On linguistic signalling games in improvised theatre

Lochlan Morrissey
BA (Hons)

School of Humanities, Languages and Social Science
Griffith University

Submitted in fulfilment of the requirements of the degree of
Doctor of Philosophy

26 June, 2017
Abstract

This thesis offers a game theoretic model of the use of language in improvised theatre. The model views actors as game theoretic agents; that is, rational agents who select certain utterances, and certain interpretations of these utterances, by calculating the utterance or interpretation that will yield the greatest utility. Actors in improvised performance have the difficult task of executing a performance that is convincingly verisimilar to a group of spectators, while simultaneously generating the plot, and the world in which it takes place, called the plot-world. The generation of a plot requires coordinated action, and cooperative planning, which must take place during the performance. Actors must therefore produce utterances that signal their intentions for the trajectory of the plot, but that conceal this information from the audience, for this planning information damages the verisimilitude of the plot. Concealment of this kind is treated in this thesis as a problem of natural language steganography. The study of steganography involves examination of situations in which information is transmitted between certain parties in such a way that the very presence of this information is kept from unwarranted access.

The model set out in this thesis is novel for a game theoretic study of natural language insofar as its primary purpose is to be a comprehensively descriptive model of a particular genre of speech, rather than a generalised, predictive model. To this end, it is based on a dataset of two hours of improvised performance. Basing the model on natural language necessitates the addition of numerous formalisms to the standard model that is used in game theoretic studies of natural language. I make two major amendments to the standard model. (1) Discourse Representation Theory is included as the semantic component of the model. This endows the model with an intensional, model theoretic semantics, and one that is able to express the nuanced meanings that are used in the actors’ concealment of information. (2) Conversations are modelled by iterated games. Iterated games are games that are made up of a number of periods, where these periods are repetitions of a specific game. Iterated games allow for the model to take into account various important facets of communication, including most saliently the cumulative effects of information throughout the course of a conversation, and the alteration of player preferences over time.

Three major aspects of the generation of improvised performances are attended to. First, frequent strategies that are used to introduce information into the plot-world are identified, and some of the linguistic characteristics of these strategies are described with respect to the game theoretic model. Second, the use of strategies that are meant to guide certain interpretations of the utterances produced in the course of the performance is dis-
cussed. Third, the actual strategies that are used to conceal information, and the precise nature of the information that is concealed, are analysed. At the end of each of these discussions, principles that describe the optimality of certain strategies with respect to the functions of those strategies are proposed. These are called heuristics, and by stating optimal means of achieving certain functions of utterances, it is proposed that they capture the norms by which improvising actors select utterances, and by which they interpret them.
This work has not previously been submitted for a degree or diploma in any university. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Lochlan Morrissey
26 June, 2017
For Emily, Dukka
Contents

Abstract ii
List of Figures ix
List of Tables x
List of Data xi
Acknowledgments xii

1 Introduction 1
  1.1 Studying nightmare situations ........................................ 1
  1.2 An overview of the thesis’s argument ................................ 5
    1.2.1 Fundamental methodological principles ............................. 5
    1.2.2 Major claims .......................................................... 7
    1.2.3 An overview of the model .......................................... 7
    1.2.4 Thesis structure ..................................................... 9

2 Theoretical grounding 11
  2.1 Language and game theory ............................................. 11
    2.1.1 Probability theory and possible worlds theory .................. 13
    2.1.2 Basic notions ...................................................... 15
    2.1.3 Signalling games .................................................. 21
    2.1.4 Iterated games .................................................... 29
    2.1.5 Conclusion .......................................................... 35
  2.2 Improvised theatre ..................................................... 35
    2.2.1 Improvisation ...................................................... 36
    2.2.2 Theatrical improvisation ......................................... 38
  2.3 Conclusion: the rationale for the study ............................ 46

3 A more expressive semantics for games 48
  3.1 Game theoretic approaches to interpretation problems .............. 50
    3.1.1 Games of partial information ..................................... 50
    3.1.2 The Iterated Best Response model ................................ 51
    3.1.3 Interpretation under error ....................................... 52
  3.2 Discourse Representation Theory .................................... 54
    3.2.1 Basic discourse representation structures ....................... 55
    3.2.2 Propositional attitudes .......................................... 61
    3.2.3 Syntax .............................................................. 64
    3.2.4 Semantics .......................................................... 67
    3.2.5 Some omissions from the standard model ....................... 69
4 Improvised performance games

4.1 Elements of improvised performances

4.1.1 Plot-worlds

4.1.2 The plot, scenes, and states

4.1.3 Actors and characters

4.2 A DRT for narratives

4.2.1 What is different about DRT for narratives

4.2.2 Character traits and Discourse Representation Structures

4.2.3 Offer–response pairs and DRSs

4.3 Improvised performance games and their iterated forms

4.3.1 Improvised performance games

4.3.2 Plots as iterated games

4.4 Conclusion and a caveat about concealment and utilities

5 Plot-world generation and consistency

5.1 The data used in this study

5.2 Information creation in a plot-world

5.2.1 Presupposition

5.2.2 Attribute ascription through sociolinguistic variation

5.2.3 Creating a plot-world

5.3 Resuming a plot-world

5.3.1 Shifting state between scenes

5.3.2 Subverting expectations

5.4 Errors of consistency within and across scenes

5.5 Heuristics for world building

5.5.1 Verisimilar worlds

5.5.2 Resuming plot-worlds

5.5.3 Plausibility

5.5.4 Heuristics in interpretation

5.6 Conclusion

6 Resolving interpretations with divergent common grounds

6.1 Context and negotiation

6.1.1 Existing game models of common ground

6.1.2 A new game model of common ground

6.1.3 Negotiation

6.2 Strategies for resolving interpretation

6.2.1 Negotiating operative interpretations

6.2.2 Operations on the state of the scene

6.3 Heuristics for resolving interpretation

6.3.1 Markedness of the operative status of an interpretation

6.3.2 Convergence through repetition

6.3.3 Response strategies

6.4 Conclusion
7 The subliminal channel in improvised theatre

7.1 Information theory

7.1.1 Information content and entropy of random variables

7.1.2 A model of information theoretic communication

7.2 Steganography

7.2.1 Information theoretic steganography

7.2.2 Game models of steganography

7.3 Natural language steganography

7.3.1 DRS with permissions

7.3.2 A formal model of natural language steganography

7.3.3 Stegogames

7.3.4 An interim summary

7.4 Strategies to conceal information

7.4.1 Declarative utterances

7.4.2 Questions

7.4.3 Implicature

7.5 The advantages and the limits of steganography

7.6 Heuristics for concealed communication

7.6.1 An a priori common ground for actors

7.6.2 Proposal, acceptance, rejection

7.6.3 A return to verisimilitude

7.7 Conclusion

8 An analysis of a scene

8.1 To recapitulate: an outline of the model

8.2 The scene

8.3 Describing the plot through state transitions

8.4 A segment of the plot as an iterated game

8.4.1 The first offer–response pair

8.4.2 The second offer–response pair

8.4.3 The third offer–response pair

8.4.4 The fourth offer–response pair

8.5 Conclusion

9 Conclusion and future work

9.1 Future work

9.2 Limitations

References

List of Symbols and Notation

Appendix A Heuristics, their definitions, and assumptions
## List of Figures

2.1 Normal-form representation of the coin-matching game. .......... 18
2.2 Normal-form representation of the Prisoner’s Dilemma. .......... 20
2.3 Graph of preferences of players in the Prisoner’s Dilemma. .......... 21
2.4 Tabular representation of the signalling game described in example 2.3. 24
2.5 Normal form of the shootout game, with characters’ preferences only. 41
2.6 Normal form of the shootout game, with actors’ preferences only. 42
3.1 Interpretative process. ............................................ 74
4.1 Graphic depiction of a plot. ...................................... 82
6.1 Tabular representation of the pea signalling game. ................. 145
6.2 A tabular representation of the Verdi versus Wagner signalling game. 151
7.1 Plot of the entropy of a coin flip. ................................. 181
7.2 Schematic diagram of a communication system. .................... 183
7.3 Information theoretic, symmetric-key stegosystem. ................ 189
7.4 Natural language stegosystem. .................................... 202
List of Tables

4.1 Summary of the possible response types. ........................................ 85
4.2 Operation of offer and response types on state. ............................... 85
5.1 Summary of transcription symbols used in this thesis. ........................ 107
6.1 Summary of the possible kinds of response. .................................... 163
# List of Data

<table>
<thead>
<tr>
<th></th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Example of a CA-style transcript, from scene 4-B1.</td>
<td>106</td>
</tr>
<tr>
<td>2</td>
<td>Beginning of 2-C1.</td>
<td>108</td>
</tr>
<tr>
<td>3</td>
<td>Beginning of 1-A1.</td>
<td>110</td>
</tr>
<tr>
<td>4</td>
<td>Beginning of 2-A1.</td>
<td>112</td>
</tr>
<tr>
<td>5</td>
<td>Beginning of 3-B1.</td>
<td>118</td>
</tr>
<tr>
<td>6</td>
<td>End of 1-A1.</td>
<td>119</td>
</tr>
<tr>
<td>7</td>
<td>Middle of 3-B1.</td>
<td>122</td>
</tr>
<tr>
<td>8</td>
<td>Beginning of 3-C1.</td>
<td>122</td>
</tr>
<tr>
<td>9</td>
<td>End of 1-A1.</td>
<td>124</td>
</tr>
<tr>
<td>10</td>
<td>Middle of 1-B1.</td>
<td>124</td>
</tr>
<tr>
<td>11</td>
<td>Beginning of 1-C3.</td>
<td>125</td>
</tr>
<tr>
<td>12</td>
<td>Middle of 1-A2.</td>
<td>127</td>
</tr>
<tr>
<td>13</td>
<td>Beginning of 3-C1.</td>
<td>128</td>
</tr>
<tr>
<td>14</td>
<td>Middle of 1-A2.</td>
<td>147</td>
</tr>
<tr>
<td>15</td>
<td>Middle of 2-A2.</td>
<td>156</td>
</tr>
<tr>
<td>16</td>
<td>Most of 3-B2, from the beginning.</td>
<td>159</td>
</tr>
<tr>
<td>17</td>
<td>Middle of 1-A3.</td>
<td>164</td>
</tr>
<tr>
<td>18</td>
<td>Beginning of 2-A2.</td>
<td>167</td>
</tr>
<tr>
<td>19</td>
<td>Beginning of 4-C1.</td>
<td>212</td>
</tr>
<tr>
<td>20</td>
<td>Beginning of 3-C1.</td>
<td>214</td>
</tr>
<tr>
<td>21</td>
<td>Middle of 2-A1.</td>
<td>215</td>
</tr>
<tr>
<td>22</td>
<td>Scene 1-A1.</td>
<td>229</td>
</tr>
<tr>
<td>23</td>
<td>Beginning of 1-A1.</td>
<td>234</td>
</tr>
</tbody>
</table>
Acknowledgments

This thesis, as I’m sure is true of all other theses, has followed a very strange, and winding, path. It was, in its earliest form, a game theoretic description of the late-Renaissance theatrical form popular in the Italian city-states, *commedia dell’arte*, whose stock characters and their linguistic habits, I believed, could allow game theoretic linguists to understand how players designed their utterances to confirm to, and to diverge from, canonically expected forms. The genre specification was dropped in favour of a more general description of improvised theatre and the linguistic strategies of the actors, though a good, solid focus remained vague and elusive, despite the protestations of my wonderful (and, needless to say, patient) supervisors, for the better part of a year. ESSLLI 2015 changed that. The exciting discussions with fellow ESSLLIers, and the profound reflections they caused, gave me the impetus (and the courage) to choose something seemingly unrelated from the previous topics I had touched on, but one which has thoroughly satisfied me intellectually.

The work that you hold in your hands, or read on your screen, Dear Reader, could not have come about without the labour of many people, in directly contributing to the text, in guiding my intellectual development, and in supporting me through these four years.

I thank my supervisors, Claire Kennedy, Andrea Schalley, and Anton Benz. Claire and I have known each other for a long time, starting when I was a fresh-faced student of Italian. Since then, Claire has been immensely supportive of my academic endeavours, as well as my various creative pursuits. Claire’s attention to detail and curiosity have been extremely important to the project. We can thank Claire for any lack of vague waffling and adherence to some sort of system of spelling in this thesis. Thank you, Claire! Likewise, Andrea and I have worked together for many years. Andrea was my Honours supervisor, and so fine a supervisor is she that I simply couldn’t have conceived of starting a doctorate without her as a supervisor. Her breadth of knowledge and extraordinary work ethic have shaped this thesis in manifold ways, and without her it would be a shadow of the work that it is. Thank you, Andrea! Finally, I thank Anton. Even before he joined the supervision team, Anton exerted a strong effect on this thesis. It was his *Introduction to Game Theory for Linguists* (written with Gerhard Jäger and Robert van Rooij) that led me to decide to work in game theoretic linguistics in the first place. He was extremely generous with his time and wisdom when I visited Berlin for five weeks, and the discussions we had dramatically changed the direction of the project, and made it much stronger. Thank you, Anton!

I thank Petter Ericson. We met at ESSLLI 2015 and immediately became fast friends, owing perhaps to the fact that Petter is also a musician, and because we both have fine beards. It was Petter who pointed out that my poorly-articulated musings were a problem of steganography, and it was Petter who offered me inspiration in two moments of feeling intellectually adrift, once giving me intellectual and moral support, and once giving me a physical space to work and stay (and, most importantly, jam!) for a week. I cannot express the profound effect Petter has had on this project, and I thank him sincerely.
I thank Helen Arnot. Her supportive wisdom and solidarity in the struggles of writing a thesis meant that I was far happier and far less lonely than I imagine I would have been had I not had such a wonderful officemate.

I thank the staff of the School of Languages and Linguistics (which has become Humanities, Languages and Social Science, but it’ll always be LAL to me). I’ve been a student and staff member of the school now for eight years, and it’s been a lovely and nurturing environment for me as a scholar, and as a person. In particular, I thank Val Beckett and Ron Magarry for their always friendly assistance, and the Italian crew—Sara Visocnik, Tiziana Miceli, Bernadine Leon, Rosalia Miglioli, and Claire Kennedy—for forcing me to keep up the language, and for being good pals!

I thank Improv Conspiracy for their willingness to allow me to record their performance, particularly at short notice.

I thank Anne Tiernan, Sian Robinson, Julie Blakey, and Jenny Menzies from the Policy Innovation Hub for giving me employment in the latter part of the thesis project, and for providing me with an extremely strong light at the end of the tunnel.

I thank Christian Romuss for his excellent proofreading, and for regular sessions of mutual, vociferous agreement.

I thank the many people who have supported me emotionally and psychologically outside academia. The Sunday afternoon West End–Dutton Park social soccer crew have helped me debrief from the busy weeks, and prepare for those to come. My musician friends, in particular Justin Bullock, Bec Lloyd-Jones, Glenn Christensen, Brett Sturdy, Stewie Webb, Joe Meyers, Petter Ericson, Rosalia Miglioli, the Brisbane Baroque Players, and l’Unicoro, have provided me with hours of profound enjoyment and have helped me grow as a musician, far beyond my expectations. Stephen Cuttriss has lent me a bandoneón, and encouraged me in learning it, which has been a saviour to me in the last few months of the doctoral project. My sailing mates, Mike Hobbs, Rory O’Connell, Craig Cuttriss, Stephen Cuttriss, Matt Prime, as well as other skippers, have graciously accepted me onto their boats as ballast crew. The best hosts anywhere, Maurizio Bontempi and Rosalia Miglioli, have kept me fed from their own orto through the years. And many more wonderful people have safeguarded my sanity and happiness throughout the project: Simon Katterl (even if he still manages to frighten me from Melbourne), Sam Brabham, Jess McLean, Sasha Hunt, Charlie Wayment, Jerath Head, Stéphane Guillou, Liz Venables, Enton Shaho, Gemma Plesman, Alena Petkov, Dana Cohen, Penny, Esla and Ivy Prime, Alexander Borkowski and the boys, Conrad and Vincent, Alexander Thomas, Bruno De Craene, Will Kane-Potaka, Joseph De Veaugh-Geiss, Carlos Sieguesmund, Antje Caspers, and the Molinos: Paola, Michele, Martina, and Benedetta.

I thank my family, whose constant love, patience, and understanding have carried me through these four years. My parents, Cheryl and Shane, have never hesitated to help me when I needed it, and a trip down to their place was always a means of recuperating and refreshing. My brothers, Jeshua and Jason, have both helped me with various problems: Jeshua, with maths and programming questions, Jason with artistic and political ones. My family-in-law, Karen, James, Will and Taylor, have also provided emotional sanctuary at various times, and have been extremely generous throughout.

Finally, I thank my wife Emily, whose love and extraordinary commitment have made my life far richer. Her steadfastness has allowed this thesis, and me along with it, to flourish. Thank you, Emily!
Chapter 1

Introduction

Estragon, sitting on a low mound, is trying to take off his boot. He pulls at it with both hands, panting. He gives up, exhausted, rests, tries again. As before. Enter Vladimir. 
estragon: (giving up again) Nothing to be done.

