colours) across the spectrum, keeping all other object features stable, yielded different responses from the classifier for the word *gelb* (yellow), where the y-axis on the figure represents probability for that particular colour value.[2]

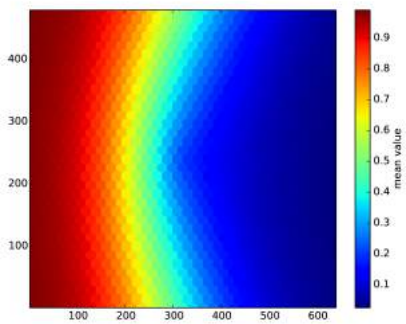


Figure 6.6: Strength of word *links* (German for *left*) predicting when given different x-coordinate values; higher values (in red) show that the classifier for *links* yields higher probabilities as the objects are farther to the left.

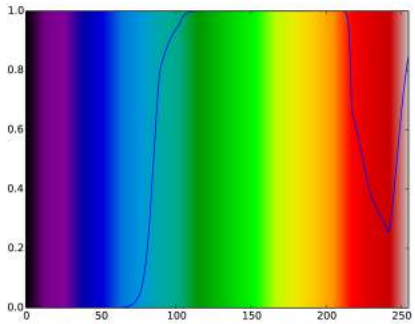


Figure 6.7: Strength of word *gelb* (German for *yellow*) predicting when given different colour (HSV) values; there is some overlap with green and red, but high probabilities are yielded for the classifier when near the yellow values.

We further looked into shape words. Figure 6.8 shows the response of the classifier for *kreuz* (cross) when given object representations where only the shape-related features (number of edges, skewness) were varied across all possible shapes (the x-axis uses here the standard labelling of pentomino pieces with letters whose shapes are similar). Interestingly, the classifier generalised the word to apply not only to objects with the cross shape, but also the Z-shape piece (the red piece in the bottom of the top right group in Figure 6.3) and others which also intuitively seem to be more similar. For a sanity check, we looked at the responses to change in colour for the word *kreuz*. As Figure 6.9 shows, this classifier does not pick out any specific colour, as it should be. This shows that the word classifier managed to identify those features that are relevant for its core meaning, ignoring the others. Further analysis is provided in Chapter 7.

[2]There is also a high probability in the black region. It could be the case that the yellow classifier learned that a low value for B is highly discriminative (black is 0 for all RGB values).

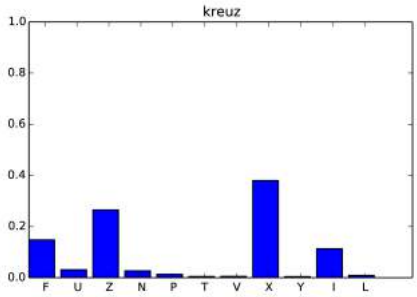


Figure 6.8: Strength of word *kreuz* (German for *cross*) predicting when given different values of number of edges.

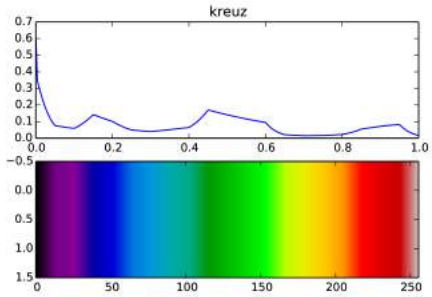


Figure 6.9: Strength of word *kreuz* (German for *cross*) predicting when given different colour (RGB) values.

6.3 Experiment 2: Resolving References made to Tangible, Real-World Objects

For all experiments in Chapter 5 and Experiment 1 in this chapter, we have only considered scenes where the objects are virtual objects appearing on a computer screen and those objects are represented symbolically without any uncertainty in their representation. In Experiment 4 of Chapter 5 and Experiment 1 of this chapter, we used a distorted version of the TAKE corpus to provide a more realistic setting where the virtual objects could be recognised with varied levels of uncertainty. In this final experiment of this thesis, we apply the WAC model to data where the objects are no longer virtual–they are real, tangible objects that can be manually manipulated.

6.3.1 Data

This experiment uses the TAKE-CV data described in Chapter 5 which consisted of 870 usable episodes. All episodes had a target referent but, 460 had references which included landmarks. An example scene is depicted in Figure 6.10. All REs were automatically transcribed using Google ASR.

6.3.2 Task & Procedure

We break down our data as follows: episodes where the target was referred directly via a 'simple reference' construction (DD; 410 episodes) and episodes where a target was referred

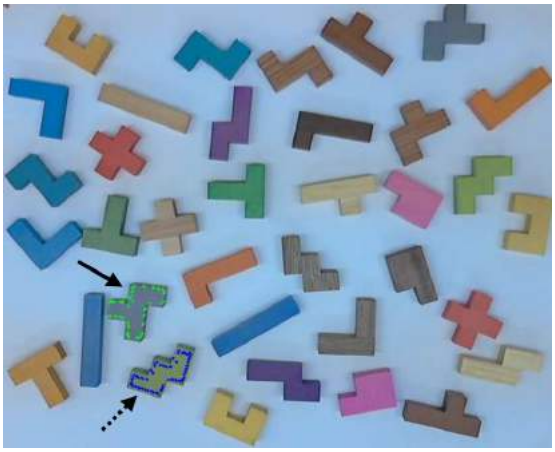


Figure 6.10: Example episode for *phase-2* where the target is outlined in green (solid arrow added here for presentation), the landmark outlined in blue (dashed arrow).

via a landmark relation (RD; 460 episodes). We also test with either knowledge about structure (simple or relational reference) provided (ST) or not (WO, for "words-only"). All results shown are from 10-fold cross validations averaged over 10 runs; where for evaluations labelled RD the training data always includes all of DD plus 9 folds of RD, testing on RD. The sets address the following questions:

- DD.WO: how well does the sr model work on its own with just words?
- DD.ST: how well does the sr model work when it knows about REs (i.e., has some notion of syntactic structure)?
- RD.ST (sr): how well does the sr model work when it knows about REs, but not about relations?
- RD.ST (r): how well does the model learn relation words after having learned sr?
- RD.ST with DD.ST (rr): how well does the rr model work (together with the sr)?

Words were stemmed using the NLTK (Bird, 2006) Snowball Stemmer, resulting in a vocabulary size to 1306. Due to sparsity, for relation words with a token count of less than 4 (found by ranging over values in a held-out set) relational features were piped into an UNK relation, which was used for unseen relations during evaluation (we assume the UNK relation would learn a general notion of *nearness*). For the individual word classifiers, we always paired one negative example with one positive example.

For this evaluation, word classifiers for sr were given the following features: RGB values, HSV values, x and y coordinates of the centroids, euclidean distance of centroid from the

centre, and number of edges. The relation classifiers received information relating two objects, namely the euclidean distance between them, the vertical and horizontal distances, and two binary features that denoted if the landmark was higher than/lower than or left/right of the target.

6.3.3 Metrics

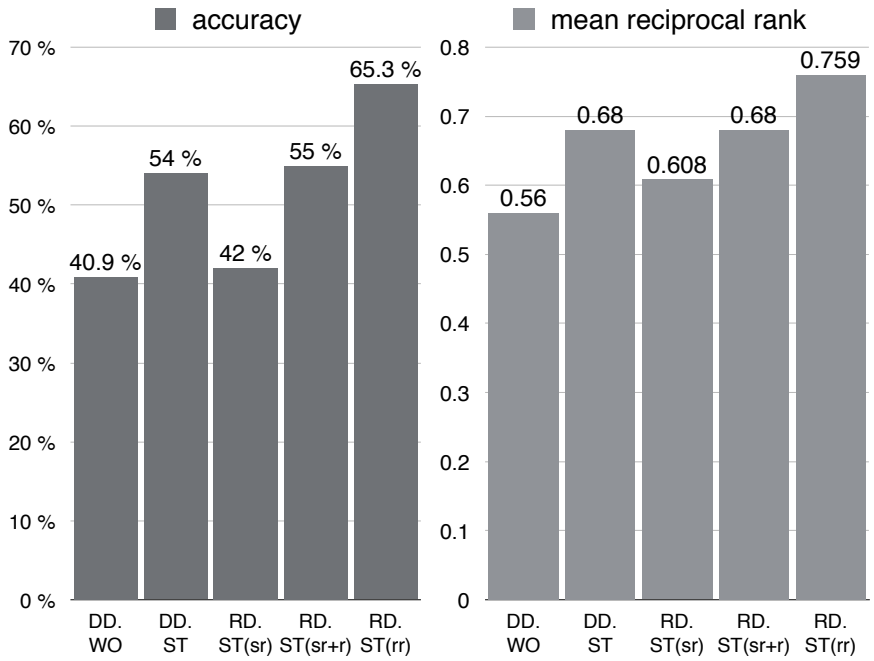


Figure 6.11: Results of evaluation in accuracy and MRR.

The RE-level metrics use accuracy and mean reciprocal rank. For incremental metrics, we show how well the model performs on a subset of episodes by calculating the average rank of the gold referent; the lower the rank the better, where 1 is the best possible average rank.

6.3.4 Results

Utterance-level Results

Figure 6.11 shows the results. (Random baseline of 1/32 or 3% not shown in plot.) DD.WO shows how well the *sr* model performs using the whole utterances and not just the REs.[3] DD.ST adds structure by only considering words that are part of the actual RE, improving the results

[3]Note that all evaluations are on noisy ASR transcriptions.

further. The remaining sets evaluate the contributions of the rr model. RD.ST (sr) does this indirectly, by including the target and landmark simple references, but not the model for the relations; the task here is to resolve target and landmark SRs as they are. This provides the baseline for the next two evaluations, which include the relation model. In RD.ST (sr+r), the model learns SRs from DD data and only relations from RD. The performance is substantially better than the baseline without the relation model. Performance is best finally for RD.ST (rr), where the landmark and target SRs in the training portion of RD also contribute to the word models.

The *mean reciprocal rank* scores follow a similar pattern and show that even though the target object was not the argmax of the distribution, on average it was high in the distribution. For all evaluations, the average standard deviation across the 10 runs was very small (0.01), meaning the model was fairly stable, despite the possibility of one run having randomly chosen more discriminating negative examples. Our conclusion from these experiments is that despite the small amount of training data and noise from ASR as well as the scene, the model is robust and yields respectable results.

Incremental Results

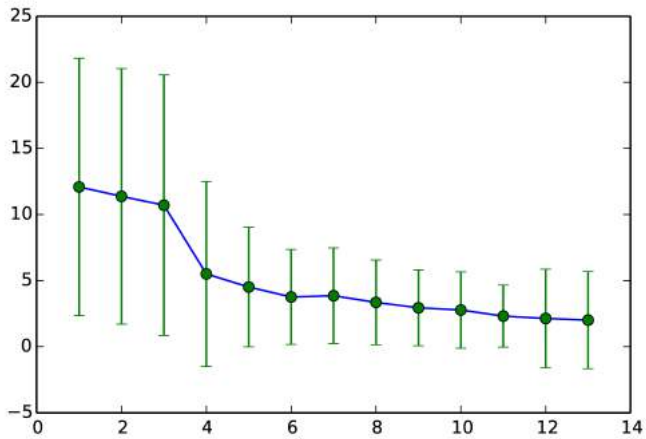


Figure 6.12: Incremental results: average rank improves over time

Figure 6.12 shows how our rr model processes incrementally, by giving the *average rank* of the (gold) target at each increment for the REs with the most common length in our data (13 words, of which there were 64 examples). A system that works incrementally would have a monotonically decreasing average rank as the utterance unfolds. The overall trend as shown in that Figure is as expected. There is a slight increase between 6-7, though very small (a differ-