Samuel Beckett, Waiting for Godot

1.1 · Studying nightmare situations

As one might expect of a thesis about improvised theatre, we begin with a tale.

Example 1.1: Lucky’s piano recital. Our heroes, dutiful Vladimir and naïve Estragon, attend a piano recital put on by their mutual acquaintance, Lucky. It is terrible. Lucky butchers the Bach, murders the Mozart, and, for good measure, wrecks the Ravel. After a marathon two-hour performance, Vladimir and Estragon, exhausted, attempt to slip away, but they bump into Lucky at the exit. After exchanging the usual polite nothings, Estragon asks Vladimir what he thought of the performance. Not wishing to offend Lucky, but also not wishing to lie, Vladimir murmurs, “it was very … expressive.” Lucky is rapt, and replies, “I don’t play accurately—anyone can play accurately—but I play with wonderful expression. As far as the piano is concerned, sentiment is my forte.” Despite Estragon’s apparently infinite naïveté, he knows that Vladimir detests mawkish playing, and so he has indicated that he disliked the performance. Lucky, on the other hand, is very pleased by the compliment, because he does not know that Vladimir dislikes overly expressive playing.

The problem that confronts Vladimir is that he wishes to convey two different things to two different interlocutors. To Estragon, he wishes to convey that he did not enjoy the performance, while to Lucky, he wants to convey the exact opposite. Vladimir has to find an utterance that simultaneously does both.

I assume that situations that fit this schema are familiar to most readers, though I hope they are not necessarily too familiar. (When I provided example 1.1 to an acquaintance at a party, as an instance of what this thesis attempts to model, he called my doctoral project
“a study of nightmare situations”). But there is one genre of naturally-occurring speech in which virtually every utterance has to convey two separate things. In improvised theatre, actors must plan and coordinate their behaviour to successfully execute a scene, but may do so only with utterances that plausibly belong to the characters they are playing in a plot-world. They therefore conceal their planning within the utterances that are ostensibly produced by characters in the plot. Actors are required to be aware at all times of the double function of any utterance produced by their fellow actors, for otherwise, a performance risks collapse.

But this is only the most extreme use of language in improvised theatre. An actor’s repertoire of linguistic strategies spans the entire linguistic faculty, for it is almost solely through the performance and interpretation of utterances that actors are able to execute performances. Furthermore, improvised theatre has a set of well-established and well-articulated aesthetic norms that constrain what actors consider best practice. This interaction of extemporaneous language production and the constraints of performing for an audience makes the use of language in improvised theatre interesting for linguistic research. It is precisely the definition of a formal linguistic model of improvised theatre that this thesis aims for.

To model improvised theatre, I employ three formal theories.

The first of these theories is game theoretic linguistics, which has seen a proliferation of studies in recent years (see, for instance, Franke 2011, Benz 2012a, van Rooij 2008). The core feature of these game theoretic linguistic models is that interlocutors are viewed as rational agents. A game theoretically rational agent selects her actions to optimise the utility that she will receive in a given situation. So, rational agents in linguistic games are interlocutors who select linguistic strategies to gain maximal utility. I propose that actors engaged in improvised theatre be viewed as rational agents who use linguistic strategies to maximise their benefit while executing a performance. While the focus of this study is somewhat novel, the questions that I address are the questions that are central to more general game theoretic research on language use. The most salient of these questions are twofold. (1) How do agents felicitously interpret utterances that they observe? Interpretation is a complex behaviour, and is influenced by multiple factors. I am especially interested in the information that is sufficient and necessary for a particular interpretation to be more plausible than others, and how this can be modelled formally. (2) How do agents produce utterances in such a way that certain interpretations are more likely to be selected by the observers of those utterances? In game theory, interpretation is modelled as a selection between possible interpretations. If the Sender of an utterance wishes the Receiver of the utterance to select a certain interpretation, the Sender must provide some evidence for that particular interpretation. As we will see, this evidence can be directly encoded into the utterance, or can be conveyed though the use of what Stalnaker (2002) terms common ground propositions: propositions that are accepted for the purpose of the conversation by the parties to the conversation, without necessarily being believed by any party.
The second theory incorporated into my model is Discourse Representation Theory (DRT) (see Kamp 1981, Kamp & Reyle 1993, 2011). DRT is a formal framework for the representation and analysis of linguistic context. DRT is predicated on the notion that any information introduced into a conversation at a given turn has an effect on the interpretation of utterances in subsequent turns. It defines Discourse Representation Structures (DRSs), which are the basic form of representing contexts. DRSs contain two sets: a set of entities, which point to individuals, objects, concepts, and so on; and a set of conditions that apply to the entities. A context is thus viewed in DRT as a register of the entities that are referenced in a conversation, and the information about those entities that is revealed. DRT also allows for the definition of an intensional model for DRSs; that is, a model that (1) includes multiple possible worlds; and (2) defines the truth of a given DRS relative to a specific possible world. The capacity for the definition of an intensional model, coupled with a method of representing fine-grained differences in utterance semantics, renders DRT extremely useful. I show that DRT can be combined with game theory to endow the latter with a more expressive semantics, while retaining the explanatory capacity of the former.

The third theory that I use is information theoretic steganography (see Cachin 1998, 2004). Steganography is the art and science of hiding information within a cover medium so that the presence of the information is able to be detected only by certain trusted parties. One of the more common methods of steganography is the concealment of information within an image’s data (see Ker 2005). Altering an image’s data necessarily alters the image. So, to successfully conceal information, the data must be altered in such a way that an observer is unable to detect any imperfections in the image. In this thesis, information theoretic steganography provides us with the ability to describe the basic tension of improvised performance that was noted above: actors must produce utterances that signal their intentions for the trajectory of the plot, but that conceal this information from the audience, since this planning information would damage the verisimilitude of the plot. Information theoretic steganography allows us to approach this situation in a way that is, I argue, unavailable when utilising only the formalisms of standard game theory.

The synthesis of these three theoretical frameworks allows for a robust descriptive model of communication in improvised theatre. Improvised theatre may appear at first a strange choice for an investigation of naturally-occurring speech. It may seem to be the case that improvised theatre is a sort of rarefied, abstracted form of spoken interaction, given its artificial constraints, and given the presence of an audience.

It is true that improvised theatre has rules that determine what is possible in a performance, which range from aesthetic values to explicit statements of what is permitted. For instance, in a particularly popular ‘theatre sport’, every line that the actors produce must be a question. Questions answer questions until eventually an actor fails, and is disqualified from the remainder of the scene. This continues until a victor emerges. This is an example of a particularly artificial constraint, but even the presence of such severe constraints is not enough to dismiss the genre entirely as an object of linguistic study. We
need not search far to find genres of naturally occurring speech with potentially severe restrictions on what is possible for such speech. For instance, political speeches of various kinds (see van Dijk 2006) and interviews (see Talmy 2010) are genres of speech that are constrained in different ways, and yet each is the subject of linguistic analysis. In fact, the speech of improvised theatre is more naturalistic than prepared speeches, since it involves interaction between multiple interlocutors, as well as spontaneous invention.

I claim that the constraints of speech in improvised theatre are not problematic, but are valuable precisely because they offer a solid basis for the study of language use. They provide demonstrable, well-defined boundaries on what an actor considers a felicitous choice. Theories abound that place limitations on what interlocutors will consider sensible actions in conversation. For instance, the Gricean programme (Grice 1975, 1989; see Davies 2007 for a good overview) with its cooperative principle and maxims, proposes a set of principles according to which interlocutors believe they can expect their fellows to behave. The expectation that an interlocutor’s fellows will behave in a certain way affects how the interlocutor will behave. As such, Grice’s cooperative principle imposes a set of constraints on the language that Gricean cooperative interlocutors will utilise. Linguists use these constraints as a lens through which to view interlocutors’ actions.

The constraints that are imposed by improvised theatre operate in a similar fashion. Actors expect their fellows to behave in certain ways, and to use utterances for specific reasons. And since they assume that these expectations are shared by their fellows, actors understand that they must conform to these expectations when they select their own utterances. Indeed, improvised theatre has the considerable advantage for analysis of natural language strategies that it is one of the few genres of naturally-occurring speech where a common ground is explicitly constructed in front of the audience and the other actors. This gives the analyst a framework which is distinguished, amongst the various genres of naturally-occurring speech, by its clarity and definiteness.

The improvised performances that interest us here are so-called longform improvisations. Instead of being aesthetically inclined to short, pithy sketches, longform performances have a tendency towards exploring the actions of characters who are more fully realised, in plot-worlds that appear more plausible than those of other types of improvised performance. Actors’ preferences for the use of language follows these tendencies, and therefore, the language of longform performances tends to be far more naturalistic than the language of other kinds of improvised theatre. This is not to claim that constraints on speech are eschewed entirely. As I show, the form requires that actors pretend to be individuals who have different traits from their own, and act as if the material conditions of the plot-world are different from those of the actual world that the performance takes place in, all the while covertly planning amongst themselves. But severe constraints, such as those in theatre sports, are generally not present in longform improvisations.

The central role that speech plays in improvised theatre, and the complex conditions under which the speech is produced, mean that a linguistic analysis of the practice will
1.2 · An overview of the thesis’s argument

In this section, I give an overview of my argumentation in this thesis. This involves: a statement of four fundamental methodological principles that underpin the investigation at hand (§1.2.1); the introduction of the three major claims that I make in the thesis, which serve to direct the course of the argument (§1.2.2); a brief overview of the components of the descriptive model, and how they interact (§1.2.3); and finally, a description of the chapters that constitute the remainder of the thesis (§1.2.4).

1.2.1 · Fundamental methodological principles

Four fundamental principles underpin the methodology that I use to create the model.

(i) A formal model of linguistic practice in improvised theatre is most useful when it is descriptive, rather than predictive.\footnote{I thank the anonymous examiner who pointed this out to me.} The model should thus provide a means of scaffolding any instance of communication in improvised theatre, by specifying the parameters that are minimally necessary for a rigorous description of the practice. A descriptive model is a natural consequence of the thesis’s aim to describe a particular genre of naturally-occurring speech. The standard, predictive approach to game theoretic analyses of language relies on abstract, introspective data of language use as its evidential basis. In general, data of this sort are selected with an established set of parameters in mind that enables for measurement of the data. For instance, the optimality of introspective linguistic data that are used in game theoretic models are often measured according to the Gricean cooperative principle and its maxims (see, for instance, Dekker & van Rooij 2000). Since we are confronted in the present project with a large amount of data, and since I aim to employ game theory to devise a faithful model of these data, my major task is the identification and formal definition of the parameters by which the data can be measured. In other words, this project approaches game theoretic problems backwards: instead of starting with a model and applying it to data, it begins with data and from them builds a model.

This is an important point to bear in mind, for two reasons. (1) Aiming for thorough description has a considerable effect on the decisions that are made in the process of devising the model. For instance, I do not propose a definition of equilibria, nor do I provide criteria for deriving such a definition, for equilibria are used as a means of predicting agent behaviour, as I demonstrate in §2.1.2 with the canonical example of the Prisoner’s Dilemma. When these kinds of decisions are made, I will refer back to this methodological principle. (2) Aiming for a descriptive, rather than a predictive, model is a departure from the standard methodology of game theoretic modelling used in formal linguistics (see, for instance, Franke 2009, Parikh 2006, Asher, Paul...
To those accustomed to this standard methodology, some parts of the model may appear suboptimal, particularly because the primarily descriptive methodology that I use, to the best of my knowledge, the first of its kind for a game theoretic study of linguistic phenomena.

A descriptive model is strongest when it is primarily data-driven. The data are taken from a dataset compiled from four performances of improvised theatre, which in total comprise two hours of performance. Each new formalism that is introduced into the model must be accountable to these data. This allows for a meaningful game theoretic description of communication in the specific genre of improvised theatre, but it also provides an in-depth study of the application of game theoretic modelling to natural language data, and so provides an exemplar for future research.

Agent behaviour is best captured by identifying heuristics. The heuristics that I propose are meant to represent expectations that actors have of the purpose of their fellows’ behaviours; they represent in aggregate the lens through which actors view the actions of their fellows. In game theory, this role is usually taken up by solution concepts, which are algorithms that prescribe certain outcomes as optimal according to given criteria. However, because of the breadth of possible strategies in improvised theatre, I will argue that no single solution concept suffices for the description of actors’ choices. Instead, multiple heuristics interoperate to affect actors’ decisions in different situations.

I propose two kinds of heuristic. The first kind relates to the specific function of certain strategies. A strategy’s function may be considered the ‘end that the utterance is aimed at achieving’. An utterance that is intended to be optimally clear, for instance, has clarity as its function. Heuristics that describe conditions of optimality are expressed as optimal stratagems for the given functions, and are therefore defined formally. The intuition behind this idea is that, if a strategy is optimal, then actors will be expecting that strategy to be played. If optimality is characterised by specific properties, then actors who observe an utterance with those properties will infer that the utterance has a specific function. The second kind of heuristic describes general tendencies that underlie actors’ decisions. These heuristics express the actors’ general assumptions about the use of naturally-occurring speech in improvised theatre. They approximate ‘norms’ of the communicative genre. Since they do not prescribe conditions of optimality, they are not expressed formally.

The final methodological principle is that I should seek to avoid, as much as possible, the introduction of redundant elements into the model. That is, the formal apparatus should have the fewest number of mechanisms possible for a complete description of the phenomena observed.

---

2The performances that constitute this dataset were recorded in adherence to Griffith University’s Human Research Ethics procedures and standards. Approval was granted for this project under protocol number LAL/07/14/HREC.
1.2.2 · Major claims

I make three major claims in the course of the thesis, which capture in essence the progression of its overall argument.

A That common ground in improvised theatre is built through iterative processes of negotiation. Actors propose information which may be accepted into or rejected from the common ground. This process of proposal and judgment of details of the plot is fundamental to an understanding of linguistic strategy in improvised theatre, for it is through this process that a plot is constructed and executed. This claim is justified in chapter 6.

B That for every possible combination of the parties to an improvised performance, multiple, divergent partial grounds are generated. Partial grounds represent the common ground information that is shared by a subset of the interlocutors. I show that this information is generally established prior to a conversation in which partial grounds are active. They are particularly important to the success or failure of interpretations in a conversation, for it is on their basis that certain interpretations become more salient than others. This claim is attended to explicitly in chapter 7, and is informed by the discussion in chapters 5 and 6.

C That actors consciously exploit differences between the partial grounds to transmit information which is inaccessible to the audience. The common ground information available to an interlocutor determines the set of interpretations that that interlocutor deems plausible. It follows that if two interlocutors observe a single utterance but the partial grounds that they share with the producer of the utterance are different, then they will be likely to differ in their judgments of which interpretations are more likely. The claim that actors are aware of this fact, and furthermore that they use this property of language to deliberately cause the observers of the utterance to differ on the proper interpretation of that utterance, becomes far more plausible than it might at first appear. This claim is made, and its formal solution implemented, in chapter 7.

Note that each of these claims is made with specific reference to the communicative situation that actors involved in improvised theatre find themselves in.

1.2.3 · An overview of the model

With the major claims in mind, I now present a broad overview of the model, in order to better position each of its components with respect to the others, and to assist in the

---

3This is a somewhat awkward term, but it captures the essence of the notion. Just as common ground propositions are commonly agreed upon by the parties to a conversation, partial ground propositions are agreed upon by a part of the group of interlocutors. The term also has overtones of exclusion, and even of collusion, from the sense of partiality as bias. The major problem with the term, however, is that grounds is in English both a count and mass noun. I always use partial ground as a count noun, while I use common ground as either a mass or count noun.
comprehension of what follows. Note that this overview purposefully eschews most of the fine detail.

The model of communication in improvised theatre operates on the foundation of a model of improvised theatre itself. Any given plot in an improvised performance takes place within a plot-world. A plot is a series of events, such that these events are possible (that is, their probability is nonzero) in the plot-world. Events are the visible actions that are performed by the actors; these include utterances, gestures, and mime, but crucially do not include interpretations, for interpretation selection is an internal process. The plot is partitioned into scenes, which usually take place in a specific location, and involve a finite set of characters. Scenes are further partitioned into states of the scene, which are of central importance to the model. States of the scene (or, simply, states) are associated with particular plot-worlds, such that only certain states are true at certain plot-worlds. It is therefore primarily at the level of states that the possibility—and, therefore, believability and verisimilitude—of a given action is determined. That is, if an action is not possible in a plot-world, it is not possible in certain states. States shift throughout a performance, through the explicit and purposeful action of actors. A typology of actions and their functions on the states of the scene is proposed as part of the model. Characters are the rational agents in an improvised performance. Characters have traits that are ascribed by the deliberate actions of the actors as part of a performance. Character consistency is of central importance to actors, and for this reason a definition of character consistency is proposed as part of the model.

The game theoretic description of an improvised performance is based on interpretation games, which are signalling games with natural language signals. Since an improvised performance is made up of a large number of conversational turns, the model of a performance is essentially an iterated interpretation game, which ultimately comes to be named an improvised performance game (IPG). Some additional amendments are made to the standard interpretation game model, to account for some of the specific characteristics of improvised theatre. An iterated game allows for a formal account of the effects of an action on the subsequent actions in the scene, and can describe the shifting preferences of an actor as the scene progresses. Iterated games have histories, which are ledgers of the actions that are performed during a play of that iterated game. Since the utility of a play of a game is determined by the actions of all actors, each player’s utility for a play of an IPG is only determined at the termination of the game. But since an actor can only see (a) visible actions (i.e., events, as described above); and (b) her own interpretations, each actor has her own private history. This implies that actors always operate under a very specific sort of incomplete information, and that actors’ estimations of utilities during the iterated game—that is, before the actual utilities are awarded at the conclusion of the game—may differ greatly from those of their fellows.

To describe the language that is used by actors, I use Discourse Representation Theory whose semantics is defined in terms of an intensional model. DRT is also used to model the common ground of the actors, which is a record of the information that has been
introduced and accepted by the actors. The process of introduction and acceptance is a process of iterated negotiation, as per claim A. The common ground is therefore a partial model of the plot-world that contains all publicly known information about that plot-world. The common ground is further decomposed into partial grounds, which represent the information that is shared and agreed to by any subset of the actors and audience, as per claim B.

Actors’ communicative acts must both communicate their intentions for the trajectory of the plot and maintain the appearance of credible communication in the plot-world. Thus, the covert communication described by claim C relies on the fact that there is always a partial ground shared by the actors that is not available to the audience. The contents of this partial ground imply that any utterance that the actors perform must be interpreted by the other actors both at the level of the characters’ communication and at the level of the actors’ planning and coordination communication. I claim ultimately that utility for a successful performance is measured by the extent to which the actors’ communication remains verisimilar; that is, by how successful the actors are in hiding their planning communication.

1.2.4 · Thesis structure
This introductory chapter has summarised the aims and scope of the thesis, foreshadowed key arguments, outlined essential methodological principles, and given a broad overview of the model. The remainder of the thesis is organised into eight chapters.

Chapter 2 provides an introduction to the relevant literature on game theoretic linguistics, and on improvised theatre. The chapter begins with descriptions of those aspects of game theory that are salient to this thesis. In the latter section of this chapter, I introduce relevant literature on improvised theatre. In particular, I highlight a recent trend of formal description of improvised theatre. The limitations of these models in terms of their capacity for the description of language are the principal motivator for the present study.

Chapter 3 introduces a major theoretical component, in Discourse Representation Theory. A means of incorporating DRT as the semantic component of the game theoretic model is also provided in this chapter.

Chapter 4 presents the provisional formal descriptive model of improvised theatre. It is an almost complete specification, in that it lacks the ability to formally account for the concealment of planning information that is almost always present in an utterance. This component requires advancement of the model in the subsequent two chapters before it can be approached directly in chapter 7.

Chapter 5 begins the exploration of the data in line with the model as proposed in chapter 4, and the introduction of heuristics. This chapter is concerned in particular with the means by which plot-worlds are built. This involves a discussion of the way
that information is introduced in improvised theatre, and identifies ways that actors ensure that the plot-world is internally consistent.

**Chapter 6** continues the analysis of the data by focussing on the means by which actors control and exploit the common ground. This chapter’s primary purpose is to introduce and prosecute claim A (that common ground is negotiated), and to formulate a mathematical means of accounting for it.

**Chapter 7** introduces the mechanism of hiding information within utterances, and completes the model of improvised theatre. It is in this chapter that I proffer claims B (that partial grounds exist) and C (that differences in partial grounds are used as the basis for concealing information from the audience). I give an overview of information theory and the information theoretic study of steganography.

**Chapter 8** provides a brief example of how a scene of improvised theatre might be approached analytically using the descriptive model that I have set out. This application of the model demonstrates its strength in the description of natural language.

**Chapter 9** concludes the thesis, and suggests directions for further research.

The general structure of the thesis is one where the theoretical foundations are laid in chapters 2 and 3, the model is proposed in chapter 4, then tested and refined throughout chapters 4, 5, and 6, before culminating in chapter 7. The model is then applied in chapter 8.
Chapter 2

Theoretical grounding

If I succeed in mentally constructing a fortress from which it is impossible to escape, this imagined fortress either will be the same as the real one—and in this case it is certain we shall never escape from here, but at least we will achieve the serenity of knowing we are here because we could be nowhere else—or it will be a fortress from which escape is even more impossible than from here—which would be a sign that here an opportunity of escape exists: we have only to identify the point where the imagined fortress does not coincide with the real one and then find it.

Italo Calvino, *The Count of Monte Cristo*

In this chapter, I establish the theoretical basis and rationale of the project. In §2.1, I introduce game theory, and especially those formalisms salient to the study of language. In §2.2, I review the literature on improvised theatre, with a particular focus on a recent trend in the literature that attempts to build rigorous, formal models of improvised theatre. The chapter concludes in §2.3 with an explicit statement of the rationale for the study.

2.1 · Language and game theory

In this section, I introduce those aspects of game theory which are relevant to models of linguistic interaction. The following sources were consulted when writing this introduction: von Neumann & Morgenstern (1947), Maynard Smith (1982), Gibbons (1992), Weibull (1997), Hargreaves-Heap & Varoufakis (2004), Vincent & Brown (2005), Benz, Jäger & van Rooij (2006), Brams (2011). This section is structured as follows. In §2.1.1, I provide a brief introduction to the elements of probability theory and of possible worlds theory that form the bedrock of game theory, and especially its application to language. The description of games begins in §2.1.2 with an exposition of the basic notions of game theory. In §2.1.3, I discuss games in which the information that is available to the players is restricted, creating the epistemological conditions for games of incomplete information. A class of
game that is singularly important to game theoretic analysis of linguistic interaction, signalling games, is introduced in this section. §2.1.4 introduces the formal apparatus that allows for iterated games. An iterated game is constituted of multiple repetitions of a particular game. The section concludes in §2.1.5 with some reflections on how gaps in the game theoretic study of language might be filled.

These kinds of games are not of equal importance to the model that is devised in subsequent chapters; signalling games and iterated games are of chief concern, while games of complete information are only discussed in this chapter. I use the introduction of games of complete information as an exposition for more basic concepts that are nonetheless present in more complex kinds of games.

It is useful to first make explicit some key assumptions that are held by game theorists and influence their reasoning process. These assumptions also serve to introduce some game theoretic nomenclature. The italicised part of the statements of assumptions 1, 2, and 3 are quoted from Hargreaves-Heap & Varoufakis (2004: 7–32), and the statements of assumptions 4 and 5 are my own.

1. Any given action performed by an agent in a game yields a given result, which is either beneficial, detrimental, or neutral to that agent. In other words, actions have consequences. Agents are aware of these consequences. The metric of benefit or detriment is called utility, and may be defined *ad hoc* to suit the situation being modelled. For instance, utilities may be quantities of physical commodities, but may also be more abstractly defined as numbers that reflect the ordering of players’ preferences concerning potential outcomes of a given game. Players prefer higher to lower utilities, and are indifferent to utilities that are equivalent. Regardless of what a utility represents, it must be a real number.

2. Players of games are rational. Game theoretic rationality refers to the tendency (and ability) of a player to select an action based on the utility that the action will yield. A player may have bounded rationality, meaning that she has limited capacity for reasoning about her moves. A player having bounded rationality implies that the player player’s uses inferential reasoning and hypothesis testing to select actions which she estimates will yield her the highest utility. The important thing to note here is that all such inferences involve a process of deliberation in which the boundedly rational player assesses what is most likely given the evidence at her disposal. That is, players never make injudicious decisions on the spur of the moment.

3. The players have mutual knowledge of rationality. In other words, a player knows that her adversary\(^1\) is rational; the adversary knows that the player knows that she

\(^1\)The term adversary is used in the game theoretic literature to mean any player other than the player under discussion. While this usage is odd, it is a standard term in game theory, *even in games where players cooperate*. I use the term in this description to be faithful to the literature, but later in the thesis I use fellow to emphasise the cooperative nature of improvised theatre.
2.1 · Language and game theory

is rational; and so on, infinitely. Consequently, each player understands that the decision-making process that she goes through in selecting an action to perform is the same as that employed by the other players.

4. A game is a model of a situation, and not necessarily a model of how players should (or will) act in that situation. The prescription of an optimal course of action for all players of a game is determined by the solution to a game, which is arrived at by solution concepts. Solution concepts are algorithms that prescribe a means by which players’ most beneficial strategies might be found. Not every game has a unique solution: depending on the solution concept that is employed, it is possible for games to have multiple potential solutions, or for games not to have any solution. Indeed, a game may have a solution when a certain solution concept is employed, and not when another is used. Furthermore, different solution concepts may differ in their identification of an optimal strategy. Solution concepts operate effectively as models of player deliberation.

5. Utilities are determined by every player’s action. When determining the best course of action, then, each player must take into account the utilities, and thereby preferences, of the other players.

2.1.1 · Probability theory and possible worlds theory

Game theory relies heavily on probability theory and possible worlds theory. Probability theory provides game theory with expressions of uncertainty, which is a core concept of games in which players lack information regarding the motives, and even the moves, of their fellow players. Possible worlds theory is used to model alternative states of the world, or alternative ways things might be. It is used in a very basic fashion in standard signalling games, and in this thesis it also becomes a central part of the more complex Discourse Representation Theoretic semantics that is defined in chapter 3.

2.1.1.1 · Probability theory

A probability space represents a measure of probability over a set of outcomes, and is defined formally in definition 2.1\(^2\). Outcomes are generally thought of as the result of an experiment. For instance, a flip of a coin produces an outcome of either heads or tails. A set of outcomes is called an event.

**Definition 2.1:** Probability space. Let \( \Omega \neq \emptyset \) be a countable set. A probability space is a double \((\Omega, \delta)\) such that \(\delta_{\Omega} : \wp(\Omega) \rightarrow [0, 1]\) is a probability distribution such that \(\sum_{o \in \Omega} \Pr(o) = 1\). \(\Delta(\Omega)\) is the set of all possible probability distributions over \(\Omega\). A probability distribution on \(\Omega\) induces a probability measure on elements of \(\wp(\Omega)\), such that if \(E \in \wp(\Omega)\),

\[
\Pr(E) \equiv \sum_{e \in E} \Pr(e), \tag{2.1}
\]

\(^2\)This definition is based on Gut (2012: 10).
where $E$ is an event. A probability distribution has full support if there is no $e \in \Omega$ such that $\Pr(e) = 0$.

Suppose that we are interested in two events $X, Y \in \varphi(\Omega)$ that are independent; that is, the probability of one event has no effect on the probability of the other. This probability is calculated by multiplying their probabilities together:

$$\Pr(X \cap Y) = \Pr(X) \Pr(Y). \quad (2.2)$$

The conditional probability of an event $X$ given that another event $Y$ has occurred, such that $X, Y \in \varphi(\Omega)$, is defined in definition 2.2.

**Definition 2.2:** Conditional probability. If $\delta$ is a probability distribution over $\Omega$, and $X, Y \in \varphi(\Omega)$ are events, then $\Pr(X|Y)$ is the conditional probability of $X$ given $Y$ as calculated by

$$\Pr(X|Y) = \frac{\Pr(X \cap Y)}{\Pr(Y)}, \quad (2.3)$$

which is defined iff $\Pr(Y) \neq 0$. A probability distribution over $\Omega$ is conditional, denoted $\delta_{X|Y}$, if each element of $\delta_{X|Y}$ is a conditional probability $\Pr(X|Y)$.

From the definition of conditional probability, we can derive Bayes’s equation,

$$\Pr(X|Y) = \frac{\Pr(Y|X) \Pr(X)}{\Pr(Y)}, \quad (2.4)$$

which holds iff $\Pr(Y)$ is nonzero. Bayes’s equation is useful in that it allows us to derive $\Pr(X|Y)$ directly from $\Pr(Y|X)$.\(^4\)

Bayes’s equation is most often used to model the update of an agent’s subjective judgment of a probability upon observation of new evidence. The nomenclature of Bayesian probabilities reflects this use. $\Pr(X)$ is called the prior probability, since it is a judgment of the probability made before the introduction of evidence; $\Pr(X|Y)$ is the posterior probability, because this represents the conditional probability of $X$’s occurrence given the occurrence of $Y$; and $\Pr(X|Y)/\Pr(X)$ is a measure of the degree to which the occurrence of $X$ correlates with the occurrence of $Y$.

Finally, suppose that we are interested in the effect of multiple conditions on a single event. According to Russell & Norvig (2010), this is defined as follows.

**Definition 2.3:** Conditional probability with multiple conditions. Let $(\Omega, \delta)$ be a probability space, and let $X, Y, Z \in \varphi(\Omega)$ be events. The probability of $X$ given $Y$ and $Z$ is solved by

$$\Pr(X|Y, Z) = \frac{\Pr(X \cap (Y \cap Z))}{\Pr(Y \cap Z)}. \quad (2.5)$$

Since $\cap$ is commutative, the inner brackets can be removed from the top term.

---

\(^3\)Based on Gut (2012: 16–17) and Grinstead & Snell (2012: 140–141).

\(^4\)The derivation is as follows: $\Pr(X \cap Y) = \Pr(Y \cap X)$, since $\Pr(X) \Pr(Y) = \Pr(Y) \Pr(X)$. Since $\Pr(X \cap Y) = \Pr(X|Y) \Pr(Y)$, $\Pr(X|Y) \Pr(Y) = \Pr(Y|X) \Pr(X)$. Therefore, $\Pr(X|Y) = \frac{\Pr(X|Y) \Pr(Y)}{\Pr(Y)}$. 

2.1.1.2 · Possible worlds theory

Possible worlds theory is predicated on the notion that there is an infinite number of possible worlds, and the world that we hold to be the actual world is simply one of these. I use the interpretation of possible worlds proffered by Stalnaker (1984: 45) as a fully specified “way things might have been”, or as a “state of affairs”. For a possible world to be fully specified, its identity must be defined by the sum total of every fact that is true in it. This means that two possible worlds differ if at least one fact is not common to both.\(^5\)

This notion of possible worlds lends itself to an interpretation of the semantics of utterances, called possible world semantics, the seminal texts of which include Hintikka (1957a,b), Kripke (1959, 1963b,a). Under this interpretation of semantics, the truth of any given statement is determined relative to a given possible world. For instance, statement (i) is true only in those possible worlds where there is an individual named John who is bald.

(i) John is bald.

There are multiple, distinct possible worlds that have this fact in common while differing in any number of other facts, such as John’s height, for instance. So, according to Stalnaker (1976), (i) denotes a set of possible worlds such that in each of the possible worlds in the set, the statement is true. Any set of possible worlds is called by Stalnaker a proposition, and so a sentence denotes a proposition. When a sentence denotes a proposition, that proposition is called the sentence’s extension by Stalnaker (1976).

In what follows, I often employ the terms possible world and state of the world as equivalent terms, for, just as a single fact alters the identity of a possible world, so too does the alteration of a fact change the state of the world. States of the world play an important part in game theory, for they allow for the formal representation of alternative situations in which certain actions become more beneficial, and others less so.

2.1.2 · Basic notions

Since games are models of situations, the simplest games we can define are games of two players, where each player knows (a) which actions are available to the other, and (b) the other player’s utilities for any given course of action, and therefore her preferences over possible outcomes. Games that satisfy these conditions are called games of complete information. Let us consider an example of such a game in which each player may select between precisely two actions.

Example 2.1: Coin-matching game. Vladimir and Estragon play a game where they each hold a coin out of sight of the other, and select whether to show heads or tails. They reveal their choice simultaneously. If the coins match, Vladimir wins. If they do not, Estragon wins. The prize for winning is to take both coins.

\(^5\)See Menzel (2016) for further information regarding the nature of possible worlds.
The winner of the game gains a coin, and so he is awarded a utility of 1, while the loser of the game loses his coin, and so the outcome yields a utility of \(-1\) for him. This game is a trivial example of a situation that can be defined by a game model, for the choices of both players are largely random, and the deliberation of each player is unsophisticated.

This is also an example of a *static* game, in that the players select and deploy their strategies simultaneously. In static games, both players make their decisions under equivalent epistemic conditions, and there is no progressive introduction of information into the game. A static game of complete information is defined as follows.\(^6\)

**Definition 2.4**: Static game of complete information. A static game of complete information is a tuple

\[
G = (N, \sigma = S_1 \times \ldots \times S_n, U : s \in \sigma \rightarrow \mathbb{R}^n)
\]

where \(N = \{1, 2, \ldots, n\}\) is the set of players, \(\sigma\) is the game’s strategy space, \(S_i\) is player \(i\)’s set of legal actions (or strategies), and \(U\) is a utility function.