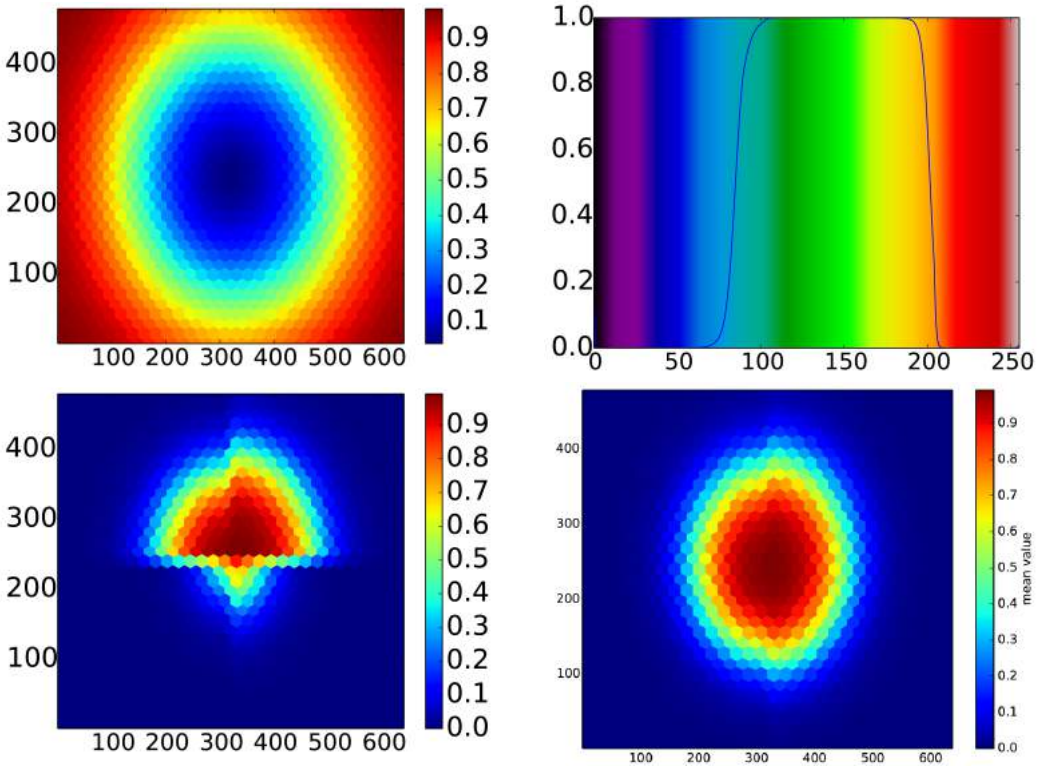


Figure 6.13: Each plot represents how well selected words fit assumptions about their lexical semantics: the top left plot *ecke* (*corner*) yields higher probabilities as objects are closer to the corner; the top right plot *grün* (*green*) yields higher probabilities when the colour spectrum values are nearer to green; the bottom left plot *über* (*above*) yields higher probabilities when targets are nearer to a landmark set in the middle; the bottom-right plot shots the UNK relation where an object set in the middle would have higher probabilities to other nearby objects.

ence of 0.09). Overall, these results seem to show that our model indeed works *intersectively* and "zooms in" on the intended referent.

6.3.5 Further Analysis

Analysis of Selected Words

As with Experiment 1, we analysed several individual word classifiers to determine how well their predictions match assumptions about their lexical semantics. For example, for the spatial word *ecke* (*corner*), we would expect its classifier to return high probabilities if features related to an object's position (e.g., x and y coordinates, distance from the centre) are near corners

of the scene. The leftmost plot in Figure 6.13 shows that this is indeed the case; by holding all non-position features constant and ranging over all points on the screen, we can see that the classifier gives high probabilities around the edges, particularly in the four corners, and very low probabilities in the middle region. Similarly for the colour word *grün*, the centre plot in Figure 6.13 (overlaid with a colour spectrum) shows high probabilities are given when presented with the colour green, as expected. Similarly, for the relational word *über* (*above*), by treating the centre point as the landmark and ranging over all other points on the plot for the target, the *über* classifier gives high probabilities when directly above the centre point, with linear negative growth as the distance from the landmark increases.

Also included is the UNK relation classifier. When words that were marked as relations had a count of 4 or less, they were treated as UNK words. The graph shows that our assumptions were correct in that the UNK relation learned that, given a landmark object set in the middle and ranging over all other points for the target object, a notion of *nearness* is learned. Thus, if an unknown relation word (i.e., was not in training) is uttered and marked as a relation word, the model is better off knowing that the two related objects are near each other, even if a more specific direction of that nearness (e.g., left of, above, etc.) is not known.

Note that we selected the type of feature to vary here for presentation; all classifiers get the full feature set and learn automatically to "ignore" the irrelevant features (e.g., that for *grün* does not respond to variations in positional features). They do this quite well, but we noticed some 'blurring', due to not all combinations of colours and shape being represented in the objects in the training set.

Analysis of Incremental Processing

(1) a. *grauer stein über dem grünen m unten links*
b. grey block above the green m bottom left
c. tc ts r l lc ls tf tf

Figure 6.14 finally shows the interpretation of the RE in Example (1) in the scene from Figure 6.10. The top row depicts the distribution over objects (true target shown in red) after the relation word *unten* (bottom) is uttered; the second row that for landmark objects, after the landmark description begins (*dem grünen m* / *the green m*). The third row (target objects), ceases to change after the relational word is uttered, but continues again as additional target words are uttered (*unten links* / *bottom left*). While the true target is ranked highly already on the basis of the target SR alone, it is only when the relational information is added (top row) that it becomes argmax.

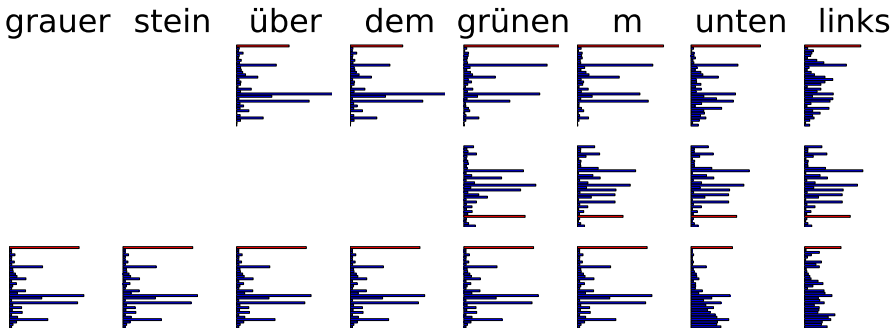


Figure 6.14: A depiction of the model working incrementally for the RE in Example (1): the distribution over objects for relation is row 1, landmark is row 2, target is row 3.

Analysis of Learning Curve

Figure 6.15 shows how well the classifier for *grün* (*green*) learned its lexical semantics after varied amounts of training data. After only a single training instance (top-left) the classifier had no chance of learning how to distinguish green objects from others. After 10 training samples (top-right), the classifier already does a better job–at the least, it can distinguish green objects from purple and blue ones, with some confusion in light blue, yellow, and some red. With 50 training samples (bottom-left), the classifier is still improving. After 100 training samples (bottom-left) the model is able to determine if an object is green, with some confusion in yellow objects and some confusion in blue objects, which might be acceptable as humans also have similar problems. What has not been shown is the set of negative training examples that were randomly chosen to help the classifier distinguish between green and non-green objects; the choice of objects could certainly make a difference in how well the classifier works.

Additional Error Analysis

Some of the individual classifiers did not learn their semantics that one might assume. After being trained on a fold of data, Figures 6.16 and 6.17 show cases where the semantics for *brown* and *blue* (respectively) were not learned as strong as for the colours shown above; brown only yields high probabilities for a very dark, narrow region, and blue would yield fairly high probabilities for objects that one would perceive as green. As shown above, it doesn't generally take much training data to learn colour classifiers, but more training examples is generally better as it makes the classifiers better at distinguishing between features.

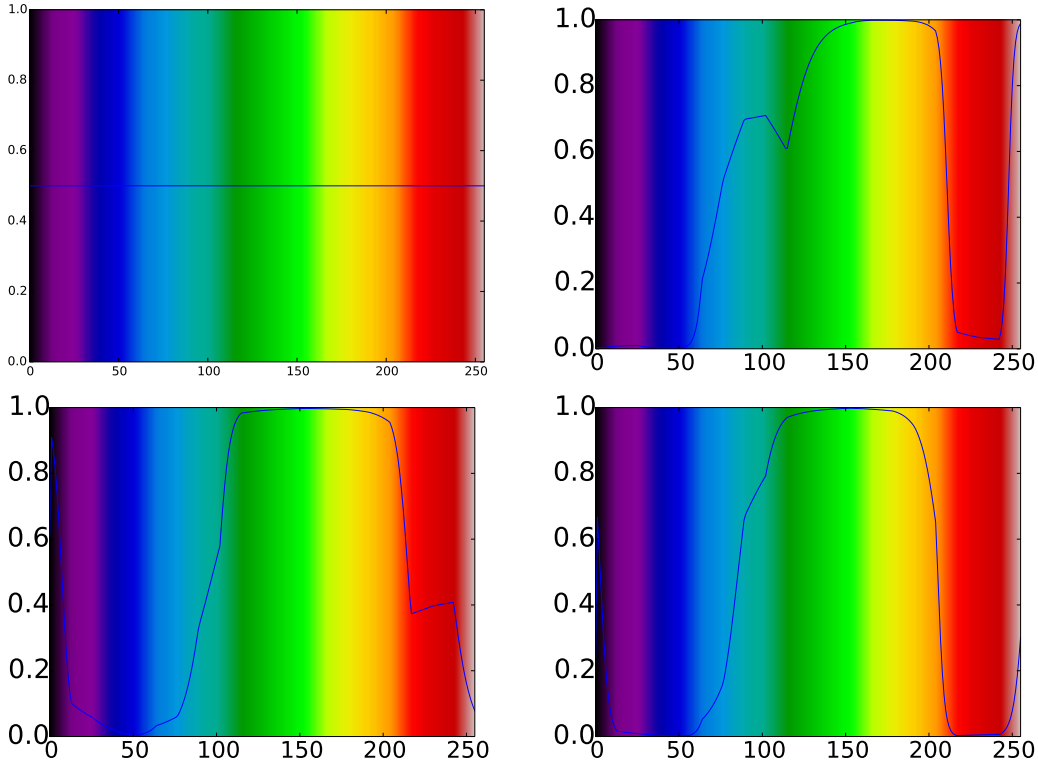


Figure 6.15: Each plot represents how well *grün* (*green*) is learned for varied amounts of training; top-left: 1 training instance, top-right: 10, bottom-left: 50, bottom-right: 100.

6.4 Discussion

The WAC model presented in this chapter is arguably a simpler approach to RR than SIUM in that there is no complicated factorised joint distribution, rather it is a simple mapping between words and objects (represented as low-level features). The model works well on the data it has been tested on, despite realistic, noisy conditions of the scene representation and RE representation. Clearly, it is a grounded model in that the classifiers themselves are given low-level features. Importantly, the model works incrementally.

Strictly speaking this model follows a linguistic relativistic notion in that the concepts that human can consider mentally are based on language. Indeed, the classifiers themselves are trained and applied such that they are independent of all others. This isn't the intent of the model; it just works out that way because of the way it is formulated. We leave for future work a formulation that can represent concepts that might not be directly linked to words as well as

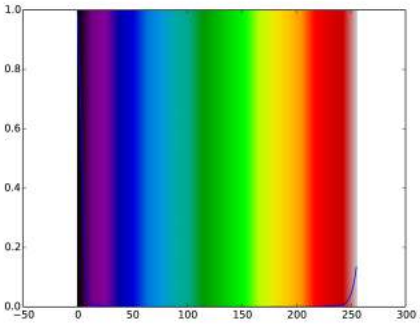


Figure 6.16: Strength of word *brown* (German for *brown*) predicting with different colour values.

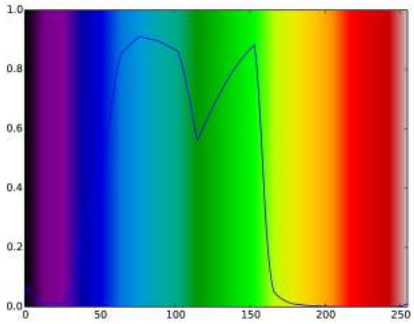


Figure 6.17: Strength of word *blau* (German for *blue*) predicting with different colour values.

a conceptual link between words and concepts.

The WAC model has been implemented as a component in an interactive dialogue system as described in Kennington et al. (2015c), albeit only using simple references. Details of the implementation are explained in Appendix B.