The set of players \(N\) is a straightforward set of integers between 1 and \(n\). A strategy \(s_i\) of player \(i\) is a deterministic plan that identifies an action in response to each of her adversaries’ possible actions (Hargreaves-Heap & Varoufakis 2004). So, for instance, in the coin-matching game, Vladimir’s strategy stipulates that (a) he should select heads if Estragon selects heads, and (b) he should select tails if Estragon selects tails. Note that this sentence describes *pure* strategies. Other kinds of strategy exist. In particular, mixed strategies are probabilistic analogues to pure strategies: instead of having a single action in response to her adversary’s actions, \(i\) has a probability distribution over her actions in response to her adversary’s actions. In this thesis I deal exclusively with pure strategies, and each of the following definitions applies only to pure strategies.

Each player \(i \in N\) has a strategy set \(S_i\), which is a collection of the legal\(^7\) strategies that are available to that player. A game’s strategy space \(\sigma\) is the Cartesian product of the \(n\) players’ strategy sets. In a play of a game, each player selects a single strategy from her strategy set. Formally, this is called a *strategy profile*.

**Definition 2.5**: Strategy profile. Let \(N\) be a set of \(n\) players, \(S_1, \ldots, S_n\) be the players’ strategy sets, and let \(\sigma\) be the Cartesian product of all strategy sets. A strategy profile \(s = (s_1, \ldots, s_n)\) is an element of \(\sigma\), and is as such an \(n\)-vector.

Given that strategies stipulate responses to actions, it is useful to have a precise method of indicating that a certain strategy responds directly to another strategy:

**Notation 2.6**: Strategy responses. Suppose we wish to discuss player \(i\)’s strategy particularly as a response to another player \(j\)’s strategy. The strategy \(s_i\) responding to \(s_j\) is denoted \(s_i(s_j)\). I alternate square brackets with parentheses where doing so makes the notation clearer: \(s_i\) in response to \(s_j\), which in turn responds to \(s_k\) is denoted \(s_i(s_j[s_k])\).

---

\(^6\)Based on Benz, Jäger & van Rooij (2006: 10).

\(^7\)Rules of games are not *per se* formalised in the game model. Instead, the scope of the strategies available to the player is constrained by implicit rules. For instance, when selecting a strategy in Paper-Scissors-Rock, a player only has precisely those three strategies available to her, as the rules of the game do not permit any other move.
The utility function $U$ assigns an $n$-vector of real numbers to each strategy profile, where the $i$th scalar of the vector is player $i$’s utility. The vector of real numbers that is output by $U$ is called a utility profile. To make this clearer, suppose there are two players $i$ and $j$ who select $a \in S_i$ and $b \in S_j$, respectively. A utility profile for the play $(a, b)$ is a pair of real numbers, for instance $(1, 2)$. $i$’s utility is the first element of the vector, and $j$’s utility is the second. Since $U$ is a function over entire plays and not single strategies, this is the formal implementation of the game theoretic assumption that all players’ strategies determine the outcome for all players of the game.

**Notation 2.7:** Utility profiles. The mapping of a specific strategy profile $s$ to an $n$-vector is denoted with a lower-case $u$: $u(s)$. Player $i$’s utility for a particular strategy profile $s$ is denoted $u_i(s)$, which is a single real number.

The utility function is defined for each $s \in \sigma$. Definitions can be expressed explicitly, giving cases for each possible value of $s$, as in equation 2.7.

$$
U(s) = \begin{cases}
(1, -1) & \text{if } s = (h, h) \\
(-1, 1) & \text{if } s = (h, t) \\
(1, -1) & \text{if } s = (t, t) \\
(-1, 1) & \text{if } s = (t, h)
\end{cases}
$$

2.7 describes the utilities for the coin-matching game of example 2.1, where Vladimir and Estragon are players 1 and 2, respectively.

A given utility can include some measure of cost, which measures the loss that the agent incurs upon performance of an action. The cost may be zero. Cost is modelled as penalty deducted from the nominal gain, though the subtraction is not usually explicitly represented. That is, a given utility is the result of the subtraction of the cost from the yield of the strategy profile, rather than the cost being a post facto evaluation once utilities have been determined.

2.1.2.1 · Representation of games

The full specification of even a simple game is burdensome. For instance, the following is the tuple for the coin-matching game of example 2.1.

$$
(N = \{1, 2\}, \sigma = (\{h, t\} \times \{h, t\}), U)
$$

where $U$ is given by equation 2.7. Clearly, such a specification for a game with many more possible actions, and therefore possible strategy profiles, strains comprehension.

For this reason, a more succinct matrix representation is preferred, called the normal form of the game (Benz, Jäger & van Rooij 2006). For instance, figure 2.1 overleaf is the
Figure 2.1: The normal-form representation of the coin-matching game.

<table>
<thead>
<tr>
<th>(u₁, u₂)</th>
<th>h²</th>
<th>t²</th>
</tr>
</thead>
<tbody>
<tr>
<td>h¹</td>
<td>(1, -1)</td>
<td>(-1, 1)</td>
</tr>
<tr>
<td>t¹</td>
<td>(-1, 1)</td>
<td>(1, -1)</td>
</tr>
</tbody>
</table>

normal form representation of the coin-matching game. The top-left cell indicates which players’ utilities are which in the utility vectors of the rest of the matrix. It also illustrates which player is the row player, and which is the column player, as the row player’s utilities are, by convention, the first of each pair. The first column and first row contain players’ moves. In §2.1.3, I introduce a tabular form that represents more signalling games.

2.1.2.2 · Game solutions

As I mentioned above, solutions to games prescribe which courses of action yield maximal utility. Formally, a game is solved if there exists at least one strategy profile s that is optimal according to a specific criterion (Allis 1994). A strategy profile that is optimal according to some solution concept is denoted s*; its elements are likewise denoted s₁, ..., sₙ. Solution concepts rely on a formal notion of preference. Players’ preferences are based on the fact that, since utilities must be real numbers, they are ordered by > (greater-than) and < (less-than) relations. If uᵢ(s') < uᵢ(s), i prefers s to s'. Preferences thus order the strategy profiles of a game.

A concept that relates closely to preference, but applies instead to individual strategies, is domination. If strategies sᵢ and sᵢ' are elements of s and s' respectively, and if uᵢ(s) > uᵢ(s'), then sᵢ dominates sᵢ'. There are two kinds of domination that are defined in game theory (Benz, Jäger & van Rooij 2006). Strict domination is a straightforward implementation of the notion of preference. To define strict domination formally, we must first introduce a new notation.

Notation 2.8: i-centric strategy profiles. Suppose that we are interested only in player i’s choice when presented with a fixed series of strategies selected by the other players. Denote this (sᵢ, sᵢ-) where sᵢ- denotes the strategies of the n - 1 other players. Set theoretically, sᵢ- ≡ s \ {sᵢ}. Thus, (sᵢ, sᵢ-) is a profile that is distinguished from (sᵢ', sᵢ-) only by i’s strategy.

Strict domination is formally defined as follows.

Definition 2.9: Strict domination. A strategy sᵢ of player i strictly dominates a strategy sᵢ', denoted sᵢ' ≺ᵢ sᵢ, iff for all strategy profiles s it holds that uᵢ(sᵢ', sᵢ-) < uᵢ(sᵢ, sᵢ-).

There is also a second definition of domination, called weak domination. For sᵢ to weakly dominate sᵢ', two conditions must be met. First, there must exist at least one strategy

---

8In this example, I have included a superscript of the index of the player who performs that move. Such a convention is only necessary when the players’ strategy sets are identical. It is merely for comprehensibility, and has no game theoretic significance.

9Based on Benz, Jäger & van Rooij (2006: 18).
profile such that $s_i$ strictly dominates $s'_i$. Second, for all other strategy profiles, the utility of playing $s_i$ is equal to or greater than that of $s'_i$. In other words, $s_i$ is at least as good as $s'_i$ in every case. The following formal definition is based on Benz, Jäger & van Rooij (2006: 18).

**Definition 2.10:** Weak domination. A strategy $s_i$ of player $i$ weakly dominates a strategy $s'_i$, denoted $(s'_i, s_{-i}) \preceq_i (s_i, s_{-i})$, iff for all strategy profiles $s$ it holds that (1) there exists at least one strategy profile such that $u_i(s'_i, s_{-i}) < u_i(s_i, s_{-i})$; and (2) for all other strategy profiles, $u_i(s'_i, s_{-i}) \leq u_i(s_i, s_{-i})$.

Since it allows players to select between strategies that are as good as one another, weak domination allows players to select the least bad option.

Two solution concepts that are important for game theoretic linguistics are **iterated domination** and **Nash equilibria**. Iterated dominance is based on the iterative application of strict domination to a utility matrix (Ho, Camerer & Weigelt 1998). The idea is this: player $i$ compares each pair of utility profiles in a utility matrix, and deletes (or disregards) those strategies that are dominated for both herself and her $n - 1$ adversaries. The same procedure is performed by all players. $i$ understands that, if strategy $s_j$ dominates $s'_j$ for player $j$, she must disregard $s'_j$ as a contingency, since $j$ will never play $s'_j$ instead of $s_j$. Effectively, $i$ must act as if $s'_j$ is not a move that is available for $j$ to play, and must therefore respond instead to members of $S_j \setminus \{s'_j\}$. The deletion is a stepwise procedure that continues until no further strategies are dominated.

Given that iterated dominance relies on deleting strategies based on strict domination, if there are strategies that are not strictly dominated, it is possible that no unique $s^*$ will be yielded by the procedure. What may result is a number of mutually undominated strategies: in this case, these games are unsolvable by iterated dominance. It is useful in these circumstances to have a solution strategy which does not enforce such stringent requirements as strict domination. A more flexible solution concept relies on calculating the Nash equilibrium of a game. Nash (1950, 1951) proved that every finite game has at least one ‘equilibrium’ solution, such that it is the case for each player that it is not beneficial for her to unilaterally alter her strategy selection. The formal definition is as follows.

**Definition 2.11:** Nash equilibrium. A strategy profile $s^* = (s_i, s_{-i}) \in \sigma$ is a Nash equilibrium if $u_i(s'_i, s^*_{-i}) \leq u_i(s_i, s^*_{-i})$, for all $i \in N$ and all alternative strategies $(s'_i, s^*_{-i}) \in \sigma$.

Unlike iterated dominance, there may be multiple Nash equilibrium solutions for a game (Weibull 1997: 15). As I discuss in §3.1.1, Nash equilibria appear in game theoretic descriptions of language.

### 2.1.2.3 An example game: The Prisoner’s Dilemma

Let us draw these definitions together with the description of a game that is often used to introduce game theory (as in, for instance, Benz, Jäger & van Rooij 2006, Gibbons 1992, Hargreaves-Heap & Varoufakis 2004): the Prisoner’s Dilemma (PD). Its popularity in introductions to game theory derives from the fact that the game depicts the process of ratio-
nalisation of strategy that the players go through, and its solution is intuitively surprising. A play of the PD goes as follows:

**Example 2.2: Prisoner’s Dilemma.** Two criminals have been captured by the police. They are questioned individually and given no chance to communicate. Each is offered the following deal. If the criminal confesses and her partner does not, she will receive no jail time, while her partner will receive a sentence of ten years. If both criminals confess, they will each receive a reduced sentence of seven years. If neither confesses, they will both be indicted on some minor offence for which the police have strong evidence, and therefore will each receive a sentence of two years.

What follows is a definition of each of the entities present in definition 2.4 on page 16 with respect to the PD. The set of players is \( N = \{1, 2\} \). The strategies available to each player are confess (\( c \)) or remain silent (\( r \)); hence \( S_1 = S_2 = \{c, r\} \) and \( \sigma = S_1 \times S_2 = \{(c, c), (c, r), (r, c), (r, r)\} \). The normal form of the game in figure 2.2 captures this information, as well as defining the game’s utility function.

<table>
<thead>
<tr>
<th>((u_1, u_2))</th>
<th>(c^2)</th>
<th>(r^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(c_1)</td>
<td>((-7, -7))</td>
<td>((0, -10))</td>
</tr>
<tr>
<td>(r_1)</td>
<td>((-10, 0))</td>
<td>((-2, -2))</td>
</tr>
</tbody>
</table>

**Figure 2.2:** The normal-form representation of the PD. The row player’s utilities are, by convention, the first of each pair. The prison sentences described in example 2.2 are represented by negative utilities, for they represent a player’s loss.

The PD is used in introductions to game theory because it illustrates the assumptions of game theory that were introduced at the start of this section. Both players are aware that their rational choices lead to certain outcomes, and those outcomes are ranked according to their relative benefit (or, as in this case, relative detriment). Importantly, this game depicts a situation in which each player’s choice affects the utilities received by all players.

Yet perhaps the principal reason that the PD is preferred by writers of introductions to game theory is that its solution is counterintuitive, and demonstrates how game theoretic modelling of scenarios is able to reveal facts about agents’ rational processes that are not apparent when these scenarios are reckoned informally. Upon casual observation, it seems obvious that the prisoners will both remain silent, since it is in their mutual interest to do so. However, let us consider more closely the deliberation of 1. Regardless of what 2 selects, choosing to confess (\( c \)) yields a greater utility for 1. If 2 selects \( c \), 1’s utility for \((c, c)\) is greater than that for \((r, c)\); that is, \(-7 > -10\). The same is true for the cases when 2 selects \( r \); \( u_1(c, r) > u_1(r, r) \), that is, \( 0 > -2 \). A solution thus becomes especially clear if we examine 1’s preference chain over \( \sigma \):

\[
(r, c) \prec_1 (c, c) \prec_1 (r, r) \prec_1 (c, r)
\] (2.9)
So, \( c \) strictly dominates \( r \), and 1 plays \( c \). However, the same process of deliberation is undertaken by 2. 2 prefers to play \( c \) regardless of what 1 plays, so she will always play \( c \). Figure 2.3 illustrates both players’ decisions, and the effect the decisions have on the solution.

Figure 2.3: The Prisoner’s Dilemma: all paths lead to \((c, c)\). An arrow from a strategy profile to another means that a player prefers the latter to the former. 1’s preference relation is represented by vertical arrows, while 2’s is horizontal. Because of the preferences of both players to mitigate loss, the solution tends towards \((c, c)\).

As it transpires the solution to the PD is \( s^* = (c, c) \). Observe that this solution is a Nash equilibrium, since neither player benefits from unilaterally altering her strategy from \( c \).

2.1.3 · Signalling games

The games so far described model scenarios where agents have a view of the situation that is identical to the modeller’s perspective. Each player knows the definition of the utility function for any given \( s \), and is thereby able to reckon the best response to each potential \( s \in \sigma \) for pure strategies, or a probability of each response for mixed strategies.

Linguistic interaction does not fit with these notions, for it is characterised by an interlocutor’s lack of information regarding the epistemic state of her fellows. Fortunately, game theory provides for the modelling of agents’ epistemic states, and especially the effect of a poverty of information on agents’ deliberative processes. Games in which at least one player lacks knowledge of her adversary’s utilities are called games of incomplete information, or Bayesian games (Harsanyi 1967). A lack of information has profound consequences, as a player may have no knowledge of the preferences of her adversaries, and perhaps also lacks information on the utility that she herself will receive from a course of action. Under these conditions, a player is not able to respond in the same way that a player of a game of complete information might. Yet, players of Bayesian games are still rational—though their rationality is bounded—and so their decisions are informed by inferential reasoning, where evidence observed informs a player’s beliefs regarding what might be the value of the utility function for any strategy profile (see Gigerenzer & Selten 2002, Kahneman 2003b,a).

To model inferential reasoning, two additional formal entities are needed: (1) a set of states of the world, and (2) player beliefs, which are calculated using Bayes’s theorem, whence the name of the class of game.

The inclusion of states of the world in the game model is predicated on a simple idea: that certain actions are felicitous only under certain conditions. For instance, as I describe in an example below, a player who is disarming a bomb is interested in cutting the correct
wire to achieve this safely. If the state of the world is one in which cutting the red wire causes the bomb to detonate, then cutting the red wire is not felicitous. The conditional nature of felicitousness of certain actions is also present in natural language, and is the basis of possible worlds semantics. If an interlocutor wishes to say something truthful, then the determination of which utterances are felicitous depends on the state of the world in which she produces the utterance.

In order for a player to be unsure about the true state of the world, we need multiple potential states of the world, and so we need some mechanism for setting states of the world. To do so, we introduce a player who completely lacks agency and rationality, called Nature. Nature’s strategy set is $\mathcal{W}$, the set of all possible worlds, and she selects her moves probabilistically. She has no ‘interest’ in the outcome of the game, since she is not awarded a utility at its conclusion, and so she does not have preferences.

A type of Bayesian game that is particularly important for game theoretic linguistics, and for this thesis, is signalling games. Signalling games were first proposed by Lewis (1969) as a means of studying communication. They are widely used in game theoretic linguistics (see, for instance, Parikh 1991, 1992, 2000, Benz 2012a, Franke 2009, van Rooij 2004a, Benz & van Rooij 2007), because they are apt for modelling the effects of the players’ epistemic states and of linguistic context on the production and interpretation of utterances. Signalling games involve two players, a Sender $S$ and a Receiver $R$, and proceed according to the following general schema.