6.5 Intension, Extension, and the Words-as-Classifiers Model

The approach to resolving REs presented in this chapter is clearly different from that of the previous chapter. Here, there are no rigid properties which make up a set of concepts to which all words must be mapped. Rather, the words themselves are the concepts/classes. More specifically, the classifiers correspond more closely to the intension of a word, which for example in Montague's approach is similarly modelled as a function, but from possible worlds to extensions (L.T.F. Gamut, 1991). The process of normalising over objects given a word can then be seen as the extension of the word in a given (visual) discourse universe of the candidate objects.

This can be reflected in an example we saw in Chapter 2:

(2) The morning star is the evening star.

Where though *morning* and *evening* would be represented by different classifiers (intensions), the entire composed REs would both still refer to the same object.

With the WAC model, the *meaning* (i.e., intension) of a word is somehow the classifier itself. Following Carnap's definition of intension and extension (again, see Chapter 2), the classifiers

can be seen as assigning objects to classes. Indeed, by definition, each word's classifier has a binary decision to make given an object: does this object belong in my class (i.e., does this object "fit")? Albeit, this binary "decision" is a soft decision as each word's classifier assigns a probability to an object based on how well it fits the word (i.e., the conceptual class; e.g., the classifier for *red* can assign objects to the class of red things by a probability).[4]

The WAC model also fits well our simple formal framework laid out in Chapter 2. Whereas SIUM needed a discrete set of concepts represented by properties, WAC needs no such rigid set; rather, they are the words themselves. Thus ϕ_w in 6.11 represents any word classifier. The variable x are the object features, thus the classifier for ϕ_w (for a word w) returns a probability that x is a member of that class of objects. These can then be part of larger FOL expressions, linking (albeit in a small way) grounded semantics with a formal logical calculus, such as was explained in Chapter 2.

$$[\![w]\!]_{obj} = \lambda x.\phi_w(x) \tag{6.11}$$

This gives evidence that the WAC model fulfils the two goals of this thesis that SIUM was not able to, namely to account for word meaning, and to fit those meanings in a larger framework. This model effectively overcomes the shortcomings of FOL explained in chapter 2:

1. the set of classes must be determined
2. the functions that assign objects to those classes must be determined
3. incremental composition

As noted, the words are the classes and the classifiers representing those words determine class membership. During application, the model resolves references made to visually present objects in an incremental fashion, thereby overcoming, to a certain degree, those shortcomings.

6.6 Chapter Summary

This chapter introduced and explained the *words-as-classifiers* model of reference resolution. The model was described as being a function from object features to a decision of how well those features fit a given word. Two types of words were modelled: single and relational.

[4] Arguably, a threshold for each word's classifier would need to be determined for class membership. However, it may not be necessary in practice; certainly it wasn't the case for the experiments presented in this chapter–the probabilities were informative enough.

Training and application of the model in single and relational types of referring expressions was explained. The model was evaluated in two experiments. Experiment 1 used the TAKE data and the model was interpolated with gaze and deixis. Experiment 2 applied to model to real-world, tangible objects. Both experiments showed the robustness of the model to noisy conditions coming from automatically-recognised speech and from the object features. The model is theoretically pleasing, as it is compositional, also theoretical notions of intension and extension are expressed via the model. It better fits into our formal framework laid out and overcomes the shortcomings of FOL as explained in Chapter 2

7

Closing Remarks: Comparisons and Outlook

In this closing chapter of this thesis, we first look at a comparison between the SIUM and WAC models presented, respectively, in Chapters 5 and 6. We compare them on a general, theoretical level, then look at how they compare in a RR task including RE-level and incremental-level comparisons. The comparisons are then followed by a conclusion to this thesis, which gives a recap of what we have shown. That is followed by a section on what further work needs to be done, and some parting thoughts on meaning and reference.

7.1 Comparing SIUM and WAC

In this section we take look at how SIUM and WAC compare in RR tasks. We specifically look at the episodes in the TAKE corpus where each model produced distributions where the referent was not the argmax by the end of the RE. We choose the TAKE corpus because we have symbolic properties as well as low-level feature representations of the scenes (though the symbolic version was what was used to obtain the REs when collecting the data).

7.1.1 General Comparison

We saw in Chapter 5 that SIUM is a generative, grounded model that learns from examples of language use and works in an update-incremental fashion. If done correctly, it can resolve

demonstratives, pronouns, and definite descriptions. It can also make use of contextual priors (i.e., saliency). In Chapter 6, we saw that WAC is discriminative, learns the semantics of individual words directly from low-level features, and can handle more complicated utterances with relations.

In principle, the two models are quite similar in that there is a mapping from objects to words. For WAC, the mapping is done directly. For SIUM, the mapping is done via a mediating variable for a set of properties. Both are update models in terms of incremental processing.

There are some programmatic differences, however. WAC can handle relational REs, where SIUM can only handle simple ones, yet SIUM can handle demonstratives, pronouns, and definite descriptions, whereas WAC can only handle the latter (i.e., directly; though see Experiment 1 of Chapter 6 where deixis was incorporated as a separate model).

These differences beg the question as to how the models compare in RR tasks and if they could benefit from being combined (e.g., by weighted interpolation). This is explored presently.

7.1.2 Comparison on Reference

Using the TAKE data, we randomly selected 100 episodes for evaluation and trained on the remaining 900. For SIUM, we used the properties (as the TAKE corpus is of virtual scenes, we know the colour, shape, quadrant, etc.), and for WAC we used raw features extracted from the distorted scenes. For this comparison, we are interested in the episodes where the models referred to the wrong object (i.e., the argmax of the distribution was not the referred object). We first look at the scenes to determine if the complexity of a scene contributed to the difficulty of resolving the referent. We then look at the REs.

We first look at the episodes that both models got wrong. There were 11 such episodes (out of 100); all 11 episodes had at least 1 other colour distractor (i.e., there was always at least one other object in the scene with the same colour as the referent), 7 scenes had distractor shapes, and all 11 had distractor objects in the same quadrant. In only one case did all three properties (colour, shape, and quadrant) match a distractor (there were 5 such cases in total in the evaluation set; i.e., 4 such cases were correct by both models).

There were 16 cases where SIUM was correct and WAC was wrong. Of those, there were 2 with no colour distractors, 3 with no shape distractors (each of those three were different shapes).

There were 5 cases where WAC was correct and SIUM was wrong. Of those 5 , there were always colour, shape, and quadrant distractors.

From these comparisons, the complexity of the scene doesn't appear to contribute to whether or not a particular model got them right based on confusability with other objects. At least, not in an obvious comparison like those done as just described.

Some patterns do emerge when looking at the REs, however. For the 11 episodes where both models are wrong, the average number of words is 21.2. Where SIUM was correct and WAC was wrong (16 episodes), the average number of words is 12.3. Where WAC was correct, SIUM wrong (5 cases), the average number of words is 15. The average number of words for the 68 episodes where both models were correct is 10.6. Thus it appears that the main contributing factor to getting an episode correct is the number of words; once extra, unneeded words are added to the RE, it confuses both models (though that confusion happens earlier for WAC). One participant in particular, who accounted for 7 of the 11 episodes that both models got wrong, had a much higher average number of words per RE than other participants.

Below are some example REs (confirmations are left out, as they don't contribute to the REs; English glosses provided in parentheses that follow each example), (1) where WAC was correct and SIUM wrong; (2) where SIUM was correct and WAC was wrong, and (3) where both were wrong.

(1) a. oben links das linke von den beiden unteren (top right the left from the two lower)
b. grün ganz oben rechts (green completely top right)
c. die rechte rote form die form ganz rechts in rot (the right red figure the figure completely right in red)

(2) a. unten rechts das grüne (bottom right the green)
b. dazu dann dieses hellblaue plus ähm nicht pluszeichen sondern das t das hellblaue t hier (thereto then this light-blue plus um not plus-symbol rather the t the bright-blue t here)
c. das einzige graue symbol links unten ja (the only gray symbol left below)

(3) a. okay oh wir haben so viele hier nimm mal bitte das lila c (okay oh we have a lot here take now please the purple c)
b. hmm lila oben rechts da mh wähle (hmm purple top right there um take)
c. oben rechts das türkise in der mitte (top right the turquoise in the middle)

These comparisons show that the two models have similar strengths and weaknesses when applied to the TAKE data. It is therefore not apparent that combining the two models in a single task would provide any benefit. It should be noted, however, that the SIUM model did have access to the properties with certainty, whereas the WAC model used low-level features from distorted images. The conclusion here is, if a dialogue system designer needs a RR component and has a virtual scene, then SIUM would be a better choice. Also, if the designer needs a framework that can use properties to incorporate additional modalities and exophoric pronouns, SIUM is the preferred choice. However, if the scene has representations of real, tangible objects

in different lighting conditions, then WAC would be a better choice, even though SIUM can be made to handle uncertainty in the representations, provided the low-level features are abstracted over as discrete properties. Moreover, if a designer has more complex REs, such as those with relations between objects, then WAC is the preferred model.

7.1.3 Learning Curves

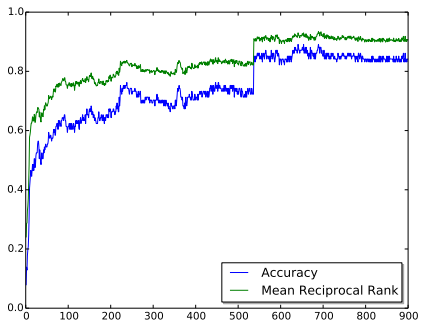


Figure 7.1: SIUM learning curve; x-axis is amount of training data, y-axis is accuracy. Accuracy is always lower than MRR.

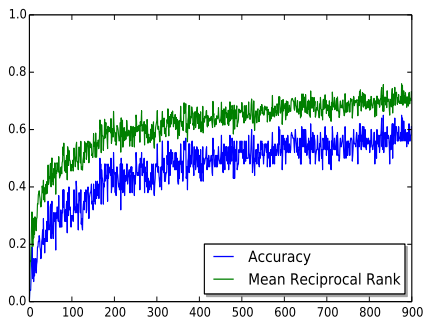


Figure 7.2: WAC learning curve; x-axis is amount of training data, y-axis is accuracy. Accuracy is always lower than MRR.

Figures 7.1 and 7.2 show how well SIUM and WAC perform on a RR task, respectively, with varied amounts of training data. The x-axis shows how much training data was applied (1-900 episodes, again from the TAKE data), the y-axis shows how well each model predicts the referent in 100 episodes, given that amount of training data. As expected, both models improve as training data is added. For SIUM, over 60% accuracy is achieved with only 100 training samples; for WAC, it takes 200 in order to reach around 50%. The comparison isn't completely fair as the SIUM has an easier job to do: SIUM is using the original scenes represented as properties whereas WAC is using the distorted scenes, represented as noisy low-level feature. The dramatic jump in accuracy for SIUM around 550 is explained by an addition of a training sample that happens to be more useful to the evaluation set; however, notice that there are no further improvements overall after that point as the model appears to level off. WAC, on the other hand, continues to improve as the amount of training data is increased.

It was shown in Chapter 6 that the individual word classifiers for the WAC model were able to at least partially discriminate their individual word semantics with as little as 10 training episodes (at least for colour words).

Incremental Comparison

% edit overhead		
utt length	WAC	SIUM
1-6	11.5	3.8
7-8	19.76	17.2
9-14	41.0	27.5
% never correct		
utt length	WAC	SIUM
all lengths	19.5	32.0

Table 7.1: % edit overhead and never correct

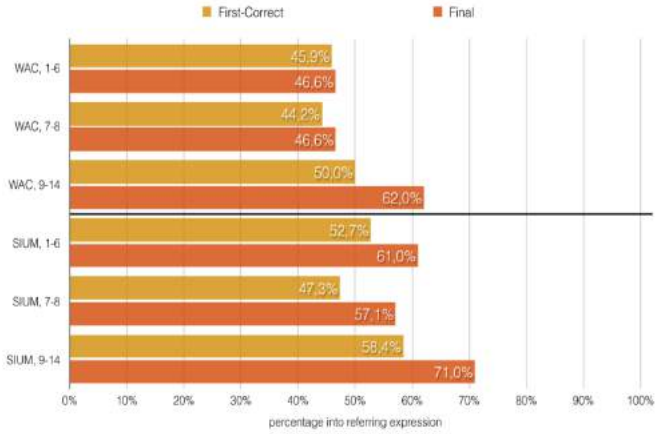


Figure 7.3: Incremental Performance

Figure 7.3 gives an overview of the incremental performance of the two models. Results here are for the hand-transcribed utterances of the distorted TAKE data. As the metrics talk about "% into expression", these metrics can of course only be computed when the eventual length of the expression is known, that is, after the fact. Moreover, to make these units comparable (as "10% into the utterance" is very different in terms of words for an utterance that consists of 2 words than for one that is 12 words long), we bin REs into the classes *short*, *normal*, *long* (1-6, 7-8, 9-14 words, respectively, as to group together REs that are of similar length).

Ideally for use downstream in a dialogue system, the reference resolver would make a first correct decision quickly, and this would also be the final decision (that is, in the graph the two bars would be close to each other, and at a low value). As the Figure shows, WAC is somewhat earlier than SIUM and on average that decision is very close to the final decision it makes, but it pays for this in a higher edit-overhead. When looking at first-final, SIUM is not that far away from WAC, but nevertheless, WAC beats the baseline across the board (and, as shown in Table 7.1, is also correct more often). WAC produces less stable predictions, which is presumably due to its response to the expression being a simple summation of the responses to its component words, whereas SIUM is a proper update model.

7.1.4 Semantic Comparison

Looking again at our original formula,

$$[\![w]\!]_{obj} = \lambda x.\phi_w(x) \tag{7.1}$$

The job of ϕ_w is to assign an object represented by features x to the class that ϕ_w represents. For example, if ϕ_w is red, then it determines if x is assigned to the class of red things. This fits closely to the notion of *intension* as we discussed in Chapter 2.

For SIUM, ϕ_w is not necessarily learned, rather all words in REs are mapped to a smaller set of properties, each of which represents its own ϕ_w. SIUM doesn't directly assign objects to classes, rather, it assigns words to classes through the $P(R|U)$ part of the model which results in a distribution over properties given words in the RE. For example, if *red* is uttered, a distribution over all properties is produced, presumably with the `red` property with the highest probability. This stochastic mapping between words and concepts (i.e., properties) is where the grounding takes place. However, there is no notion of assigning objects to their respective classes. That is done in $P(R|I)$, but in the cases where there is no uncertainty, objects are assigned classes based on rules (e.g., how a virtual game board is laid out). The set of concepts must also be pre-defined.

For WAC, the concepts *are* the words (i.e., w), thus no set of concepts need to be pre-defined. Thus each individual word can replace ϕ_w and the classifier for that word determines how well an object x fits into the class or concept represented by the word. Thus WAC, in a small way, fulfils Harnad's theory that bottom-up perceptual approaches can be "hooked up to symbolic" ones in a more direct way than SIUM. This being the case, then WAC could potentially be fit into larger semantic frameworks. Certainly, not all words are perceptually-grounded. Other words that are based on the meanings of other words which are not grounded in visual perception could also be represented in a word-as-classifier approach, provided the features for such a meaning exist, as they did for visually present objects.

Composition for the two models (i.e., the *and* $\wedge$ operator in FOL) is different for the two models. For SIUM, the nature of the prior $P(I)$ necessitates a multiplicative composition as it updates the distribution from word to word. The WAC model uses an additive composition (i.e., a weighted interpolation in the form of an average) for the simple references and multiplicative for the relational model. These constituted a simple composition that works with this kind of intersective semantic composition as explained in Chapter 2. We leave composition of more complex semantic phenomena to future work.

7.2 Conclusion

We began this thesis with a look at the background of reference and meaning, beginning with SDSs, where a RR component would fit into a SDS followed by a general approach to RR using FOL. The shortcomings were noted (i.e., defining the set of classes, how to assign objects to those classes, and how to perform RR incrementally), and partially addressed by an appeal to grounded semantics. We then toured through the philosophy on how meaning and reference are related, finding motivation on using reference as a task for learning meaning. We set out to overcome the shortcomings of FOL and traditional approaches to RR by automatically determining the classes that objects could be assigned to, using a learned, grounded semantics that could determine if an object belongs to a specific class, and performing these tasks incrementally.

We reviewed the relevant literature with this in mind, noting that such an approach to incremental reference resolution and meaning representation had not yet been accomplished. Hence, the need for the work in this thesis. We then turned to a generative approached in the *simple incremental update model* of RR. We saw that SIUM worked well when varying different aspects of the model, from how the REs were represented, to how the world was represented, and how the two were connected. Though the model performed well in several experiments ranging over 3 different languages and despite the fact that the model does perform a kind of grounding, the model did not quite satisfy the original problem of determining the set of possible classes (rather, the model required a set of properties to be pre-defined), nor did it quite work as a mechanism of assigning objects to those classes (though it did to some degree), but SIUM improves over previous work in that it works in an update-incremental fashion. The model was also able to handle the two types of REs in which we were interested: definite descriptions and demonstratives (as well as, to a lesser degree, exophoric pronouns).

We then turned to the *words-as-classifiers* model which mapped directly from low-level object features to words; thus the need for mediating properties was removed. The set of classes is determined by the language itself as the words are treated as individual classes. The model represents each word as a classifier, which acts as the mechanism for probabilistically assigning objects to classes. The model further specified how the classifiers were to be applied in a reference resolution task and processed in an update-incremental fashion. It works robustly when used for definite descriptions (including relational REs). Though it can be incorporated with a separate model for resolving deixis (demonstratives, in a general way), we leave modelling demonstratives in WAC for future work.

Given the experiments and evaluations in Chapters 5 and 6, we can conclude that the overarching goal of modelling and implementing a practical component of incremental RR has been

thesis aim	SIUM	WAC
the model can resolve referring expressions update-incrementally	✓	✓
the model learns, given data, a mapping between visually present objects	✓	✓
given novel REs and novel scenes, the model can generalise	✓	✓
the model can handle definite descriptions	✓	✓
the model can handle demonstratives	✓	
the model can be implemented as a component in a SDS	✓	✓
the model can be evaluated to show that it can handle noise in ASR and scene	✓	✓
formulate the model in such a way that word meanings can be accounted for		✓
fit the model into a semantic framework		✓

Table 7.2: Thesis aims addressed by SIUM and WAC.

realised. For completeness, Table 7.2 shows a listing of these aims and whether SIUM or WAC realises that particular aim. Both SIUM and WAC work in an update-incremental fashion. Both models work robustly against noise from ASR due to spontaneous (though in somewhat limited tasks), often ungrammatical speech. The WAC model further performed well under noisy conditions of representing the objects. Both models can be fused with separate models of deixis and gaze, and SIUM can incorporate deixis and gaze information as properties. Both models ground, in their own way, aspects of REs with aspects of the world (by object properties for SIUM or by low-level object features for WAC). Both models are fast and can perform their tasks in real time. These models improve upon previous work which was either not grounded, not incremental, or could not function in real time.

A lesser goal of this thesis is to fit these models into a larger semantic framework which is shown above in the semantic comparison. With the *words-as-classifiers* model, we substantiate the claim made in Dahlgren (1976):

> Extensions determine intensions, though in a complex way, and not the other way around.

The work presented in this thesis, I assert, brings us a step closer to understanding meaning of words (i.e., visual words) and how that meaning is derived from interaction with the world. Specifically, through the *words-as-classifiers* model, we have estimated meanings of words using the features of objects referred to in the real world. Furthermore, the model is not just a theoretical model, rather it has been implemented and tested in RR tasks with some degree of success.

7.3 Further Work

Though the two models presented in this thesis go beyond previous work in several respects, they by no means completely handle all kinds of REs under all circumstances. Both models, for example, assume that a different module does the difficult, yet important task of segmenting the REs (including quantifier scope) identifying that they are in fact REs, and identifying possible relations between them, is left to another component entirely. At the moment, only WAC can handle REs that have relations, though neither model has, at the moment, any way of handling negation. There are other important aspects of language that can be found in REs that the models might not handle directly, but I think that these models present a way of resolving references to a fair portion of the kinds of REs that people come across day-to-day.

Specifically, we are interested in looking further into the following:

- *Compositionality*. Composition for both models happens on the level of extension, not necessarily on the level of composing words into phrases, etc. Improved composition would mean composing the way the meanings are represented, e.g., through a semantic calculus.

- *Quantification*. At the moment, we are assuming reference to a single, visually present object. Definite descriptions preface such a reference type with the word *the*. But other words denote different types of reference, for example *a* implies that any object falling in a certain class (or set of classes) can be referred, *all* implies that all objects falling in a certain set of classes be referred, or even specific numbers like *two*, etc.

- *Language Generation*. Being a generative model, SIUM is a natural candidate for potential NLG research; in particular for generation of REs. It was mentioned in Chapter 3 how SIUM is similar in principle to the NLG model in Mast et al. (2014), but actually using SIUM in a NLG task is left for future work. We have looked into using the WAC classifiers for generation of REs by clustering the classifier coefficients with some initial promising findings.

- *Reference Domains*. At the moment, the two models require that the reference domain, i.e., the set of candidate objects, be pre-specified and visually present. More would could be done to relax this; a reference could be made to an object that is not visually present, but later perceived.[1] Also, abstract entities that aren't visually present, but represent an object that could be imagined (e.g., a unicorn).

[1]Though see recent work in this area that uses both SIUM and WAC models in Han et al. (2015).

- *Non-visual Reference*. Though the focus of this thesis has been to visually present objects, the models could be usable in referring to entities that are not perceivable, at least in the way objects are. For example, referring to a particular person, city, or idea (e.g., health or democracy). This could potentially be done using SIUM if there are properties to each entity, and those properties (though not visual; e.g., for a city, a property could be the country it is in, it's population, nicknames, etc.) could be learned to ground with certain words that refer to those entities. For WAC, the features would need to be determined. Possibly fitting WAC into a formal framework would also allow it to be usable in more abstract situations like referring to non-visual entities.

- *Demonstratives (and Pronouns)*. While SIUM can handle the types of REs that we are interested in, at the moment WAC can only handle definite descriptions. It could be made to handle demonstratives by incorporating additional features (e.g., that a hand is pointing, and coordinates to where it is pointing, etc.) and it is unclear as to how pronouns could be incorporated, though a binary feature like that which was used for SIUM could also be used.

- *Further Fitting into a Formal Framework*. WAC represents individual words as classifiers. We have seen how application of an object to those classifiers can happen using lambda calculus. However, those classifiers in turn could be fit into a larger semantic framework, thus combining the benefits of the WAC model for grounded semantics, as well as formal frameworks which provide necessary scoping, relations, etc, without, for example, grounding non-content words like *the*.

- *Learning through Association*. Using words as classifiers, they need to be presented with visual features. If, for example, we have a richer set of features that can distinguish animals from each other using WAC, but the model has no direct acquaintance with tigers, then it would be feasible to ask *what is a tiger?* and if the answer *a tiger is a kind or large cat*, then the model could possibly take the classifier for cat, copy it for a new tiger classifier, and then adjust the weights for size. If additional information, such as colour and the fact that tigers generally have stripes is also explained, then the weights that represent those features could also be adjusted, thus gaining an intensional notion of what a tiger is without ever having seen one.

- *Learning by Discovery in Interactive Dialogue*. Both models are trained on a corpus and evaluated offline. It would be developmentally motivating, for example, to use the WAC model and begin with no classifiers, but after interacting with a human who points to objects, makes definite descriptions to objects, etc., that the model learns through

experience the meanings of words online as it interacts with a human.

7.4 Parting Thoughts on Meaning and Reference

It is quite an amazing thing that a speaker of some language can communicate with other speakers with relative ease. Some word uses and accents might differ, but the core of grammar and a high percentage of words are shared across all communities which speak that language. When I say blue, I can draw other English speakers' attention to blue things, or when I describe a blue object, English speakers can imagine that colour. If I were to call blue things by a different word, say, *fliff*, e.g., *look at that fliff car over there*, then other English speakers wouldn't know what I am talking about–indeed nobody would because society hasn't agreed that fliff denotes things of a certain colour. Thus, it is argued, that meaning isn't something in our heads, rather it is something that is represented by a larger group (as noted in Chapter 2).

Yet somehow an individual of a language community has an approximation of what she takes to be the meaning of a particular word, such as *blue* in the English community. Certainly, a human cannot learn a language spoken by a community without interacting with a member of that community. Meanings of words are agreed upon based on their uses and those meanings can change in a community over time.