1. A state of the world $w \in \mathcal{W}$ is set probabilistically by Nature with probability $\Pr(w)$, and only $S$ observes it.
2. $S$ sends a message $m$ from a message set $\mu$.
3. $R$ receives $m$, and performs an action $a \in A$.

Despite its apparent simplicity, the general form of signalling games is suitable for modelling a linguistic context and the effects that are exerted by that context over the actions of both the Sender and the Receiver. Consider the following situation, which constitutes a typical signalling game.

**Example 2.3:** Vladimir phones a friend. *Through a series of improbable events, Vladimir finds himself attempting to defuse a bomb. He phones Estragon, who very conveniently happens to be an expert on explosives. Beneath the shell of the bomb are two wires: a red and a blue wire. Cutting the wrong wire will detonate the explosive. Estragon asks some questions about the bomb’s form, and with perfect aplomb, tells Vladimir which of the wires to cut. Vladimir follows Estragon’s advice, and saves the day.*

This scenario conforms to the signalling game schema above. The state of the world determines which wire will detonate the device, and Vladimir is unsure of the state of the

---

10The constitution of a message is defined according to the needs of the game model. A message may be anything that conveys information. Some examples are: the setting of prices and use of advertising for commodities (Milgrom & Roberts 1986), physical characteristics of animals (Maynard Smith & Price 1973, Maynard Smith 1982), and, as we will see, natural language utterances.
world. Meanwhile, Estragon is an expert in explosives, and so he can view the state of the world. Estragon sends a signal to Vladimir, to reveal which state of the world is effective, and Vladimir follows with an action, cutting one of the wires, which is based on his belief which is formed by observing that signal. As I demonstrate in the following paragraphs, this vignette also captures effects of belief, trust, and information about a possible world encoded within natural language utterances. It therefore represents a complete (if fanciful) utterance context.

Let us model this example formally. There are two possible world-states, one in which cutting the blue wire will disarm the device and in which the red wire will cause the bomb to explode (\(w_b\)) and a state of the world where the reverse is true (\(w_r\)); and so, \(\mathcal{W} = \{w_b, w_r\}\). Suppose that there are also two possible utterances that Estragon can produce: he may suggest that Vladimir cut the red wire (\(m_r\)) or that he cut the blue wire (\(m_b\)); and so, \(\mu = \{m_r, m_b\}\). Likewise, Vladimir can choose to cut the red wire or the blue wire (\(a_r\) and \(a_b\), respectively), so that \(A = \{a_r, a_b\}\). Vladimir is uncertain about the true state of the world, and so is unsure about what he should do. It is clear that a given course of action is felicitous only within certain states. Vladimir ought to cut the red wire iff, in that state of the world, it will defuse the bomb instead of causing it to explode.

In games of incomplete information, players formulate beliefs regarding the true state of the world. Player beliefs are simply probability distributions over sets of possibilities. Each player has beliefs concerning different aspects of the scenario, but we are for the moment only interested in the Receiver’s beliefs regarding the world-states. Vladimir, the Receiver, has no idea which wire to cut, and so he believes \(w_b\) and \(w_r\) to be equiprobable.

The Receiver’s beliefs regarding the possible world-states before she receives the utterance are called her prior beliefs (based on Gibbons 1992: 187–188 and Franke 2009: 26–27).

**Definition 2.12:** Receiver prior beliefs. Let \(\mathcal{W} = \{w_1, \ldots, w_k\}\) be a set of world-states. A prior belief of \(R\) is a probability distribution \(\rho_R(w) = \delta_{\mathcal{W}} \in \Delta(\mathcal{W})\) such that \(\forall w \in \delta_{\mathcal{W}} : Pr(w) \in (0, 1]\), and \(\sum_{w \in \delta_{\mathcal{W}}} Pr(w_k) = 1\). An element of \(\rho_R\) is denoted \(Pr_R(w_k)\).

Both the Sender and Receiver have other beliefs than this single definition, but I delay their introduction until the discussion of solutions to signalling games.

I am now able to give a formal definition of a linguistic signalling game.

**Definition 2.13:** Linguistic signalling game. A linguistic signalling game is a tuple

\[
\langle N = \{S, R\}, \mathcal{W}, \rho_R, \sigma = \mathcal{W} \times \mu \times A, \text{den}, U \rangle
\]  

(2.10)

where \(\mathcal{W}\) is a set of world-states; \(\rho_R\) is the Receiver’s prior belief; \(\mu\) is the message set containing possible moves of \(S\); \(A\) is the action set containing possible moves of \(R\); \(\text{den} : \mu \to \mathcal{W}\) is a semantic denotation function that associates messages to sets of worlds for which they are true; and \(U : \sigma \to \mathbb{R}^n\) is the utility function of the game.

Note that the implementation of semantics in this definition, as a function from \(\mu\) directly to \(\mathcal{W}\), is only provisional, and I introduce a more complex mechanism for the inclusion of semantics in chapter 3.
To represent signalling games succinctly, I adapt a tabular scheme from Franke (2009), which defines each of the entities in definition 2.13 in a more digestible manner. Figure 2.4 defines the signalling game that models the example we have been examining (example 2.3) with arbitrary utilities: both players benefit from Vladimir cutting the correct wire, and suffer greatly otherwise.\footnote{These utilities might seem quite conservative for a game of life and death. But recall that in games where the benefits are not actual units (commodities, for instance), the relative values of the utilities is salient, rather than their absolute values.} The first column lists the worlds that are in the world set $W$, and the second column displays the Receiver’s probabilistic beliefs regarding those worlds. The third and fourth columns state whether a particular $m$’s denotation includes the world of the row; they also list $S$’s possible moves. The rightmost two columns define the utilities for the actions that $R$ may perform, respective to the various states of the world. They tell us that (a) if $a_r$ is selected in $w_r$, or if $a_b$ is selected in $w_b$, then each player receives a utility of 10; and (b) the players receive −10 otherwise.

\begin{table}[h]
\begin{tabular}{cccccc}
\hline
\text{\(w \in W\)} & \text{\(\Pr_S(w)\)} & \text{\(m_r\)} & \text{\(m_b\)} & \text{\(a_r\)} & \text{\(a_b\)} \\
\hline
\(w_r\) & 0.5 & \(\in\) & (10, 10) & (−10, −10) \\
\(w_b\) & 0.5 & \(\in\) & (−10, −10) & (10, 10) \\
\hline
\end{tabular}
\caption{Tabular representation of the signalling game described in example 2.3.}
\end{table}

2.1.3.1 · Solutions for linguistic signalling games

Signalling games are useful for modelling interaction in game theory, both because they model the effects of the situation on an interlocutor’s deliberation, and because the message may be a natural language utterance. The second fact complicates the analysis somewhat, since such signals must be interpreted by $R$, and solutions to linguistic signalling games must implement interpretation. A simple way to do so is the use of a Lewis utility function, proposed by Lewis (1969), which is defined as follows.

\textbf{Definition 2.14: Lewis utility.} Let \((w, m_S, a_R)\) be a strategy profile of a signalling game. For each world \(w \in W\), it is the case that

\[ u(s) = \begin{cases} 
(1, 1) & \text{if } s = (w, m_S(w), m_R(m_S[w])) \\
(0, 0) & \text{otherwise.} 
\end{cases} \]  

(2.11)

That is, the only beneficial strategy profiles are those where $S$’s message is a response to the state of the world, and where $R$’s action is a response to $S$’s message and the correct world-state. It is important to observe that $w$ is only visible to $S$. Thus, linguistic signalling games are successful iff the correct world-state is transmitted from $S$ to $R$. There is no cost associated with sending signals in Lewis’s signalling games. This continues to be common
practice in game theoretic linguistics, with the notable exception of Parikh (2000, 2006, 2007), whose formulation equates cost with utterance complexity. Unless otherwise stated, I assume that messages have zero cost.

Game theoretic beliefs. Despite the apparent simplicity of the Lewis utility function, even the most basic linguistic signalling games (and Bayesian games more generally) have complex solutions, since solutions for these games rely on the players’ belief structures and rely on Bayesian inference more than solutions to games of complete information do. I have already defined one component of the game theoretic belief structure, namely the Receiver’s prior beliefs (definition 2.12). There are two other components that I wish to address: the Receiver’s posterior beliefs, and the Sender’s beliefs (the formal definitions of which are based on Gibbons 1992: 187–188 and Franke 2009: 26–27). Although these are formally defined, they are not included in the formal definition of a game, and are used only in solution concepts. That is to say that these definitions of belief are more important in analyses of the deliberations of players than in the definition of the situation that per se constitutes a game.

Let us resume our model of example 2.3. Vladimir has not only prior beliefs, but also a belief that he forms as a result of observing the message that is sent by the Sender. This is called his posterior belief.

**Definition 2.15:** Receiver posterior beliefs. Let \( W = \{w_1, \ldots, w_k\} \) be a set of world-states, \( \rho_i(w) \) be a prior belief of player R, and \( m \in \mu \) a message sent by player S. A posterior belief of R is a fully supported probability distribution \( \rho_R(w|m) = \delta_{W|\mu} \in \Delta(W) \) such that each member of \( \delta_{W|\mu} \) is the Bayesian probability of a \( w \) given \( m \), \( \Pr_R(w|m) \).

R’s posterior beliefs are a Bayesian update of her prior beliefs, meaning that she incorporates observations about the world into his estimation of the probabilities of which world-state might be true. Vladimir, for instance, updates which world he considers more likely—and thereby which wire to cut—when Estragon tells him which wire will disarm the bomb.

The Sender also has beliefs. Although she is certain of the state of the world, she is uncertain of what the Receiver is going to do, and so she holds beliefs regarding the effect that her utterance will cause.

**Definition 2.16:** Sender beliefs. Let \( A = \{a_1, \ldots, a_k\} \) be a set of legal actions for R. A Sender’s beliefs are a probability distribution with full support \( \rho_S \equiv \delta_A \in \Delta(A) \).

The Sender’s belief is not conditional, because it describes his guess of the Receiver’s prior belief. As such, it provides the Sender with a basis for deciding the effect that any given signal will have on the Receiver’s choice. For instance, if the Sender believes that the Receiver will play a certain action with absolute certainty, then her utterance selection means less than if she believes that the Receiver is equally likely to play any of her available actions. In the case of Vladimir and Estragon, Estragon believes correctly that Vladimir is
completely uncertain which wire to cut, and therefore that the probability of cutting either of the wires is 0.5. This means that he expects that the probability of Vladimir cutting one of the wires will become greater if he tells Vladimir to cut that wire. Since the Sender’s belief is a prediction of what the Receiver’s prior belief is, Estragon can attempt to steer the Vladimir’s inferential reasoning.

Each player thus has some uncertainty regarding the actions that will be performed by the other player, a property which is common to Bayesian games. Players are not able to formulate responses reliably if they are unable to predict their adversaries’ behaviour. For this reason, players in Bayesian games have only expectations, not information, of the value of the utility function for a given strategy profile. What results is that each player has, for each $s \in \sigma$, an expected utility (from Benz, Jäger & van Rooij 2006: 6).

**Definition 2.17**: Expected utility. Let $W$ be a set of worlds, $i$ a player, and $\rho_i$ her beliefs. The expected utility for player $i$ of strategy profile $(w, s_i, s_{-i})$ is calculated by

$$
\hat{u}_i(s_i, s_{-i}) \equiv \sum_{w \in \rho_i} \Pr(w) \times u(w, s_i, s_{-i}).
$$

(2.12)

The chief consequence of this definition is that, since $\Pr(w)$ is the multiplier, strategies in more probable worlds tend to yield utilities which are more distant from zero than the same strategies in a world with a low probability.

**Bayesian Nash equilibria**. Bayesian Nash equilibria (BNE) are conceptually close to Nash equilibria in games of complete information, in that they model a player’s indifference towards unilateral alteration of her strategy. But since Bayesian games are characterised by players’ belief structures, BNEs—especially those in signalling games—have properties that games of complete information do not (Gibbons 1992: 187–8). For a strategy profile $s$ to be a signalling BNE, the following requirements must be met.

1. $R$ must have a posterior belief. That is, $R$ must update her beliefs regarding the world-state upon observing $S$’s message. As such, a $\rho_R(w|m)$ must exist for each $w \in W$ and for each $m \in \mu$. This condition enforces sequential rationality. If an agent is sequentially rational, she factors observed moves into her deliberation. In signalling games, the Sender is sequentially rational iff she uses her observation of Nature’s selection of $w$ to influence the choice of message that she sends. Likewise, the Receiver is sequentially rational iff she updates her beliefs as a result of the message that is sent to her. Note that she cannot observe Nature’s selection of $w$, and therefore she can still be sequentially rational even if she does not respond to the correct $w$. The condition of sequential rationality may appear to be arbitrary, so let us consider the argument of Farrell & Rabin (1996). If the Receiver does not update her beliefs after observation of the message—in other words, if the message has no effect on her selection of a strategy—then an equilibrium exists where the Sender
produces some meaningless babble. Since it does not matter what the Sender produces, she may as well produce something which avoids even the cost of articulating phonemes!

(2) For each \(m \in \mu\), \(R\)'s action \(a^*(m)\) must maximise her expected utility, given her belief \(\rho_R\).

(3) For each \(w \in \mathcal{W}\), the Sender’s message \(m^*(w)\) must maximise her expected utility, given the Receiver’s strategy \(a^*(m)\).

Let us define these requirements formally.

**Definition 2.18: Signalling Bayesian equilibrium.** A pure-strategy, perfect Bayesian equilibrium in a signalling game is a pair of strategies \(m^*(w)\) and \(a^*(m)\) and a belief \(\rho(w|m)\) satisfying the following requirements:

1. \(\rho_R(w|m)\) must exist;
2. \(a^*(m_i)\) must solve \(\max_{a \in A} \sum_{w \in \mathcal{W}} \Pr(w|m) \times u_R(w, m, a)\);
3. \(m^*(w)\) must solve \(\max_{m \in \mu} u_S(w, m, a^*(m))\); and
4. for each \(m \in \mu\), if there exists \(w \in \mathcal{W}\) such that \(m^*(w) = m\), then the Receiver’s belief at the must follow from Bayes’s rule, and from the Sender’s strategy: \(\rho_R(w|m) = \frac{\Pr(m|w) \Pr(w)}{\Pr(m)}\).

The definitions that I have given so far in this section suffice to define a simple signalling game with linguistic strategies and that can capture linguistic context, although in a very simple fashion. I now turn to an example of how such a model might be defined.

### Example of a linguistic game model

Let us reconsider the bomb scenario. Recall that if \(R\) plays \(a_r\) in \(w_r\), or \(a_b\) in \(w_b\), both players receive a utility of 10. If \(R\) plays \(a_b\) in \(w_r\), or \(a_r\) in \(w_b\), both players receive \(-10\).

When one views the tabular representation of the game, the solutions are obvious: \((w_r, m_r(w_r), a_r(m_r[w_r]))\) (if it is the red wire that will defuse the bomb, Estragon tells Vladimir to cut the red wire, and Vladimir complies) and \((w_b, m_b(w_b), a_b(m_b[w_b]))\) (the same is true, mutatis mutandis, if the blue wire will disarm the device). These solutions are BNEs, since \(\rho_R(w|m)\) exists and it is a Bayesian probability, and since both \(a^*(m)\) and \(m^*(w)\) are maximal in these solutions; I justify these assertions shortly. Since we do not learn much from these solutions, they are not particularly interesting. But, as we will see with most games, even prosaic solutions offer insight into agents’ processes of deliberation.

At the beginning of the game, Estragon (\(S\)) views the world-state. This is the game theoretic equivalent of stating that he is an expert in some domain—in the present case, explosives—and so he knows which wire disarms and which detonates the device. Since Estragon believes that Vladimir (\(R\)) is completely uncertain about which wire to cut, Estragon believes that Vladimir will cut the wire that he is told to cut, which is to say
Estragon believes that Vladimir trusts his expertise. Vladimir’s trust is expressed in the values of his Bayesian update. For instance, if Estragon utters $m_r$, then Vladimir’s posterior belief states that $\Pr(w_r|m_r) = 1$, and that $\Pr(w_b|m_r) = 0$. That is, Vladimir takes what Estragon says to be absolutely true. Vladimir’s decision is simple. If $w_r$, he should send $m_r$, and if $w_b$, then he should send $m_b$.

On the other hand, Vladimir is not an expert in explosives, which is equivalent to stating that he is unable to see the world-state. He is completely unsure and so he assigns an equal probability to both worlds, so that $\Pr_R(w_r) = \Pr_R(w_b) = 0.5$.