What does meaning have to do with reference? Besides the list of philosophers in Chapter 2 who linked meaning with some kind of referent, there are more recent ideas. For example, Putnam endorsed Burge's claim that reference is psychologically more primitive and more ubiquitous than language use (see Burge (2010)). This links back to the introductory chapter, indeed the introductory paragraph where I pointed out that humans interact with objects almost continually. When you put your shirt on this morning, that required reference to an object; albeit non-linguistic. When you ate your last meal, you somehow brought the food to your mouth. The food itself and the utensil that you used to perform the task both required reference to objects, even if those references were unspoken. Thus how we perceive and interact with the world is tied to objects and reference to them. It is no wonder, then, that among children's first communicative attempts are references to visually present objects (Wittek and Tomasello, 2005).

Meaning, at least in linguistic terms, is perhaps the mechanism that can assign something to a category. Categorization, I assume, begins long before language can be produced and it is potentially those categories that can link language with percepts. For example, a child plays with a ball and knows something about its shape and the fact that it can roll (these can also be seen as properties of the ball). Thus to categorize something as a ball, it would need to have those properties, or at least some kind of approximation to them. Later, when the word *ball*

is acquired, it is linked due to co-occurrence with physical balls. Thus via extension (i.e., the balls), the intension (i.e., the mechanism for assigning objects to the class `ball`) is estimated, based on how that word is used by other speakers.

This shouldn't be controversial, as it doesn't deviate much from the accounts that philosophers have given thus far. In terms of future research, it seems to me that finding and representing meaning will need to take place in dialogue settings, and that if any system stands a chance at finding and representing meaning, it should probably begin with reference.

A

Implementation of SIUM

A.1 Overview

Recall the final equation for SIUM:

$$P(I|U) = P(I) \sum_{r \in R} P(R = r|U) P(R = r|I) \tag{A.1}$$

The goal is to recover $P(I|U)$, which is a distribution over I (i.e., the values in I sum to one). During decoding, a speaker utters a word in U. This triggers a component for $P(R|U)$ into action by producing a distribution over R, given U. That distribution over R is then fed into a component for $P(R|I)$ which steps through all of I, summing together all results from the distribution over R that each I has. That results in a distribution over I, which is what is desired. This is stored as $P(I)$, which is then used as a prior for later steps.

Though fairly straight forward to implement (one only needs objects in a scene, their properties, the referring expressions, and knowledge about which object is referred), in the following section we detail a Java implementation that is freely avaialable.[1]

[1] `https://bitbucket.org/bakuzen/sium`

A.2 Java Implementation

The Java implementation consists of three main objects: LingEvidence (U), Mapping ($P(R|U)$), Context ($P(R|I)$), and a Grounder (the marginalization and combination with $P(I)$).[2]

We will set forth a simple example that uses the following properties: `red`, `green`, `X`, `Z`. There are 3 objects:

- o_1: `red`, `X`
- o_2: `red`, `Z`
- o_3: `green`, `Z`

Where o_x make up I, the properties make up R, and the above list shows what properties belong to which object, which is $P(R|U)$. This is used to fill a Context object:

```
Context scene = new Context();
scene.addPropertyToEntity("o1", "red");
scene.addPropertyToEntity("o1", "X");
scene.addPropertyToEntity("o2", "red");
scene.addPropertyToEntity("o2", "Z");
scene.addPropertyToEntity("o3", "green");
scene.addPropertyToEntity("o3", "Z");
```

The referring expression we will use is *the red cross*. These words fill individual LingEvidence objects.

```
LingEvidence rew = new LingEvidence();
rew.addEvidence("w1", "the")
```

LingEvidence objects are maps, where the key (e.g., above "w1"), can represent the kind of evidence that is being presented. For example, "w1" means *word one*, which maps to a unigram. One can add as many bits of evidence as desired, for example, ngrams (e.g., by adding "w2" for the previous word, etc.), semantic abstractions, etc. A list of these (e.g., ArrayList) make up an entire referring expression.

These can be used for training or for evaluation (i.e., decoding). When used for training, use a Mapping object and give it words in referring expressions represented as LingEvidence objects, as well as scenes represented as contexts:

```
Mapping mapping = new Mapping();
```

[2]Some of the objects implement type generics if something other than Strings are used.

```
//for each word in a referring expression as "rew"
mapping.addEvidenceToTrain(rew, scene.getPropertiesForEntity(gold));
```

Where the `gold` variable is an object ID of the known referred object (e.g., "o1"). When all of the training data has been added, call

```
mapping.train()
```

There are several types of mapping that have already been implemented that are named based on the classifier that does the learning: MaxEntMapping,[3] NaiveBayesMapping, and CoocurrenceMapping. Others can easily be added by implementing the Mapping interface.

For decoding, the scene and the referring expression are packaged into the same objects as described above.

```
Grounder grounder = new Grounder();
// for each word in a referring expression as "rew"
grounder.groundIncrement(scene, mapping.applyEvidenceToContext(rew))
```

The final line above returns a Distribution, which represents $P(I|U)$, which is up-to-date with the current word in the referring expression. I is determined by the set of identifiers (i.e., the unique set of strings that were given as the first argument of scene.addPropertyToEntity()).

A.3 InproTK Module

The objects above are used directly in a module that is distributed with InproTK. Following the terminology explained in Schlangen and Skantze (2011) and Chapter 2 in this thesis, the module's left buffer can take any kind of IU whose payload will be included as a value in LingEvidence. Contexts can be updated by calling a method externally (e.g., by another module). The module can handle the ADD, REVOKE, and COMMIT edit types. The module will produce an IU that has a distribution over I at each word increment.

There are two versions implemented, one for using ngrams, where the length of the ngram can be specified in a config file. The other version can use an implementation of RMRS as described in Peldszus and Schlangen (2012); Peldszus et al. (2012). Examples of using both exist in the InproTK distribution.[4]

[3] https://opennlp.apache.org/
[4] https://bitbucket.org/inpro/inprotk

B

Implementation of WAC

B.1 Overview

The goal is to produce a distribution over a set of visible objects.

For our example, we will use the same scene as in Appendix A. However, for this example the scene will not have properties. Rather, it will have low-level features of RGB values and number of edges:

- o_1: $R : 240$, $G : 20$, $B : 30$, $E : 10$
- o_2: $R : 240$, $G : 20$, $B : 30$, $E : 8$
- o_3: $R : 20$, $G : 240$, $B : 30$, $E : 8$

Knowledge about which object is being referred, namely o_1, we can train classifiers for each word in the referring expression *the red cross* by giving each classifier the four features for object o_1 and a random other object, say o_3. If used during application, the trained classifiers for the three words of the referring expression are taken and each set of above features are given to each of those classifiers so they can provide a "fit" probability, which are then normalised over (over the objects), and averaged over (over the words in the referring expression).

Though fairly straight forward to implement (for example, using Python's sklearn's linear_model), we will describe a Java implementation.

B.2 Java Implementation

The Java implementation extends several useful objects from SIUM, described in Appendix A: Context, and Grounder. Added to those is the WordsAsClasifiersRR. The library used for logistic regression is Apache Mahout which requires a Vector object that contains the low-level features.[1]

First, we must represent the scene using a Context object, much like SIUM. However, it is differently done here:

```
Context scene = new Context();
scene.addPropertyToEntity("o1", "R:240");
scene.addPropertyToEntity("o1", "G:20");
scene.addPropertyToEntity("o1", "B:30");
scene.addPropertyToEntity("o1", "E:10");
...
scene.addPropertyToEntity("o3", "E:8");
```

Note that the values are preceded by a string representing their feature type, followed by a colon (e.g., "R:"), which is differnet from that of SIUM which just receives the value. The referring expression we will use is *the red cross*.

For training, simply create a WordsAsClasifiersRR object and add Vectors to it as positive and negative training samples (there is a utility that can convert from Contexts to Vectors, and a utility that will randomly choose an object identifier that is not the referred object):

```
WordsAsClassifiersRR wacrr = new WordsAsClasifiersRR();
Vector positive = VectorUtils.makeVector(context.getPropertiesForEntity(g
Vector negative = VectorUtils.makeVector(context.getPropertiesForEntity(F

wacrr.addPositiveTrainingSample(word, positive);
wacrr.addPositiveTrainingSample(word, negative);
```

Where the `gold` variable is an object ID of the known referred object (e.g., "o1") and `word` is a string with a word in the referring exprssion (not a LingEvidence object). When all of the training data has been added, call

```
wacrr.train()
```

For application,

```
wacrr.newReferringExpression();
```

[1] `http://mahout.apache.org/`

```
// for each word in a referring expression as word:
wacrr.groundIncrement(context, word);
```

Like SIUM, the last line above returns a Distribution over the objects in the Context.

B.3 InproTK Module

The InproTK module extends the module for SIUM, thus the ADD, REVOKE, and COMMIT procedures follow accordingly. The module similarly accepts Context objects to represent the scenes, where the object information is added to the Context as explained above. The module is included with InproTK, but Apache Mahout must be installed in order for it to run.

Bibliography

Barbara Abbott. *Reference*. Oxford University Press, Oxford, England, 2010.

Gregory Aist, James Allen, Ellen Campana, Lucian Galescu, Carlos Gallo, Scott Stoness, Mary Swift, and Michael Tanenhaus. Software architectures for incremental understanding of human speech. In *Proceedings of CSLP*, pages 1922—-1925, 2006.

Gregory Aist, James Allen, Ellen Campana, Carlos Gomez Gallo, Scott Stoness, and Mary Swift. Incremental understanding in human-computer dialogue and experimental evidence for advantages over nonincremental methods. In *Pragmatics*, volume 1, pages 149–154, Trento, Italy, 2007.

Anne Anderson, M Badger, Ellen Gurman Bard, Elizabeth Boyle, G Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowto, Jan McAllister, J Miller, Catherine Sotillo, Henry S. Thompson, and Regina Weinert. The HCRC Map Task corpus. *Language and Speech*, 34 (4):351–366, 1991.

Yoav Artzi and Luke Zettlemoyer. Weakly Supervised Learning of Semantic Parsers for Mapping Instructions to Actions. *Tacl*, 1(1):49–62, 2013.

Layla El Asri, Romain Laroche, Olivier Pietquin, and Hatim Khouzaimi. NASTIA: Negotiating Appointment Setting Interface. In *Proceedings of LREC*, pages 266–271, 2014.

Jon Barwise and Robin Cooper. Generalized Quantifiers and Natural Language. *Formal Semantics: The Essential Readings*, 4(2):75–126, 2008.

Jon Barwise and John Perry. Situations and Attitudes. *The Journal of Philosophy*, 78(11): 668–691, 1981.

Timo Baumann and David Schlangen. The InproTK 2012 release. In *NAACL-HLT Workshop on Future directions and needs in the Spoken Dialog Community: Tools and Data (SDCTD 2012)*, pages 29–32, 2012.

Timo Baumann, Michaela Atterer, and David Schlangen. Assessing and Improving the Performance of Speech Recognition for Incremental Systems. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 380–388, Boulder, USA, jun 2009.

Niels Beuck and Wolfgang Menzel. Structural prediction in incremental dependency parsing. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 7816 LNCS, pages 245–257, 2013.

Steven Bird. NLTK. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics, 2006.

Richard A. Bolt. "Put-that-there". *ACM SIGGRAPH Computer Graphics*, 14(3):262–270, 1980.

Hélène Bonneau-Maynard, Christelle Ayache, F Béchet, a Denis, a Kuhn, Fabrice Lefèvre, D Mostefa, M Qugnard, S Rosset, and J Servan S. Vilaneau. Results of the French Evalda-Media evaluation campaign for literal understanding. In *Proceedings of the Fifth Int. Conf. on Language Resources and Evaluation (LREC)*, volume 5, pages 2054–2059, 2006.

George Boole. *An Investigation of the Laws of Thought, On wich are Founded the Mathematical Theories of Logic and Probabilities*. Dover Publications, 2005.

Susan E. Brennan. Processes that shape conversation and their implications for computational linguistics. *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics - ACL '00*, pages 1–11, 2000.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 136–145, 2012.

Elia Bruni, Nam Khanh Tran, and Marco Baroni. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47, 2014.

Guido Bugmann, Ewan Klein, Stanislao Lauria, and Theocharis Kyriacou. Corpus-based robotics : A route instruction example. In *Proceedings of Intelligent Autonomous Systems*, pages 96–103. Citeseer, 2004.

Tyler Burge. *Origins of objectivity*. Oxford, 2010.

Hendrik Buschmeier, Timo Baumann, Benjamin Dosch, Stefan Kopp, and David Schlangen. Combining Incremental Language Generation and Incremental Speech Synthesis for Adaptive Information Presentation. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, number July, pages 295–303, Seoul, South Korea, jul 2012. Association for Computational Linguistics.

George Butterworth and Paul Morissette. Onset of pointing and the acquisition of language in infancy, 1996.

John Canny. A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698, jun 1986.

Rehj Cantrell, Matthias Scheutz, Paul Schermerhorn, and Xuan Wu. Robust spoken instruction understanding for HRI. *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on*, 1:275, 2010.

Rudolf Carnap. *Meaning and Necessity, A Study in Semantics and Modal Logic.* University of Chicago Press, 1988.

Joyce Y Chai, Lanbo She, Rui Fang, Spencer Ottarson, Cody Littley, Changsong Liu, and Kenneth Hanson. Collaborative effort towards common ground in situated human-robot dialogue. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 33–40, Bielefeld, Germany, 2014.

David L. Chen and Raymond J. Mooney. Learning to Interpret Natural Language Navigation Instructions from Observations. *AAAI Conference on Artificial Intelligence*, (August):859–865, aug 2011.

Hamid R. Chinaei, Brahim Chaib-draa, and Luc Lamontagne. Learning user intentions in spoken dialogue systems. *Icaart*, 2009.

Noam Chomsky. *Knowledge of language: Its nature, origin, and use*. Praeger, 1986.

Alonzo Chruch. A Formulation of the Simple Theory of Types. *The Journal of Symbolic Logic*, 5(2):56–68, 1940.

Herbert H Clark. *Using Language*. Cambridge University Press, 1996.

Robin Cooper. Records and record types in semantic theory. In *Journal of Logic and Computation*, volume 15, pages 99–112, 2005.

Robin Cooper. Type Theory and Semantics in Flux BT - Philosophy of Linguistics. *Handbook of the Philosophy of Science*, 14:271–323, 2012.

Ann Copestake. Semantic composition with (robust) minimal recursion semantics. In *Proceedings of the Workshop on Deep Linguistic Processing - DeepLP '07*, page 73, Morristown, NJ, USA, 2007. Association for Computational Linguistics.

Maria Cotelli, Barbara Borroni, Rosa Manenti, Valeria Ginex, Marco Calabria, Andrea Moro, Antonella Alberici, Marina Zanetti, Orazio Zanetti, Stefano F. Cappa, and Alessandro Padovani. Universal grammar in the frontotemporal dementia spectrum. Evidence of a selective disorder in the corticobasal degeneration syndrome. *Neuropsychologia*, 45(13): 3015–3023, 2007.

Kenny R. Coventry, Thora Tenbrink, and John Bateman. *Spatial Language and Dialogue*. Oxford University Press, 2009.

Deborah A. Dahl, Madeleine Bates, Michael Brown, William Fisher, Kate Hunicke-Smith, David Pallett, Christine Pao, Alexander Rudnicky, and Elizabeth Shriber. Expanding the scope of the ATIS task: The ATIS-3 corpus. In *Proceedings of the workshop on Human Language Technology*, HLT '94, pages 43–48, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics.

Kathleen Dahlgren. *Referential semantics*. Ph.d, University of California, Los Angeles, 1976.

Robert Dale and Ehud Reiter. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263, 1995.

Charles B. Daniels. *Definite descriptions*. Oxford University Press, 1990.

Vera Demberg and Frank Keller. A psycholinguistically motivated version of TAG. In *Proceedings of the 9th International Workshop on Tree Adjoining Grammars and Related Formalisms.*, pages 25–32, Tübingen, 2008.

Vera Demberg, Asad Sayeed, Angela Mahr, and Christian Müller. Measuring linguistically-induced cognitive load during driving using the ConTRe task. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 176–183, Eindhoven, The Netherlands, 2013.

David DeVault, Iris Oved, and Matthew Stone. Societal Grounding Is Essential to Meaningful Language Use. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 747. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

David DeVault, Kenji Sagae, and David Traum. Can I Finish?: Learning When to Respond to Incremental Interpretation Results in Interactive Dialogue. In *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, number September, pages 11–20. Association for Computational Linguistics, 2009.

Marco Dinarelli, Alessandro Moschitti, and Giuseppe Riccardi. Discriminative reranking for spoken language understanding. *IEEE Transactions on Audio, Speech and Language Processing*, 20(2):526–539, 2012.

Haris Dindo and Daniele Zambuto. A probabilistic approach to learning a visually grounded language model through human-robot interaction. In *IEEE/RSJ 2010 International Conference on Intelligent Robots and Systems, IROS 2010 - Conference Proceedings*, pages 790–796, Taipei, Taiwan, 2010. IEEE.

Keith S Donnellan. Reference and Definite Descriptions. *Philosophical Review*, 75(3):281, 1966.

Frank A. Drews, Monisha Pasupathi, and David L Strayer. Passenger and cell phone conversations in simulated driving. In *Journal of Experimental Psychology: Applied*, volume 14, pages 392–400, New Orleans, USA, 2008a.

Frank A. Drews, Monisha Pasupathi, and David L Strayer. Passenger and cell phone conversations in simulated driving. *Journal of Experimental Psychology: Applied*, 14(4):392–400, 2008b.

Jens Edlund, Joakim Gustafson, Mattias Heldner, and Anna Hjalmarsson. Towards human-like spoken dialogue systems. *Speech Communication*, 50(8-9):630–645, 2008.

Paul Elbourne. On the semantics of pronouns and definite articles. In *WCCFL 20: Proceedings of the 20th West Coast Conference on Formal Linguistics*, pages 164–177, 2001.

Paul Elbourne. Demonstratives as individual concepts. *Linguistics and Philosophy*, 31(4): 409–466, 2008.

Nikos Engonopoulos, Martin Villalba, Ivan Titov, and Alexander Koller. Predicting the resolution of referring expressions from user behavior. In *Proceedings of EMNLP*, pages 1354–1359, Seattle, Washington, USA, 2013. Association for Computational Linguistics.

Raquel Fernández. Rethinking Overspecification in Terms of Incremental Processing. In *Proceedings of the PRE-CogSci 2013 Workshop on the Production of Referring Expressions*, 2013.

Raquel Fernández, Tatjana Lucht, and David Schlangen. Referring under restricted interactivity conditions. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 136–139, 2007.

Charles J. Fillmore. Pragmatics and the description of discourse. *Radical pragmatics*, pages 143–166, 1981.

Charles J. Fillmore. Frame Semantics. *Encyclopedia of Language & Linguistics*, 129(1996): 613–620, 2006.

Charles J Fillmore and Collin F Baker. Frame semantics for text understanding. *Text*, pages 3–4, 2001.

Jerry Fodor and Zenon Pylyshyn. Connectionism and cognitive archirecture: A critical analysis. *Cognition*, 28(1):3–71, 1988.

Gottlob Frege. Über Sinn und Bedeutung. *Erkenntnis*, 100(1):1–15, 1892.

Annemarie Friedrich and Manfred Pinkal. Discourse-sensitive Automatic Identification of Generic Expressions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1272–1281, Beijing, China, 2015. Association for Computational Linguistics.

Kotaro Funakoshi, Mikio Nakano, Takenobu Tokunaga, and Ryu Iida. A Unified Probabilistic Approach to Referring Expressions. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, number July, pages 237–246, Seoul, South Korea, jul 2012. Association for Computational Linguistics.

Konstantina Garoufi and Alexander Koller. The Potsdam NLG systems at the GIVE-2.5 Challenge. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 307–311. Association for Computational Linguistics, 2011.

Jonathan Ginzburg. *The Interactive Stance*. Oxford University Press, 2012.

Susan Goldin-Meadow. Pointing sets the stage for learning language - And creating language. *Child Development*, 78(3):741–745, 2007.

Dave Golland. *Semantics and Pragmatics of Spatial Reference*. PhD thesis, 2015.

Peter Gorniak and Deb Roy. Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, 21:429–470, 2004.

Peter Gorniak and Deb Roy. Probabilistic Grounding of Situated Speech using Plan Recognition and Reference Resolution. In *In Proceedings of the Seventh International Conference on Multimodal Interfaces (ICMI 2005)*, pages 138–143, 2005a.

Peter Gorniak and Deb Roy. Speaking with your sidekick: Understanding situated speech in computer role playing games. In *Proceedings of Artificial Intelligence and Digital Entertainment*, pages 138–143, 2005b.

Jana Götze and Johan Boye. Deriving Salience Models from Human Route Directions. In *Proceedings of IWCS 2013 Workshop on Computational Models of Spatial Language Interpretation and Generation (CoSLI-3)*, number 270019, pages 7–12, Potsdam, Germany, mar 2013. Association for Computational Linguistics.

Zenzi Griffin and Kathryn Bock. What the eyes say about speaking. *Psychological science*, 11 (4):274–279, 2000.

Sergio Guadarrama, Lorenzo Riano, Dave Golland, Daniel Gouhring, Yangqing Jia, Dan Klein, Pieter Abbeel, and Trevor Darrell. Grounding spatial relations for human-robot interaction. In *IEEE International Conference on Intelligent Robots and Systems*, pages 1640–1647, 2013.

Jeanette K Gundel, Nancy Hedberg, and Ron Zacharski. Cognitive status and the form of referring expressions in discourse. *Language*, 69(2):274–307, 1993.

Stefan Hahn, Marco Dinarelli, Christian Raymond, Fabrice Lef??vre, Patrick Lehnen, Renato De Mori, Alessandro Moschitti, Hermann Ney, and Giuseppe Riccardi. Comparing stochastic approaches to spoken language understanding in multiple languages. *IEEE Transactions on Audio, Speech and Language Processing*, 19(6):1569–1583, 2011.

Dilek Hakkani-Tür, Malcolm Slaney, Asli Celikyilmaz, and Larry Heck. Eye Gaze for Spoken Language Understanding in Multi-modal Conversational Interactions. In *Proceedings of the 16th International Conference on Multimodal Interaction - ICMI '14*, pages 263–266, 2014.

Ting Han, Casey Kennington, and David Schlangen. Building and Applying Perceptually-Grounded Representations of Multimodal Scene Descriptions. In *Proceedings of SEMDial*, Gothenburg, Sweden, 2015.

Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3): 335–346, 1990.

Timothy Hazen. *Topic Identification, Spoken Language Understanding: Systems for Extracting Semantic Information From Speech*. John Wiley & Sons, 2011.

Jibo He, Alex Chaparro, Bobby Nguyen, Rondell Burge, Joseph Crandall, Barbara Chaparro, Rui Ni, and Shi Cao. Texting while driving: is speech-based texting less risky than handheld texting? In *Proceedings of the 5th Internation Conference on Automotive User Interfaces and Interactive Vehicular Applications (Automotive UI,13), OCtober 28-30, 2013, Eindhoven, The Netherlands*, pages 124–130, 2013.

Yulan He and Steve Young. Semantic processing using the hidden vector state model. *Computer Speech and Language*, 19(1):85–106, 2005.

Silvan Heintze, Timo Baumann, and David Schlangen. Comparing Local and Sequential Models for Statistical Incremental Natural Language Understanding. In *Computational Linguistics*, pages 9–16. Association for Computational Linguistics, 2010.

Charles T. Hemphill, John J. Godfrey, and George R. Doddington. The ATIS Spoken Language Systems Pilot Corpus. In *Proceedings of the DARPA Speech and Natural Language Workshop*, pages 96–101, 1990.

Matthew Henderson, Blaise Thomson, and Steve Young. Word-Based Dialog State Tracking with Recurrent Neural Networks. In *Sigdial'14*, pages 292–299, Philadelphia, PA, U.S.A., 2014. Association for Computational Linguistics.

Jaakko Hintikka. Quantifiers vs. Quantification Theory. *Dialectica*, 27(3-4):329–358, 1973.

Geoffrey E. Hinton, James L. Mcclelland, and David E. Rumelhart. *Disctributed representations, Parallel distributed processing: explorations in the microstructure of cognition, vol. 1: foundations."*, volume 2. 1986.