Since moves in signalling games are played sequentially, Vladimir updates his beliefs when he observes the message. Vladimir’s belief that Estragon is a domain-expert is encoded in his posterior belief. If Vladimir believes that Estragon is credible, he updates the probabilities of the worlds in $\mathcal{W}$ so that they conform to the advice that he is given. So, $\Pr_R(w_r|m_r) = 1$, $\Pr_R(w_r|m_b) = 0$ and likewise for $w_b$. His selection of action thereby becomes akin to a game of complete information, since he has become aware of which world-state is true.

The probabilities in this scenario are simple, and so it follows that the expected utilities are too. As formulated above, $R$’s beliefs dictate that if he observes $m_r$, he holds that $w_r$ is the world state with probability 1, and likewise for $m_b$ and $w_b$. Suppose that $S$ plays $m_r$. $\Pr_R(w_r|m_r) = 1$, and therefore $R$’s expected utility is

$$\hat{u}_R(w_r, m_r, a_r) = (1 \times 10) + (0 \times (-10)) = 10.$$  

(2.13)

Likewise, since $\Pr_R(w_b|m_r) = 0$,

$$\hat{u}_R(w_r, m_r, a_b) = (0 \times 10) + (1 \times (-10)) = -10.$$  

(2.14)

The same is true for each combination of world states, messages, and actions. What results is that the expected utility for cutting the wire that the Sender says to cut is always greater than cutting the other wire.

As this example demonstrates, signalling games provide a model of the linguistic conditions under which an utterance is interpreted. However, it is most often not the case that an utterance in a normal conversation is followed by an action in the sense that has been discussed. Instead, utterances are followed by utterances, and this occurs over a number of iterations. As such, if we are interested in modelling a conversation of natural language, it seems appropriate to stipulate that the Receiver’s move be a natural language utterance from the same set as the Sender’s move, $\mu$. However, I propose that it is more felicitous to stipulate that the Receiver’s move be an interpretation, rather than an utterance. Generally in game theoretic models, if the signals are natural language utterances, interpretation of the Sender’s utterance is always included, but the interpretation manifests indirectly, as an effect on the action that the Receiver selects. Modelling the explicit selection of an
interpretation allows for finer-grained models. The precise form of interpretations will be discussed in §3.2.

2.1.4 · Iterated games

In each of the examples that have been discussed so far in this chapter, the games have been played exactly once. Players of these games are aware that they have precisely one decision to make, and that the utility depends on that decision alone. In the abstract situations that we have examined, this has been sufficient, but ideally we would like to build models of conversations that last many turns, so that the long-term effects of the accumulation of information on players’ strategies can be studied.

To this end, I introduce iterated games (also called repeated games). The study of iterated games has a rich history in economic games (see Fudenberg & Maskin 1986 and Abreu, Pearce & Stacchetti 1990 for broad applications of iterated games in this field), and has, more recently, found application in linguistic research (McCready 2015, Asher, Paul & Venant 2017, Asher & Paul 2016, Asher & Quinley 2012, McCready & Asher 2013). An iterated game is simply a game that is derived from playing a certain game, called the stage game, a specified number of times. Iterated games are most often used in economics to model the effects of cooperation between firms, especially over long time periods (see, for instance, Fudenberg & Maskin (1986), Abreu (1988), Abreu, Pearce & Stacchetti (1990), Binmore & Samuelson (1992) for papers indicative of such applications). In this subsection, I introduce the aspects of iterated games that are salient to their use in this thesis. The formulation here is based on the models devised by Mertens (1986).

2.1.4.1 · Iterated games of complete information

Let us first address iterated games where players of the stage game have complete information. Since the class of game that is studied in this thesis is a game of incomplete information, I introduce these later.

The stage game. Let the stage game be a static game of complete information \( \langle N, \sigma, U \rangle \), such that \( N = \{1, \ldots, n\} \) is a set of players, \( \sigma = S_1 \times \ldots \times S_n \) is a strategy space, and \( U : \sigma \to \mathbb{R} \) is a utility function. We assume that \( S_i \) is finite for each \( i \in N \).

The iterated game. Let \( G^T \) be a game obtained by \( T \) repetitions of the stage game \( g \); \( T \) is called \( G^T \)'s horizon. If \( T = \infty \), the game is infinitely iterated, and we denote the game \( G^\infty \). I discuss infinitely iterated games in more detail below. Each repetition is called a period of the game, where any given period of the game is indexed by an integer. To reference an arbitrary period, I will use a superscript \( t \).
A history $\eta^t$ is a series of actions that precede period $t$, so that $\eta^t = (a_1, \ldots, a_{t-1})$. $\mathcal{A}^t$, which should be read as the set $\mathcal{A}$ to the $(t-1)$th power, is the set of all possible $t$-histories.

In studies of iterated games, it is usually the case that pure strategies for player $i$ are defined as sequences of functions $s_i(1), \ldots, s_i(t), \ldots, s_i(T)$, such that there is a function for each period. The pure strategy for $i$ in period $t$ is specified by a function of the actors of all players in all antecedent periods.

**Definition 2.19:** $i$’s pure strategy at period $t$. A pure strategy for player $i \in \mathcal{N}$ at period $t$ is defined by

$$ s_i(t) = \begin{cases} a \in \mathcal{A}_i & \text{if } t = 1 \\ \mathcal{A}^{t-1} \rightarrow \mathcal{A}_i & \text{if } 1 < t \leq T \end{cases} \quad (2.16) $$

where $T$ is the horizon of $G^T$.

$\mathcal{S}_i^T$ is the set of all strategies for $i$ over the $T$ periods, and $\sigma^T \equiv \bigtimes_{i \in \mathcal{N}} \mathcal{S}_i^T$ is the strategy space of $G^T$. A strategy profile $s$ of an iterated game is an element of $\sigma^T$ in the same way that $s \in \sigma$ for non-iterated games.

I do not utilise the definition of strategy given in definition 2.19 in my model; instead, common ground models the effects of strategies in earlier periods on the selection of strategies in later ones. I have introduced it here simply to highlight one of the key advantages of a model that uses iterated games: the ability to account for cumulative effects of knowledge on players’ strategies. A strategy at period $t$ depends on the actions that the actor selected during the $t-1$ periods that preceded it. The stepwise introduction of further information, and its effects on player decisions, may be represented deterministically by this extremely simple function.

The utility of an iterated game is simply the sum of the utilities that are awarded for the $T$ iterations of the stage games.

**Definition 2.20:** Utility of an iterated game. Let $G^T$ be an iterated game with horizon $T$. The utility for player $i$ of $G^T$ is evaluated by

$$ u_i(\eta^T) = \sum_{t=1}^{T} u_i(s^t), \quad (2.17) $$

where $\eta^t$ is a particular history of $G^T$.

I assume in what follows that players are aware of the utilities that they have been assigned at the conclusion of each period. This means that the selection of an agent in a given period is strongly influenced by the actions that she and her fellows have selected in the preceding periods.

\[\text{That is, the Cartesian product} \quad \mathcal{A}^t = \mathcal{A} \times \ldots \times \mathcal{A}. \quad (2.15)\]
Note that I have not included a discount factor. In the study of iterated games in the field of economics, it is often the case that the utility for any given period of the iterated game is multiplied by a number in the interval $(0, 1)$. This is called the discount factor, and it is meant to represent an agent’s or firm’s relative weighting of that period in the iterated game. In general, relative to a given period $t$, the utilities of distant periods are weighed less than periods immediately following $t$. This means that profits and losses that are further in the future are of lesser concern to the players than those that are nearer. The discount factor is also used to formally account for long-term benefits of cooperation between firms. I have omitted this detail because it is irrelevant to formulations of iterated games for the analysis of improvised theatre. What the omission of the discount factor means is that agents value utilities that are gained in the present equivalently to those that are to be gained in the future. And so, agents will tend not to use strategies that delay benefit, but they instead generally use strategies that maximise their utility at each period.

Infinitely iterated games. The designation that an iterated game is infinitely long is a misnomer. An infinitely iterated game does not need to be played an infinite number of times. It may end, but it does so at a random value of $T$. As such, the players are unaware of whether any given period will be the final period of that game. So, in effect, the players of an infinitely iterated game must act as if the game continues indefinitely. I demonstrate below that this is a satisfactory understanding of a conversation. At a given point in a conversation, the interlocutors do not know when it will end, and must select their strategies accordingly. For instance, this fact is taken as the basis for arguably the most important theory of linguistic pragmatics, the Cooperative Principle formulated by Grice (1975: 45): “Make your contribution such as required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged”. The foundation of all Gricean cooperation depends on an interlocutor’s awareness of the stage of the conversation, and the accepted purpose or direction of the exchange. This fits precisely with the definition of strategy selection in iterated games, where the selection of a strategy at $t$ depends on the moves that have preceded that period. In iterated interpretation games, any utterance and any interpretation would be conditioned by those of each that have preceded that particular period. With iterated games, then, we can represent changes in direction and shifts in shared goals that arise over time as an effect of conversation. I show that this is an especially salient property of the kind of improvised theatre that constitutes the dataset used throughout this thesis.

If games are truly infinite, it is not feasible for players to calculate their utilities, which is a vital component of the players’ selection of actions in subsequent periods of the game. As such $i$’s utility is given by evaluating the limit of the utility for a history $\eta^T$ as $T$ tends to infinity. In practice, conversations are finite, and so I eschew this overly complicated method of calculating utility, and develop my own measure of utility that is appropriate for improvised theatre, which I introduce in chapter 4.
2.1.4.2 · Incomplete information in iterated games

The stage game that will be used in the description of improvised theatre is a signalling game. As such, players do not possess complete information, and we therefore must include provisions for incomplete information. Such provisions are defined here, and are based on the research of Aumann, Maschler & Stearns (1995) in their monograph on the subject. However, what I introduce is highly simplified, and represents the sufficient and necessary degree of complexity for the project at hand. Also, I only discuss games with incomplete information on all sides, such that all players lack information about the strategies of their fellows.

Private histories. An iterated game of incomplete information differs from an iterated game of complete information in that the history of the game is split so that each player has her own private history. A player’s private history represents the strategies that she has observed. As I show, certain strategies are visible only to the player who selects them. For instance, interpreting an utterance is an action that is visible only to the agent who does the interpreting, so interpretations should only be in the history of the player selecting the interpretation, and should be excluded from other players’ histories. And so, differences emerge in players’ histories. Unlike iterated games of complete information, there is no history of the game as a catalogue of the strategies played up to a given period, for no player has access to it. Instead, the history represents the total set of combinations of private histories.

Definition 2.21: Private histories. Let $G^T$ be an iterated game with $T$ periods, and let $g$ be a stage game of $G^T$. The private history of player $i$ at period $t$, denoted $\eta^t_i$, is an element of $(\sigma_g)^{t-1}$ (that is, $\sigma_g$ to the $(t-1)$th power). At the beginning of an iterated game, the players’ histories are all equivalent to $\emptyset$. The history of $G^T$ at $t$ is the Cartesian product of the players’ private histories, such that $\eta^t \equiv \times_{i \in N} \eta^t_i$.

Providing a definition of private histories for each player allows for disjunct private histories, and therefore for the effects of differences in information available to each player. Recall that in games of incomplete information, players can be confused about which strategies were played prior to their move. For instance, in signalling games the Receiver does not know which state of the world has been selected by Nature, and so is uncertain of her best response to the Sender’s strategy. Similarly, in iterated games with incomplete information, players are confused to some extent by the moves that have preceded theirs; but this confusion is instead between histories rather than single moves.

According to Aumann, Maschler & Stearns (1995), at the conclusion of a stage game with incomplete information, each player is informed of her private history, which may be distinct from the actual moves that were played in that turn. This conceit means that “except for the information implicit in [the players’ histories], no information regarding the [actual strategies selected] is ever revealed to anybody” (Aumann, Maschler & Stearns 1995: 179). This may appear at first to be an unrealistically stringent restriction on the
information that is available to the players. But observe that this is the implementation of
the incompleteness of information that is present in signalling games to the more complex
evironment of iterated games. Players have the choice between strategies in each period,
but they also must choose which play of the game at a given period is optimal with respect
to their preferences.

Note that I have opted here for a formulation of incomplete information such that no player
has information regarding the true strategy profile that is played. This differs
somewhat from non-iterated signalling games, where only the Receiver has incomplete
information, since she does not view the world-state. I have done so to give a basis for a
redefinition of histories in the following chapter, which deals with games where players
select interpretations.

Beliefs. Beliefs can change in the course of iterated games in that the beliefs of each player
at any period are conditional upon her personal history. The pure strategies defined in
definition 2.19 above are a crude precursor to this redefinition of belief, in that an action
in a given period is a function that relies on the action of the period before.

I propose the following redefinitions of belief for iterated games. The Receiver’s prior
belief is redefined so that it is conditional on her history if her history is nonempty.

**Definition 2.22:** Receiver prior beliefs. Let $G^T$ be an iterated game, let $\eta$ be a set of histories
for $G^T$, and let $W = \{w_1, \ldots, w_k\}$ be a set of world-states. A prior belief of $R$ is a probability
distribution such that

$$\rho_R(w) \equiv \begin{cases} \delta_W & \text{if } \eta^t_i = \emptyset \\ \delta_{W|\eta^t_S} & \text{otherwise.} \end{cases}$$

(2.18)

The Receiver’s posterior beliefs are naturally an update of her prior beliefs upon observa-
tion of the message that is produced by the Sender.

**Definition 2.23:** Receiver posterior beliefs. Let $G^T$ be an iterated game, $\eta$ be a set of histo-
ries, $W = \{w_1, \ldots, w_k\}$ be a set of world-states, $\rho_R(w)$ be a prior belief of $R$, and $m$ a
message sent by $S$. A posterior belief of $R$ is a probability distribution such that

$$\rho_R(w|m, \eta^t_R) \equiv \begin{cases} \delta_{W|m} & \text{if } \eta^t_S = \emptyset \\ \delta_{W|m, \eta^t_R} & \text{otherwise.} \end{cases}$$

(2.19)

The Sender’s belief is changed only by rendering her expectations of the behaviour of the
Receiver conditional on the history.

**Definition 2.24:** Sender’s belief. Let $A = \{a_1, \ldots, a_k\}$ be a set of legal actions for $R$. A
Sender’s belief is a probability distribution such that

$$\rho_S \equiv \begin{cases} \delta_A & \text{if } \eta^t_S = \emptyset \\ \delta_{A|\eta^t_S} & \text{otherwise.} \end{cases}$$

(2.20)
These three definitions stipulate that as long as there is a history, the history has an effect on the beliefs of the agents. In practice, the only time the history is empty is prior to the first period of the iterated game.

**Utility.** The utilities of an iterated game of incomplete information are calculated by summation of the utilities for the stage games. However, in each period no player knows which strategies have been truly played. So, their knowledge of the utilities is an estimation based on their private histories. And since agents know that they are playing a game of incomplete information, they know that this is the case. In the model defined by Aumann, Maschler & Stearns (1995), knowledge of the actual utility that has been accrued throughout the iterated game is withheld until its conclusion. While this might appear to be an unrealistic constraint, it is simply the application to iterated games of the principle outlined above for single games of incomplete information, whereby a player can only expect that her actions will yield a particular utility, until the end of the game, when she receives her actual utility. In the next chapter, I provide a definition of utility that is more sensitive to the desires of actors engaged in improvised theatre.

### 2.1.4.3 · Iterated games in linguistics

As indicated above, iterated games have found some application in models of language use. Asher, Paul & Venant (2017) define a class of infinitely-iterated games called Message Exchange (ME) games, chiefly to model interactions where the participants are not game-theoretically cooperative. ME games are adapted from Mazur-Banach games (see Soare 2016), which are games specified by an infinite set of strings, with a winning condition (Asher, Paul & Venant 2017: 372). The adaptations that Asher, Paul & Venant (2017) make include specifying a set of strings and a winning condition for each player. ME games are infinitely repeated, and so the termination of the game does not occur at a fixed turn. Instead, the game ends when a player selects a strategy that satisfies the winning condition wins the overall game. The player that selects this strategy wins; or, if there are no strategies that satisfy the winning condition, neither player wins. A model that is closer in form to the sort of iterated game that I define in later chapters is proposed by McCready (2015), whose model is concerned primarily with how interlocutors establish reliability throughout a conversation. McCready (2015) uses iterated signalling games to establish a means by which cooperativity may arise even if interlocutors are antagonistic, by defining a mechanism by which interlocutors are punished for uncooperative or untrustworthy speech.

Although I do not employ the formalisms of these models in my own model, they provide alternative methods of modelling conversations that rely on iterated games, and validate it as a modelling approach. However, since improvised theatre has characteristics that are highly specific to the genre, a more specified suite of definitions must be provided for a satisfactory, descriptive model of this communicative practice.
2.1.5 · Conclusion

In this section, I have introduced those aspects of game theory that are salient to formal treatments of natural language, and especially to what follows in this thesis. I have introduced the basic elements of game theory, which allows for the development of formal models that capture agents’ decisions when they stand to gain or lose some measure of utility. Signalling games have been described as the basic class of game that is utilised throughout this thesis. As has been shown, such games are able to model simple linguistic interactions, where a Sender selects between utterances to communicate the world-state, aiming to cause a Receiver to perform certain actions that are beneficial to both players. Finally, the formal apparatus of iterated games has been introduced. I use iterated games in later chapters to grant the capability for the description of conversations that are constituted of multiple signalling games.