Jerry Hobbs. An Improper Treatment of Quantification in Ordinary English. *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics ({ACL})*, 49:57–63, 1983.

William J Horrey and Christopher D Wickens. Examining the impact of cell phone conversations on driving using meta-analytic techniques. *Human factors*, 48(1):196–205, mar 2006.

Julian Hough, Casey Kennington, and Jonathan Ginzburg. Incremental Semantics for Dialogue Processing : Requirements , and a Comparison of Two Approaches. In *Proceedings of IWCS*, pages 206–216. Association for Computational Linguistics, 2015.

Kai-yuh Hsiao, Soroush Vosoughi, Stefanie Tellex, Rony Kubat, and Deb Roy. Object schemas for responsive robotic language use. *Proceedings of the 3rd international conference on Human robot interaction - HRI '08*, 20(4):233, 2008.

Ryu Iida, Shumpei Kobayashi, and Takenobu Tokunaga. Incorporating Extra-linguistic Information into Reference Resolution in Collaborative Task Dialogue. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistic*, pages 1259—-1267, Uppsala, Sweden, 2010.

Ryu Iida, Masaaki Yasuhara, and Takenobu Tokunaga. Multi-modal Reference Resolution in Situated Dialogue by Integrating Linguistic and Extra-Linguistic Clues. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, number 2003, pages 84–92, 2011.

Ellen A. Isaacs and Herbert H. Clark. References in conversation between experts and novices. *Journal of Experimental Psychology: General*, 116(1):26–37, 1987.

Yoko Ishigami and Raymond M. Klein. Is a hands-free phone safer than a handheld phone? *Journal of Safety Research*, 40(2):157–164, 2009.

Hans Kamp. *From discourse to logic : introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory*. Number 42. Springer Science & Business Media, 1993.

David Kaplan. Demonstratives: an essay on the semantics, logic, metaphysics, and epistemology of demonstratives and other indexicals. *Themes from Kaplan*, 135(March):481–564, 1989.

Andrew Kehler. Cognitive Status and Form of Reference in Multimodal Human-Computer Interaction. In *AAAI 00, The 15th Annual Conference of the American Association for Artificial Intelligence*, pages 685–689, 2000.

John Kelleher, Fintan Costello, and Josef Van Genabith. Dynamically structuring, updating and interrelating representations of visual and linguistic discourse context. *Artificial Intelligence*, 167(1-2):62–102, 2005.

Wulf Kellerwessel. Referenztheorien in der analytischen Philosophie. page 480 S., 1995.

Casey Kennington and David Schlangen. Markov Logic Networks for Situated Incremental Natural Language Understanding. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 314–322, Seoul, South Korea, 2012. Association for Computational Linguistics.

Casey Kennington and David Schlangen. Situated incremental natural language understanding using Markov Logic Networks. *Computer Speech & Language*, 2014.

Casey Kennington and David Schlangen. Simple Learning and Compositional Application of Perceptually Grounded Word Meanings for Incremental Reference Resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 292–301, Beijing, China, 2015. Association for Computational Linguistics.

Casey Kennington, Spyros Kousidis, and David Schlangen. Interpreting Situated Dialogue Utterances: an Update Model that Uses Speech, Gaze, and Gesture Information. In *SIGdial 2013*, number August, pages 173–182, 2013.

Casey Kennington, Kotaro Funakoshi, Yuki Takahashi, and Mikio Nakano. Probabilistic multi-party dialogue management for a game master robot. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction - HRI '14*, pages 200–201, Bielefeld, Germany, 2014a. ACM.

Casey Kennington, Spyros Kousidis, Timo Baumann, Hendrik Buschmeier, Stefan Kopp, and David Schlangen. Better Driving and Recall When In-car Information Presentation Uses Situationally-Aware Incremental Speech Output Generation. In *Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '14*, pages 1–7. ACM, 2014b.

Casey Kennington, Spyros Kousidis, and David Schlangen. Situated Incremental Natural Language Understanding using a Multimodal, Linguistically-driven Update Model. In *CoLing 2014*, pages 1803–1812, 2014c.

Casey Kennington, Spyros Kousidis, and David Schlangen. InproTKs: A Toolkit for Incremental Situated Processing. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 84–88, Philadelphia, PA, U.S.A., 2014d. Association for Computational Linguistics.

Casey Kennington, Livia Dia, and David Schlangen. A Discriminative Model for Perceptually-Grounded Incremental Reference Resolution. In *Proceedings of the 11th International Conference on Computational Semantics*, number 2002, pages 195–205. Association for Computational Linguistics, 2015a.

Casey Kennington, Ryu Iida, Takenobu Tokunaga, and David Schlangen. Incrementally Tracking Reference in Human / Human Dialogue Using Linguistic and Extra-Linguistic Informa-

tion. In *HLT-NAACL 2015 - Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings of the Main Conference*, pages 272–282, Denver, U.S.A., 2015b. Association for Computational Linguistics.

Casey Kennington, Maria Soledad Lopez Gambino, and David Schlangen. Real-world Reference Game using the Words-as-Classifiers Model of Reference Resolution, In *Proceedings of SEMDIAL*, 2015c.

Hatim Khouzaimi, Romain Laroche, and Fabrice Lefevre. An easy method to make dialogue systems incremental. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, number June, pages 98–107, Philadelphia, PA, U.S.A., 2014. Association for Computational Linguistics.

Joohyun Kim and Raymond J Mooney. Adapting Discriminative Reranking to Grounded Language Learning. In *Acl-2013*, pages 218–227, 2013.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. In *arXiv preprint arXiv:1411.2539*, pages 1–13, 2014.

Thomas Kollar, Stefanie Tellex, Deb Roy, and Nicholas Roy. Toward understanding natural language directions. *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on*, page 259, 2010.

Alexander Koller, Kristina Striegnitz, Donna Byron, Justine Cassell, Robert Dale, Johanna Moore, and Jon Oberlander. The first challenge on generating instructions in virtual environments. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 5790 LNAI, pages 328–352, 2010.

Alexander Koller, Maria Staudte, Konstantina Garoufi, and Matthew Crocker. Enhancing Referential Success by Tracking Hearer Gaze. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 30–39, Seoul, South Korea, jul 2012. Association for Computational Linguistics.

Spyridon Kousidis, Thies Pfeiffer, Zofia Malisz, Petra Wagner, and David Schlangen. Evaluating a minimally invasive laboratory architecture for recording multimodal conversational data. In *Proceedings of the Interdisciplinary Workshop on Feedback Behaviors in Dialog, Interspeech Satellite Workshop*, pages 39–42, 2012.

Spyros Kousidis, Casey Kennington, and David Schlangen. Investigating speaker gaze and pointing behaviour in human-computer interaction with the mint.tools collection. In *SIGdial 2013*, number August, pages 319–323, 2013.

Spyros Kousidis, Casey Kennington, Timo Baumann, Hendrik Buschmeier, Stefan Kopp, and Stefan Schlangen. Situationally Aware In-Car Information Presentation Using Incremental Speech Generation: Safer, and More Effective. In *Proceedings of the Workshop on Dialogue in Motion (DM), EACL 2014*, pages 68–72, 2014.

Emiel Krahmer and Kees van Deemter. Computational Generation of Referring Expressions: A Survey. *Computational Linguistics*, 38(1):173–218, 2012.

Angelika Kratzer. Situations in natural language semantics. *The Stanford Enyclopedia of Philosophy*, (Spring 2014), 2011.

Saul Kripke. Speaker's Reference and Semantic Reference. *Midwest Studies in Philosophy*, 2 (1):255–276, 1977.

Staffan Larsson. Formal semantics for perceptual classification. *Journal of Logic and Computation*, 25(2):335–369, dec 2015.

Fabrice Lefevre. Dynamic bayesian networks and discriminative classifiers for multi-stage semantic interpretation. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, volume 4, pages 13–16. IEEE, 2007.

Michael Levit and Deb Roy. Interpretation of spatial language in a map navigation task. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 37(3):667–679, 2007.

Percy Liang, Michael I. Jordan, and Dan Klein. Learning Dependency-Based Compositional Semantics. In *Computational Linguistics*, pages 1–94, Portland, Oregon, 2013. Association for Computational Linguistics.

Pierre Lison. *Structured Probabilistic Modelling for Dialogue Management*. PhD thesis, University of Oslo, 2013.

Changsong Liu, Jacob Walker, and Joyce Y Chai. Ambiguities in Spatial Language Understanding in Situated Human Robot Dialogue. In *Dialog with Robots Papers from the AAAI Fall Symposium*, pages 50–55, 2010.

L.T.F. Gamut. *Logic, Language, and Meaning, vol. 2: Intensional logic and logical grammar*, volume 2. Chicago University Press, Chicago, 1991.

Chansong Lui, Rui Fang, and Joyce Yue Chai. Towards Mediating Shared Perceptual Basis in Situated Dialogue. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, number July, pages 140–149, Seoul, South Korea, jul 2012. Association for Computational Linguistics.

Yi Ma, Antoine Raux, Deepak Ramachandran, and Rakesh Gupta. Landmark-based location belief tracking in a spoken dialog system. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 169–178, Seoul, South Korea, jul 2012. Association for Computational Linguistics.

Matt MacMahon, Brian Stankiewicz, and Benjamin Kuipers. Walk the talk: Connecting language, knowledge, and action in route instructions. *Aaai*, pages 1475–1482, 2006.

François Mairesse, Milica Gašić, Filip Jurčíček, Simon Keizer, Blaise Thomson, Kai Yu, and Steve Young. Spoken language understanding from unaligned data using discriminative classification models. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 4749–4752. IEEE, 2009.

Małgorzata Marciniak, Agnieszka Mykowiecka, and Katarzyna Głowińska. Anotowany korpus dialogów telefonicznych. In *Anotowany korpus dialogów telefonicznych*, pages 217–230. Akademicka Oficyna Wydawnicza EXIT, 2010.

Robin M. Masheb, Carlos M. Grilo, and Marney A. White. *An examination of eating patterns in community women with bulimia nervosa and binge eating disorder*, volume 44. Cornell University Press, 2011.

Vivien Mast, Daniel Couto Vale, Zoe Falomir, and Mohammed Elahi Fazleh. Communication, Referential Grounding for Situated Human-Robot. In *Proceedings of SemDial*, Edinburgh, Scotland, 2014.

Cynthia Matuszek, Liefeng Bo, Luke S Zettlemoyer, and Dieter Fox. Learning from Unscripted Deictic Gesture and Language for Human-Robot Interactions. In *Proceedings of AAAI 2014*. AAAI Press, 2014.

Suzanne P McEvoy, Mark R Stevenson, Anne T McCartt, Mark Woodward, Claire Haworth, Peter Palamara, and Rina Cercarelli. Role of mobile phones in motor vehicle crashes resulting in hospital attendance: a case-crossover study. *BMJ (Clinical research ed.)*, 331(7514): 428, 2005.

Marie Jean Meurs, Frédéric Duvert, Fabrice Lefevre, and Renato De Mori. Markov logic networks for spoken language interpretation. *Information Systems Journal*, (1978):535–544, 2008a.

Marie Jean Meurs, Fabrice Lefevre, and Renato De Mori. A Bayesian approach to semantic composition for spoken language interpretation. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, number 33549, pages 1161–1164, Brisbane, 2008b.

Marie Jean Meurs, Fabrice Lefèvre, and Renato De Mori. Spoken language interpretation: On the use of dynamic Bayesian networks for semantic composition. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 4773–4776, 2009a.

Marie Jean Meurs, Fabrice Lefevre, and Renato De Mori. Learning bayesian networks for semantic frame composition in a spoken dialog system. In *... : The 2009 Annual Conference of the ...*, pages 61–64, Boulder, Colorado, USA, 2009b.

Ivan V. Meza-Ruiz, Sebastian Riedel, and Oliver Lemon. Accurate statistical spoken language understanding from limited development resources. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 5021–5024. IEEE, 2008.

John Stuart Mill. *A system of logic, ratiocinative and inductive, being a connected view of the principles of evidence and the methods of scientific investigation.* London, 1846.

Teruhisa Misu, Antoine Raux, Ian Lane, and Moffett Field. Situated Language Understanding at 25 Miles per Hour. In *SIGdial 2014*, number June, pages 22–31, Philadelphia, PA, U.S.A., 2014. Association for Computational Linguistics.