The examples that have been used in this exposition demonstrate the capability that game theoretic models have for the description of language. In Vladimir’s Bomb Dilemma, for instance, the game theoretic model was able to represent a very basic semantics of the message, as well as the communication of facts about the state of the world by the Sender to the Receiver. Game theoretic linguistics shows promise as a mathematically rigorous, general theory of language use, and may prove particularly valuable as a theory that can accommodate a number of different parts of an interlocutor’s linguistic faculty in its purview. However, it becomes apparent early in the chapters that follow that, in order to thoroughly account for the data, certain extensions and amendments must be made to the standard game theoretical framework currently in use. The extensions include the incorporation of other theoretical frameworks into the game theoretical model. Specifically, I integrate Discourse Representation Theory and information theoretic steganography into the game theoretic model to enable an adequate description of improvised theatre.

2.2 · Improvised theatre

Now that the foundations have been laid for the game theoretic part of the model, let us turn to a discussion of improvised theatre, which will be the genre of speech that is described by the model in subsequent chapters.

In this section, I introduce literature on improvisation in general in §2.2.1. I use existing research on improvisation, particularly musical improvisation, to provide some indication of the value of studying situations that involve improvisation for understanding their cognitive aspects. In this subsection, I also briefly discuss what is perhaps the most fully-realised mathematical model of improvisation (Pressing 1984), the limitations of which have influenced the design of the model presented in this thesis. In §2.2.2, I discuss improvised theatre in particular. I examine recent work that views improvised theatre as a series of decision theoretic problems, and which has some affinity with the project I have undertaken. I provide an overview of the approach to improvised theatre that is taken in existing work, particularly in relation to the formation and assignment of characters.
§2.2.2 makes clear an important motivating factor for the present study in relation to its contribution to the understanding of language in improvised theatre, for it points out the relative poverty of research into this topic. While other aspects of improvisation and improvised theatre have been analysed, there is very little thorough work on the selection and interpretation of utterances in improvised performance. Finally, in §2.2.2.4, I introduce the specific form of improvised theatre that make up the dataset, called the Harold, or Harold longform.

2.2.1 · Improvisation

Definitions of improvisation are generally formulated by researchers to suit the field they are working in. Moorman & Miner (1998: 700–2) collate a number of definitions, and generally define improvisation as any endeavour in which “composition converges with execution”, where composition covers planning, writing, and devising. Well-executed improvisation is considered to involve acting spontaneously, and creatively, within certain structural constraints, which may confine the improvisers’ choices to varying degrees.

The vast majority of scholarship regarding improvisation as artistic practice describes musical improvisation (see Pressing 1988: 141–7 for a very comprehensive overview up until that year), and so most of the discussion in this section will examine musical improvisation. Artistic uses of improvisation are generally restricted to certain genres or styles within artistic fields. Research into improvisation as an artistic practice tends to focus on its quality as a process of continuous, serial composition (Mendonça & Wallace 2004). This is often contrasted to more ‘traditional’ composition, a “discontinuous process of creation and iteration ... completed in instalments by a composer” (Sarath 1996: 2). An improviser’s decisions are, like a composer’s, governed by structural (see Johnson-Laird 1991, 2002), aesthetic (see Järvinen 1995), and technical considerations (see Palmer 1997). The structural improvisation process may additionally include audience feedback and expectations, as well as feedback from the other members of the improviser’s performing group (see Schmuckler 1990).

From interviews with jazz musicians, Sawyer (1992) notes that two major characteristics of jazz performance are two tensions in the performance of improvised material. First, there is the tension between innovation and the structural expectations of the genre. Sawyer (1992: 258) observes that “although many [jazz musicians] talk of being faithful to the [...] jazz domain, there is also a recognition of the importance of breaking with tradition, of going beyond the domain definition,” where the domain is recognised as “the scales, harmonic patterns, and rules which make up the definition of jazz”. Second, there is a tension between a musician’s “own personally developed patterns or structures, and the need to continually innovate at a personal level, to continue growing musically” (Sawyer 1992: 259). I emphasise Sawyer’s observations regarding tensions in jazz music, because they are apparent in other forms of improvised art. For instance, an actor may wish to push the boundaries of expression in improvised theatre, but must do so within a framework of
intelligibility—so that the audience is not ‘lost’ by the complexity of the plot—and must do so believably—so that the audience does not dismiss the characters and their motivations as implausible. Similarly to Sawyer’s second tension, an actor improvising as a member of a company may wish the plot of the play to follow a certain path that is not agreed upon by the other actors in the company. I return to this tension in my discussion of improvised theatre in §2.2.2.

Studies that examine the cognitive processes of improvisation are particularly useful in formalising improvisation as a strategic process. Mendonça & Wallace (2004: 1383) state that “it may be possible to build cognitively-grounded, flexible models of improvisation” given that, in their own experimental tests on jazz musicians’ creative strategies, they found that “players [...] reasoned similarly regardless of instrument, tune or the group to which a player belonged”. Cognitive models of improvisation appropriately focus on the constraints put on improvisatory artists, and their processes of working within these constraints. Yokochi & Okada (2005) rigorously timed and mapped the creative process of a Chinese ink painter working on a large painting on Fusuma doors in a Buddhist temple. Their study strongly suggests that constraints, especially the audience’s interaction with the artist, dramatically influence the artist’s creative process, and the strategies used by the artist (Yokochi & Okada 2005: 249, table 2):

Based on [the painter’s] knowledge and skills, he draws pictures in a fairly patterned way. However, random lines that [an] audience [draws] can create new constraints for his drawing and force him to create a new style of pictures.

(Yokochi & Okada 2005: 252)

A similarly cognitively-based study of effective decision making in relation to emergency management protocols by Mendonça & Wallace (2007) suggests that improvisational processes, similar to those observed in jazz musicians, operate in emergency management situations. The parallel is drawn insofar as, similarly to artistic improvisers, responders to an emergency are constrained by time, past decisions, and available human and material resources. The cognitive aspects of improvisation also feature heavily in research into computer-based interactive models of emergent narrative; see, for instance, FearNot! (Aylett et al. 2005, 2006). Such research discusses the use of improvisation as a method of increasing efficiency by increasing player agency. Programs are taught to improvise, and dynamically generate plot in response to the player’s intervention in the programmed actions of the non-player characters, that is, characters who exist in the game, but are not directly controlled by the player (see Cavazza, Charles & Mead 2002a,b). Other studies deal with the theory of computer uses of improvisation; I mention these specifically in the context of improvised theatre below.

The most complete formal model of improvisation concerned with the players’ cognition has been formulated by the psychologist Pressing (1988), and specifically refers to musical improvisation. A short summary is given here so that indications can be given for how it has influenced the model of improvised theatre that I devise. In Pressing’s model,
an improvisation is an ordered set of *event clusters*. Event clusters are discrete sections of musical aspects, where musical aspects include: produced and sensed sounds; cognitive representation of the sounds in terms of musical theoretic and expressive dimensions; and timing of muscular actions (Pressing 1988: 154). Pressing’s model proposes that “production of \[an event cluster\] $E_{i+1}$ occurs primarily on the basis of long-term factors […] and by evaluation of the effects and possibilities of $E_i$”. $E_{i+1}$ is formulated from a decomposition of the musical aspects of $E_i$, and the assignment of these aspects into a number of categories, the specifics of which do not interest us here. These are then compared with the history of the elements of these categories, and on the basis of this history, the new event cluster is selected and executed.

Pressing’s formal model of improvisation provides a launchpad for my own venture into formalisation, and some of the concepts he develops are used in my model. For instance, event clusters are the basis of my formulation of games in improvised theatre, where they are reframed as constituents of offer–response pairs, a concept I introduce in §2.2.2.2. In particular, the ability of Pressing’s model to capture the effect of prior events on subsequent decisions is something that I wish to carry over to my own model, and it is here that we find a particularly strong affinity between Pressing’s model and the theory of iterated games.

### 2.2.2 Theatrical improvisation

Improvised theatre is a genre of theatre in which certain elements of the plot are not determined before the execution of the performance. The range of possible kinds of improvised drama includes: plays whose scenes, and general plots, are planned beforehand, but whose dialogue and action within each scene are improvised; scenes in which the performers take some referent (e.g. audience members’ suggestions) and build a plot around them; and completely free improvisation, where the plot, structure, and characters are generated on stage. For instance, performances of theatre sports involve short scenes with stringent restrictions on what is possible (e.g., that each line must begin with the letter of the alphabet consecutive to the one before it), while the general class of longform improvisation tends to have fewer rules and, as the designation implies, a longer duration. The content of improvised performances is also varied. Most audiences are familiar with comic improvisation (especially Anglophone audiences as a result of the television programme *Whose Line is it Anyway?*), though improvised performance may also be serious, or tragic. Since they occur without planning, improvised performances tend to involve less stagecraft (that is, props, sets, costumes) than other forms of theatre, and therefore the content often relies on techniques—chiefly miming—which require a greater degree of the audience’s suspension of disbelief than in non-improvised performances. Improvised performances also tend to occur in places where traditional theatre is impossible; for instance, the performances that constitute the dataset I use for my analysis took place in the back room of a suburban pub.

The framework for analysis of improvised theatre that I use in this thesis is informed by a recent research programme that seeks to use the cognition of actors in improvised the-
Improvised theatre to inform the design of machine agents. The papers that form part of this programme include Cavazza, Charles & Mead (2002a,b), Arinbjarnar & Kudenko (2009), Baumer & Magerko (2009, 2010), Magerko et al. (2009), Fuller & Magerko (2010), Riedl & Young (2010). The findings of this research programme are useful because they aim for mathematical rigour, and their models are based on computer scientific formalisms and metaphors. Like the current project, the machine agent programme is driven by data of actual performances. But the account provided by these papers is silent on the question of language use in improvised theatre. I perceive this as a major flaw in the programme. The primary way that actors execute a performance is by producing utterances. And so it behoves an account that aspires to generality and formal rigour to explore in depth the use of language. Unfortunately, the machine agent programme does not provide such an exploration. While it does provide a good foundation for a better formal understanding of improvised theatre, it is lacking with respect to its understanding of language in improvised theatre.

The machine agent programme views an actor’s major task in improvised performance as the generation of a plot for an audience, without revealing the mechanics of its generation to the audience (Magerko et al. 2009). The programme uses a theory of plot proposed by Chatman (1980) that identifies two constituents of a plot: the story, which is constituted of events and existents, the latter of which are further broken up into characters and environment; and the discourse, which is made up of the manifestation (the form that the plot takes: play, novel, etc.), and the structure of the plot. I use Chatman’s theory as a framework in this section, which reviews scholarship concerning three major aspects of improvised theatre: character formation and assignment, plot generation (which covers the plot’s events, environment, and structure), and communication and organisation between actors.

2.2.2.1. Character formation and assignment

A character is identified by her “traits and consistency” (Baumer & Magerko 2009: 145). According to Baumer & Magerko (2009: 145), traits are those aspects of a character that distinguish her from other characters. The authors identify four such aspects: relationships, goals, history, and attributes. A character’s relationships are the ways in which she views the other characters, present or not in the performance, relative to herself; her goals are those outcomes she considers more beneficial than other possible outcomes; her history contains those events which precede, or occur during, the action of the play, and which therefore affect her preferred outcomes; and her attributes are her physical, mental, and/or emotional qualities. The consistency of a character is “a matter of whether or not the character remains verisimilar within the narrative” (Baumer & Magerko 2009: 145). Consistency is characterised by the character’s traits not altering dramatically as the result of inputs—that is, plot events and actions of other characters—unless these inputs dictate that the traits shift, such as when one character impersonates another to fool a third, for instance—though the act of impersonating is, too, prompted by some original goal of the character. Bruce et al. (2000: 4002) suggest that successful dramas have “plots driven by
characters who make purposeful actions towards their goals”; Riedl & Young (2010: 220) identify that “goal-oriented behaviour is a primary requirement for believability,” on an equal footing with how “the internal attributes of a character such as personality, emotion, desires, and intentions manifest themselves through the [character’s] decisions”.

Actors must ensure that their characters behave as if they are realistic, verisimilar individuals who exist in the world of the scene, which I call the plot-world. A character, if she is to be a credible individual in a possible world, must act as if driven by her own goals, and as if possessing traits that are not necessarily shared by the actor who portrays her. Perhaps most importantly, a plausible character should attempt to solve the problems that she is faced with using a rational decision-making process similar to the rational process that is used by the actor portraying her. However, unlike the character’s process of deliberation, an actor’s is not bound by the narrative universe, as she is able to view the world from a perspective which is ostensibly impossible to her character (hearing conversations between other characters from the wings, for example). Clearly such discrepancies in knowledge are particularly marked when the actors have planned that some event should befall to a character, while the character is unaware of the impending event. A quotation that is particularly instructive with respect to this tension is reported by Magerko et al. (2009: 122). It comes from an interview with an actor about a performance given immediately before the interview.13

So I started thinking like, ’What do I make my wife?’ He wishes I didn’t have my wife and I’m thinking, ‘Should I not like my wife or should I be in love with my wife?’ And that’s going to base how I feel about what he’s saying to me. My opinion about the woman I’m married to. So I could’ve sided with him and said, which I almost said; I almost said, ’I wish I wasn’t married to her either. She’s a real [expletive].’ But I decided to take the other stance and going like, ’Why? What’s wrong with my wife?’ and then thinking like, ‘I’m completely finding sort of a game within the scene. I’m going to be the total outsider from this trio of friends. I’m going to be the guy whose opinion differs from them, no matter what.’

In this excerpt from the interview, the actor directly addresses his decisions regarding the choice of attributes to ascribe to the character. What is important to note here is that the actor wishes to construct a character whose motivations have a plausible basis in his traits, and likewise wishes that those traits be plausible in themselves. To do so, he considers the epistemology of the character, insofar as it is credible in the plot-world.

A plausible character also has motivations in the plot-world. Just as knowledge can differ between a character and the actor who plays her, so too can their motivations differ. This may lead to a paradoxical tension between the motivations of the actor, namely to perform a successful show, and those of the character, the satisfaction of whose goals could

13The censorship of an expletive in square brackets in the quote is the authors’. I do not censor any of my data.
produce an unsuccessful show. This tension has been noted in the literature, if only curso-
& Magerko 2009, 2010). In particular, Sawyer (2003: 68) notes that actors face the task
of simultaneously “enacting characters” and “negotiating the emerging frame”, and there
may be contradictions that arise as a result. This tension is made clearer if we consider it
in a game theoretic light, as I propose below.

(i) A character never selects her own actions: they are instead selected by the actor
playing her;

(ii) the character should, however, be considered game theoretically rational;

(iii) the actor, who is also game theoretically rational, has knowledge of both the nar-
rative and metanarrative elements of the performance, the latter of which are not
available to the character: the actor, then, must rationalise a particular course of
action relative to both the narrative and metanarrative situations;

(iv) since she has information that is more imperfect than the actor’s, the character’s
preferences over the outcome of a plot may differ from the actor’s.

Let us consider a more concrete example, which also shows how this tension might
be better understood using game theory. Imagine that a situation were to arise with a
Western-style showdown between the protagonist Inge (i) and the antagonist Mad Jim
(j), played by the actors Alice (a) and Bob (b), respectively. i and j hold a gun aimed at
the other, and if either shoots, the other will die. There are thus two possible actions:
shoot (x), or abstain (y), and if i selects x or y, j dies or survives, respectively. The same
is true, vice versa, of j’s choice. We assume that neither character wants to die, and that
both characters wish to kill the other. Therefore, the utilities are calculated such that if a
character’s opponent is killed, she gains a point of utility; and if she dies herself, she loses
a point of utility. If we consider only the characters’ utilities, the utilities of the game are
expressed by figure 2.5. No optimal strategy emerges for either player.

But this is not the full picture. The characters themselves do not make decisions, but it
is instead the actors who do. We assume that actors are cooperatively generating the plot,
an assumption that is discussed and justified below. We therefore assume that the actors’
utilities are aligned. Suppose that Mad Jim has done enough throughout the plot to arouse
the hatred of the audience, and that only through his death will the audience be satisfied.

<table>
<thead>
<tr>
<th>(u_i, u_j)</th>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>(0, 0)</td>
<td>(1, -1)</td>
</tr>
<tr>
<td>y</td>
<td>(-1, 1)</td>
<td>(0, 0)</td>
</tr>
</tbody>
</table>

Figure 2.5: Normal form of the shootout game, with characters’ preferences only.
(\(u_a, u_b\)) x y
\[
\begin{array}{cc}
x & (−1, −1) & (1, 1) \\
y & (−1, −1) & (0, 0) \\
\end{array}
\]

Figure 2.6: Normal form of the shootout game, with actors’ preferences only.