Geoffrey Nunberg. Indexicality and deixis. *Linguistics and Philosophy*, 16(1):1–43, 1993.

Marcus Nyström and Kenneth Holmqvist. An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data. *Behavior research methods*, 42(1):188–204, 2010.

Buß Okko, Timo Baumann, and David Schlangen. Collaborating on Utterances with a Spoken Dialogue System Using an ISU-based Approach to Incremental Dialogue Management. In *Proceedings of SIGdial*, pages 233–236, Tokyo, Japan, sep 2010.

Barbara Partee. Montague grammar and transformational grammar. *Linguistic Inquiry*, 6(2): 203–300, 1975.

Giuseppe Peano. *Studii di logica matematica*. Attit della Academia delle scienze di Torino, Classe di scienze fisiche, matematiche e naturali, 1897.

Andreas Peldszus and David Schlangen. Incremental Construction of Robust but Deep Semantic Representations for Use in Responsive Dialogue Systems. In *Proceedings of the Workshop on Advances in Discourse Analysis and its Computational Aspects*, pages 59–76, Mumbai, India, dec 2012. The COLING 2012 Organizing Committee.

Andreas Peldszus, Okko Buß, Timo Baumann, and David Schlangen. Joint Satisfaction of Syntactic and Pragmatic Constraints Improves Incremental Spoken Language Understanding. In *Proceedings of the 13th EACL*, pages 514–523, Avignon, France, apr 2012. Association for Computational Linguistics.

Paul Pietroski. *The character of natural language semantics*. Oxford University Press, 2003.

Luis Pineda and Gabriela Garza. A Model for Multimodal Reference Resolution. *Computational Linguistics*, 26:139–193, 2000.

Steven Pinker and Alan Prince. On language and connectionism: analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28(1-2):73–193, 1988.

Massimo Poesio. An incremental model of anaphora and reference resolution based on resource situations. *Dialogue & Discourse*, 2(1):1–52, 2011.

Massimo Poesio and Renata Vieira. A Corpus-Based Investigation of Definite Description Use. *Computational Linguistics*, 24(2):47, jun 1997.

Zahar Prasov and Joyce Y. Chai. What's in a gaze? In *Proceedings of the 13th international conference on Intelligent user interfaces - IUI '08*, page 20, 2008.

Zahar Prasov and Joyce Y Chai. Fusing Eye Gaze with Speech Recognition Hypotheses to Resolve Exophoric References in Situated Dialogue. In *Computational Linguistics*, number October, pages 471–481, 2010.

Kari Pulli, Anatoly Baksheev, Kirill Kornyakov, and Victor Eruhimov. Real-time computer vision with OpenCV. *Communications of the ACM*, 55(6):61, 2012.

Matthew Purver, Arash Eshghi, and Julian Hough. Incremental Semantic Construction in a Dialogue System. In *Proceedings of the 9th International Conference on Computational Semantics (IWCS)*, pages 365–369. Association for Computational Linguistics, 2011.

Hilary Putnam. Naturalism, Realism, and Normativity. *Journal of the American Philosophical Association*, (1):312–328.

Hilary Putnam. Meaning and reference. *The Journal of Philosophy*, 70(19):699–711, 1973.

Hilary Putnam. The Meaning of Meaning. *Minnesota Studies in the Philosophy of Science*, 7: 131–193, 1975.

Willard Van Orman Quine. Reference and Modality. In *From a Logical Point of View (second, revised edition)*, pages 139–159. 1980.

Antoine Raux, Brian Langner, and Dan Bohus. Let's go public! taking a spoken dialog system to the real world. In *in Proc. of Interspeech* . . . , pages 885–888. Citeseer, 2005.

Hilke Reckman, Jeff Orkin, and Deb Roy. Learning meanings of words and constructions, grounded in a virtual game. *Semantic Approaches in Natural Language Processing*, (2004): 67, 2010.

Deb Roy and Ehud Reiter. Connecting language to the world. *Artificial Intelligence*, 167(1-2): 1–12, 2005.

Bertrand Russell. On Denoting. *Mind*, XIV(4):479–493, 1905.

Jacqueline Sachs, Barbara Bard, and Marie L Johnson. Language learning with restricted input: Case studies of two hearing children of deaf parents. *Applied Psycholinguistics*, 2 (01):33–54, 1981.

Susanne Salmon-Alt and Laurent Romary. Reference resolution within the framework of Cognitive Grammar. *International Colloqium on Cognitive Science*, abs/0909.2:1–25, 2001.

Giampiero Salvi, Luis Montesano, Alexandre Bernardino, and José Santos-Victor. Language bootstrapping: Learning word meanings from perception-action association. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 42(3):660–671, jun 2012.

Allison Sauppé and Bilge Mutlu. Robot deictics. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction - HRI '14*, pages 342–349, 2014.

David Schlangen. *A Coherence-Based Approach to the Interpretation of Non-Sentential Utterances in Dialogue*. PhD thesis, University of Edinburgh, 2004.

David Schlangen and Gabriel Skantze. A General, Abstract Model of Incremental Dialogue Processing. In *Proceedings of EACL*, 2009.

David Schlangen and Gabriel Skantze. A General, Abstract Model of Incremental Dialogue Processing. In *Dialogue & Discourse*, volume 2, pages 83–111, 2011.

David Schlangen, Timo Baumann, and Michaela Atterer. Incremental Reference Resolution: The Task, Metrics for Evaluation, and a Bayesian Filtering Model that is Sensitive to Disfluencies. In *Proceedings of the 10th SIGdial*, number September, pages 30–37, London, UK, 2009. Association for Computational Linguistics.

Matthias Schlesewsky and Ina Bornkessel. On incremental interpretation: Degrees of meaning accessed during sentence comprehension. *Lingua*, 114(9-10):1213–1234, 2004.

William Schuler, Stephen Wu, and Lane Schwartz. A Framework for Fast Incremental Interpretation during Speech Decoding. *Computational Linguistics*, 35(3):313–343, 2009.

John R Searle. Speech Acts. *Speech acts: An essay in the philosophy of language*, 3, 1976.

John R. Searle. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(03):417, 1980.

Ethan Selfridge and Iker Arizmendi. Integrating incremental speech recognition and pomdp-based dialogue systems. In . . . *Discourse and Dialogue*, pages 275–279, Seoul, South Korea, jul 2012. Association for Computational Linguistics.

Jude W. Shavlik, Raymond J. Mooney, and Geoffrey G. Towell. Symbolic and Neural Learning Algorithms: An Experimental Comparison. *Machine Learning*, 6(2):111–143, 1991.

Alexander Siebert and David Schlangen. A simple method for resolution of definite reference in a shared visual context. In *Proceedings of the 9th SIGdial Workshop on . . .*, number June, pages 1–4, Columbus, Ohio, 2008. Association for Computational Linguistics.

Gabriel Skantze and Anna Hjalmarsson. Towards Incremental Speech Production in Dialogue Systems. In *Word Journal Of The International Linguistic Association*, pages 1–8, Tokyo, Japan, sep 1991.

Gabriel Skantze and David Schlangen. Incremental dialogue processing in a micro-domain. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics on EACL 09*, (April):745–753, 2009.

Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. Grounded Compositional Semantics for Finding and Describing Images with Sentences. *Tacl*, 2(April):207–218, 2014.

Philipp Spanger, Masaaki Yasuhara, Ryu Iida, Takenobu Tokunaga, Asuka Terai, and Naoko Kuriyama. REX-J: Japanese referring expression corpus of situated dialogs. *Language Resources and Evaluation*, 46(3):461–491, dec 2012.

Michael J. Spivey, Michael K. Tanenhaus, Kathleen M. Eberhard, and Julie C. Sedivy. Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. *Cognitive Psychology*, 45(4):447–481, 2002.

Michael Spranger and Luc Steels. Emergent Functional Grammar for Space. In L Steels, editor, *Experiments in Cultural Language Evolution*, volume 3 of *Advances in Interaction Studies*, pages 207–232. John Benjamins, 2012.

Luc Steels. Perceptually grounded meaning creation. In *Proceedings of the Second International Conference on Multiagent Systems*, pages 338–344, 1996.

Luc Steels and Tony Belpaeme. Coordinating perceptually grounded categories through language: a case study for colour. *The Behavioral and brain sciences*, 28(4):469–489; discussion 489–529, 2005.

Luc Steels and Frederic Kaplan. Situated grounded word semantics. In *IJCAI International Joint Conference on Artificial Intelligence*, volume 2, pages 862–867, 1999.

Luc Steels and Martin Loetzsch. Perspective Alignment in Spatial Language. In Kenny R Coventry, Thora Tenbrink, and John A Bateman, editors, *Spatial Language and Dialogue*, pages 70–88. Oxford University Press, 2009.

Scott C. Stoness, Joel Tetreault, and James Allen. Incremental parsing with reference interaction. *Proceedings of the Workshop on Incremental Parsing Bringing Engineering and Cognition Together - IncrementParsing '04*, pages 18–25, 2004.

Scott C Stoness, James Allen, Greg Aist, and Mary Swift. Using Real-World Reference to Improve Spoken Language Understanding. In *Aaai*, 2005.

Peter Strawson. On referring. *Mind*, 59(235):320–344, 1950.

Yasuyuki Sumi, Masaharu Yano, and Toyoaki Nishida. Analysis environment of conversational structure with nonverbal multimodal data. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction on - ICMI-MLMI '10*, page 1. ACM, 2010.

Satoshi Suzuki and Keiichi Abe. Topological structural analysis of digitized binary images by border following. *Computer Vision, Graphics and Image Processing*, 30(1):32–46, 1985.

Michael K Tanenhaus and Michael J Spivey-Knowlton. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632, 1995.

Alfred Tarski. The concept of truth in formalized languages. In A Tarski, editor, *Logic, semantics, metamathematics*, pages 1–3. Oxford University Press, 1956.

Joel Tetreault and James Allen. Semantics, Dialogue, and Reference Resolution. In *Catalog-04: 8th Workshop on the Semantics and Pragmatics of Dialogue*, 2004.

Takenobu Tokunaga, Ryu Iida, Asuka Terai, and Naoko Kuriyama. The REX corpora : A collection of multimodal corpora of referring expressions in collaborative problem solving dialogues. In *Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 422–429, 2012.

David Traum, David Devault, Jina Lee, Zhiyang Wang, and Stacy Marsella. Incremental dialogue understanding and feedback for multiparty, multimodal conversation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 7502 LNAI, pages 275–288. Springer, 2012.

Gokhan Tur, Dilek Hakkani-Tür, and Larry Heck. What is left to be understood in atis? In *2010 IEEE Workshop on Spoken Language Technology, SLT 2010 - Proceedings*, pages 19–24, Berkeley, California, 2010. IEEE.

Gokhan Tur, Li Deng, Dilek Hakkani-Tür, and Xiaodong He. Towards deeper understanding: Deep convex networks for semantic utterance classification. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 5045–5048. IEEE, 2012.

Peter D Turney and Patrick Pantel. From Frequency to Meaning: Vector Space Models of Semantics. *Artificial Intelligence*, 37(1):141–188, 2010.

Kees van Deemter, Albert Gatt, Roger P G van Gompel, and Emiel Krahmer. Toward a Computational Psycholinguistics of Reference Production. *Topics in Cognitive Science*, 4(2): 166–183, 2012.

Adam Vogel and Dan Jurafsky. Learning to Follow Navigational Directions. In *Acl*, number July, pages 806–814, 2010.

Jason D. Williams. Incremental partition recombination for efficient tracking of multiple dialog states. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 5382–5385, 2010.

Terry Winograd. *Procedures as a representation for data in a computational program for understanding natural language*. PhD thesis, Massachusetts Institute of Technology., 1971.

Angelika Wittek and Michael Tomasello. Young children's sensitivity to listener knowledge and perceptual context in choosing referring expressions. *Applied Psycholinguistics*, 26(04): 541–558, 2005.

Luke S Zettlemoyer and Michael Collins. Online learning of relaxed CCG grammars for parsing to logical form. *Computational Linguistics*, (June):678–687, 2007.

Luke S. Zettlemoyer and Michael Collins. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. *arXiv preprint arXiv:1207.1420*, abs/1207.1, 2012.