Let us furthermore assume that the actors recognise this, and that they prefer that this be the case. Figure 2.6 represents the actors’ utilities for this game. There is a clear optimal strategy here, and that is for \(a\) to shoot, and for \(b\) to abstain.

We can see, then, that the interests of the actor \(b\) and her character \(j\) do not coincide. \(b\)’s optimal strategy is for \(j\) to die, and \(j\)’s optimal strategy is to survive, and to shoot \(i\). Since it appears that actors prefer to create characters that are believable, \(b\) must somehow allow \(j\) to die without it seeming intentional. This can only come about by means of cooperative action: \(a\) must make \(i\) shoot in such a way that \(b\) can act as if \(j\) is unprepared. This, in turn, requires planning and communication. But here we find a further paradox, namely that if the actors communicate explicitly about the plot and its resolution on the stage, this, too, violates the consistency of the characters. In game theoretic terms, the actors must account for, and act according to, two sets of preferences over the possible outcomes, and therefore must anticipate and adjust their strategies accordingly.

This tension is overcome in the theatre studies literature by deferring to the preferences of one of the agents: either the actors or the characters. Halpern, Close & Johnson (1994), Johnstone (1989, 1999) propose that the preferences of the actors should occupy a position of primacy, while Spolin (1963) advocates for the deferral to the preferences of the characters. In the model that I present in this thesis, I opt for the primacy of the actors, for the following reason. This thesis is concerned primarily with language use, rather than faithful description of the metaphysics of theatre. Since it is actors who actually select the utterances that constitute a plot, it is their deliberative process we must attend to, rather than those of the characters’. Incorporating both the characters’ and the actors’ preferences into a game theoretic model is a direction for future research.

2.2.2.2 · Plot generation

Once characters have been established, these characters’ motivations drive the generation of a plot. Let us return to the definition of plot proposed by Chatman (1980) as being constituted of events and existents, where existents include characters and environment. The environment is the setting of a particular narrative, and therefore constrains the possible actions of the characters. A narrative is a series of events. A series of events is partitioned into situations, called states of the scene by Baumer & Magerko (2009: 144). The definition that Baumer & Magerko (2009) provide of states of the scene is that of a state as it usually appears in narratology, namely “the condition of a system (or part thereof) at a given point of operation; a set of elements characterised by a number of properties and re-
lations at a given time” (Prince 2003: 92). So, a state of the scene is identified by a specific configuration of existents and of the environment, and of the attributes that are ascribed to these. It is not possible for any part of a plot to be outside of a state. A change from one state of the scene to another is called a state shift. While a precise mechanism for shifting state is not explicitly stated, Baumer & Magerko (2009) imply that the introduction of new existents, or of new attributes of those existents that are already present, causes the state of the scene to shift. For the moment, this notional definition will suffice. In chapter 4, I provide a formal definition of states and of the mechanisms that cause them to shift.

Actors in improvised theatre, according to Baumer & Magerko (2010: 169), achieve the instantiation of events by the use of offer–response pairs (o–r pairs), defined as "the core narrative move in improvised theatre (as opposed to the sole canonical offer)". Importantly, then, the core narrative move requires action by multiple agents. The study of theatre, improvised or not, has a long history of this view of communication; see Bakhtin (1981), Bate- son (1972), Hymes (1962, 1963, 1976), Silverstein (1976, 1993), Sawyer (1992, 2003), Sawyer & DeZutter (2009), for example. Indeed, anthropologists and literary theorists consistently assert that theatre is in its essence an interactional practice, which necessarily takes place in a space between humans, be they members of the audience (see, for instance, Raz 1983, Schechner 1985, 1988, Bennet 1997) or the actors themselves (see Goffman 1956, Burke 1969, Turner 1982, MacAlloon 1984). It behoves this thesis to be faithful to this view of communication, for it is widely considered to be fundamental to a proper understanding of interaction in improvised theatre.

Offers in o–r pairs range from literal offers (e.g., i asks j if she would like a drink) to i making some assertion to which j responds in the affirmative or the negative. Baumer & Magerko (2010) posit that actors behave as if there are two kinds of offers: active and passive, which are terms taken from Ryan (1991). Active offers shift the state of the scene, while the latter do not. For clarity, I prefer to use the terms state-shifting moves (SSMs), and state-maintaining moves (SMMs). States may be shifted either by the introduction of vital information through the performance of an action which is different enough from those expected in the current state, or by meta-narrative events, such as the conclusion of a scene.

Baumer & Magerko (2010) identify four types of responses, which are measured by two criteria: first, whether the offer is responded to in the affirmative or negative; and second whether the response follows the narrative path projected by i’s offer. If the first of these criteria is met, then the response is affirmative, and otherwise it is negative. If the second criterion is met, then the response is called an augmentation, and if not, then it is called a redirection. The four types are known by the names used for them by improvisers: Yes, And (affirmative, augment), Yes, But (affirmative, redirect), and No, But (negative, redirect), and No, And (negative, augment) (Baumer & Magerko 2010: 169). Baumer & Magerko (2009) claim that there is no evidence of responses that can be categorised as negative augmentations in their data. I agree in cases where the utterance responds to an SMM. SMMs generally rely on the information entailed by the utterance having been
established in a scene, and it does not seem possible to negate pre-existing information and maintain the state of the scene. However, I claim in section §6.2.2 that, if the response follows an SSM, negative augmentations are not only allowable by the formal apparatus that I devise—since a Receiver may disagree with what the Sender intends to convey while preserving the intended effect on the state of the scene—but that they in fact appear in my data. Furthermore, I claim that it is impossible to respond to an SSM with an affirmative redirection. Since SSMs introduce information into the scene, it is impossible to imagine an utterance that would simultaneously accept the new information and maintain the state.

These categories may become clearer with some example o–r pairs. The examples that I use centre around a single offer, which is given in (2).

\begin{equation}
\text{Pozzo is coming over for dinner.}
\end{equation}

Imagine that there are two situations. In the first situation, the premise that Pozzo comes over for dinner has already been established. That is, (2) does not introduce new information into the scene, and so the utterance of (2) as an offer is an SMM. The three utterances in (3) are responses to (2) as an offer is an SMM.

\begin{enumerate}
\item Affirmative augmentation: “Yes, I know.”
\item Affirmative redirection: “Yes, but he’s also bringing his mother.”
\item Negative redirection: “But how? Pozzo is dead!”
\end{enumerate}

(3-a) is an affirmative augmentation because it accepts the assertion of (2), and it does not shift the state, for if Pozzo’s coming to dinner has already been established, then we can assume that the actor producing (3-a) is aware of this fact. (3-b) is an affirmative redirection because it both accepts the premise, and it shifts the state by introducing new information into the scene. (3-c) is a negative redirection because it both rejects the premise and shifts the state.

In the second situation, that Pozzo comes over for dinner has not been established, and so (2) is an SSM. The responses in (4) are typical of the kinds of responses.

\begin{enumerate}
\item Affirmative augmentation: “Yes, I know.”
\item Negative augmentation: “No, he’s coming to lunch.”
\item Negative redirection: “No, he’s not.”
\end{enumerate}

(4-a) is an affirmative augmentation because it accepts both the premise and the shift of state. (4-b) is a negative augmentation because it rejects the premise, but does shift the state. (4-c) is a negative redirection because it rejects the premise while not allowing the state of the scene to shift.

\section*{2.2.2.3 · Communication and organisation between actors}

Plots of improvised dramas are generated by the construction of states, which are made up of series of o–r pairs. Importantly, this positions the utterance as the central element
of improvised theatre, and it is only by the performance and interpretation of utterances that an improvised performance may occur. In chapter 4, I reframe the description of these processes as strategic. I argue that actors purposefully select o–r pairs, and shift or maintain states, based on their preferences over the outcomes.

Aside from the inter-character communication described above by o–r pairs, the actors themselves must navigate the problem of inter-actor communication and organisation while on stage, and while being accountable to an audience. O–r pairs are ostensibly communication between characters; however, they are also necessarily the actors’ only method of communication between themselves while on stage. While state shifting occurs at the level of the narrative (the character), the strategic intention in driving the plot toward certain outcomes occurs at the level of the meta-narrative (the actor). Improvising actors are in a state of constant covert bargaining, cajoling, refusal, and acceptance in their attempts to steer the plot in a direction which they consider successful.

Little research has been conducted regarding this aspect of improvised theatre. Fuller & Magerko (2010: 2) remark that “misunderstandings and miscommunications are common in improv [sic] because coordination between improvisers is not an explicit act (i.e. improvisers do not directly communicate their intentions in a scene outside of what occurs in the performance on stage)”. They frame the successful communicative acts of actors as strategies which promote what they term “cognitive convergence”, which is a “state of agreement of assumptions between two or more people and is necessary for shared mental models to exist” (Fuller & Magerko 2010: 4). Fuller and Magerko also identify what they call ‘repair strategies’, which fit well with linguistic analysis of improvised theatre: they include verification and clarification requests, blind offers where an actor introduces intentionally vague information into the scene so that her fellow actors will incorporate this knowledge into their understanding of the scene, and so on.

I aim to show in the following chapters that any formal, rigorous account of improvised theatre must be a complete and detailed account of the linguistic strategies that are used by actors during an improvised performance.

2.2.2.4 · Harold longform

The data that are used throughout this thesis are recordings of performances of Harold longform, or the Harold. The Harold is a codified form of improvised theatre whose characterisation results chiefly from two aspects (Leep 2008, Halpern, Close & Johnson 1994, Fotis 2014, Hauck 2013).

First, there is a set structure that each performance of a Harold follows. In a Harold, there are two kinds of scenes: proper scenes, and theatre games. A normal Harold is organised into three groups of three proper scenes, with a short theatre game before each of the three groups. Theatre games are not part of the dataset that is used to develop my model. This decision has been made as a result of the lack of a plot in theatre games, which are unrelated to the other scenes in the performance. The proper scenes in group one are denoted A1, B1, C1, those in group two A2, B2, C2, and those in group three A3, B3, C3. As
a result of this structure, a given narrative may be distributed across multiple scenes, or may occur in only one. However, most narratives are resumed at least once, so they are spread across two scenes, since such a practice is encouraged in the Harold (see Halpern, Close & Johnson 1994).

Second, any given Harold begins, and is based on, an audience suggestion. In the data that I have collected, each suggestion was a single world; suggestions used included abacus and terracotta, for instance. This suggestion forms the basis of the ensuing performance, and importantly does so without the actors leaving the stage to consider the idea. In each of the performances that were recorded, the actors uttered a series of words that were chosen by free association, which began with the audience suggestion. These words formed the basis of the themes that were explored throughout the performance. The themes that are generated in this early stage of the performance are important to the actors, and they return to them often in the course of a performance. Indeed, these themes are treated as the core of the performance.

Harold longform is selected for the present project for three reasons. First, since this project is concerned primarily with language use in improvised theatre, more naturalistic styles of speech in improvised theatre are preferred to more artificial styles. The scenes of the Harold are unencumbered by the stringent rules on actor conduct that apply in other genres of improvised theatre. For instance, theatre sports is a form of improvisation that involves short scenes, each of which has restrictions on what actors are allowed to do (such as Johnstone 1989). Harold performances generally do not have such restrictions, and so the language use therein is closer to naturalistic conversation than, for instance, theatre sports. Second, the exploration of themes is fundamental to the practice of performing Harold longform. This means that the actors’ choices involved in a performance are more readily analysed; and that the actors themselves are more likely to establish common ground that concerns a certain topic, and thereby resemble more closely naturally-occurring speech. Third, and perhaps the simplest reason, is that longform performances tend to be of a greater duration, and therefore a greater amount of data was able to be collected. The larger dataset ensures that the ensuing analysis is accountable to a greater number of potential dramatic situations.

2.3 · Conclusion: the rationale for the study

In this chapter, I have laid the theoretical groundwork for the thesis, by introducing the two major components thereof. First, I have introduced game theory, with a specific focus on the extent to which it is used to define formal models of linguistic interaction. Beyond the established formalisms of game theory that are currently used for linguistic models, I have introduced iterated games, which allows for models of conversations with multiple turns. Iterated games have been introduced because they form the basis of the formal model of improvised theatre that will be introduced in chapter 4. I have also reviewed the literature on improvisation, and particularly on improvised theatre, that takes a formal, and rigorously descriptive approach.
In particular, my exposition has focussed on a recent trend in improvised theatre research that attempts to build cognitive models of agency in improvised theatre. While this approach has some success in formally modelling improvisation, it largely ignores the linguistic aspects of the practice. It is the lack of linguistic research into improvised theatre that this thesis wishes to address. In particular, I aim to do this through the use of a formal, game theoretic framework that builds on the formal projects that have been undertaken thus far.

As I have alluded to in this chapter, game theory is able to capture the semantic reasoning of interlocutors in simple circumstances. While not touched on in this chapter, it is also apt at predicting the pragmatic reasoning of interlocutors (see, for instance, Franke 2011). Building on the work of the theoreticians introduced above in my exposition of game theory, it is a secondary goal of this thesis to improve, through the elucidation of a particular genre of communication, the case for the viability of applications of game theory to actual data. I propose that the challenges that are presented when reconciling theory with practice highlight new directions for research that are difficult to imagine without such an application. But further than this, I aim to demonstrate that game theory is a tool for the study of language that can be successfully implemented alongside other theoretical frameworks that are based both in linguistics and outside it.
Chapter 3

A more expressive semantics for games

I believe, said Epistemon, that this is the language of the Antipodes, and such a hard one that the devil himself knows not what to make of it. Then said Pantagruel, Gossip, I know not if the walls do comprehend the meaning of your words, but none of us here doth so much as understand one syllable of them.

François Rabelais, *Gargantua*

We saw in the previous chapter that including utterance semantics in game theoretic models of language normally involves the definition of a semantic denotation function. The semantic denotation function maps each utterance in \( \mu \) onto a proposition; that is, a set of possible worlds in \( \mathcal{W} \). Functions of this kind appear in different guises in numerous game theoretic studies (see, for instance, Jäger & Ebert 2009, Franke 2011, Jäger 2014, Benz & van Rooij 2007). There are some variations on this general method. For instance, the two-step formulation used by Benz (2012a,b) defines a system in which the Sender selects a formula in a first-order language, the formula is mapped onto a set of utterances that can express that formula, and that utterance is mapped onto a proposition such that the utterance is true at each world in the proposition. Despite the variations between these models, however, the implementation of semantics that is used generally in game theory relies on the mapping of utterances onto propositions.

In my model, I employ a different approach to the semantics of linguistic strategies and include Discourse Representation Theory (DRT) (Kamp & Reyle 1993, 2011) as the semantic component of the game model. Using DRT to model the semantics of utterances endows the game theoretic model with a mechanism for a high-fidelity description of the strategies that are available to the actors. For instance, I show in §5.2.1 that actors use sentences that purposefully contravene rules for the proper resolution of anaphoric expressions to imply that the plot-world of the performance is a fully realised possible world. A representation of this strategy is not possible using the semantic denotation method, for it does not permit an expression of the precise words that constitute the utterance. On the other hand, DRT specifies a formal language for representations of utterances as Discourse Representation
Structures (DRSs). DRSs are sensitive to alterations in the exact words that make up an utterance. Alongside this ability, DRT includes a model theoretic, intensional semantics that permits the retaining of the possible world semantics that has been used so far. DRT can therefore produce the same utterance-to-proposition mappings as a semantic denotation function, with the added sensitivity to world-level changes of sentences.

But DRT’s real value to a game theoretic model of interpretation comes in its ability to represent the emergence of a common ground throughout a conversation. DRT allows for the DRSs of multiple utterances to be merged, so that a game model can account for the iterative addition of information into the common ground, and the effects that the common ground exerts on subsequent interpretations. This ability of the theory makes it invaluable to the questions raised by this thesis, which all have a common focus: how actors in improvised theatre produce and interpret utterances.

I should emphasise that one could conceivably use only first-order sentences to represent meaning in the model that I propose; indeed, DRT specifies a means of translating DRSs into well-formed formulae of a first-order language (see Kamp & Reyle 2011). But DRT is preferable for the current project for a number of reasons. (a) The box notation of DRT (discussed below in §3.2.1) provides a highly intuitive means of representing semantic structures, which becomes invaluable when representing complex attitude ascriptions, and concealed meanings within utterances. Furthermore, it provides a unified semantic representation for utterance semantics, the interpretations of utterances, and the contents of the common ground. (b) The explicit concept of a DRS being a partial model of a given possible world (which is also discussed in §3.2.1) provides an intuitive link between the plot-worlds of improvised theatre and the possible worlds of an intensional model. (c) The relation of discourse entities to entities in a given possible world is of central importance to the truth of a given DRS, and therefore to DRT in general. The centrality of this relationship suits improvised theatre well, for it is primarily the plausibility of the characters’ behaviour that determines whether a performance is successful. (d) Since proper DRSs (which are defined below in §3.2.3) have an unambiguous scope, they are able to act as self-contained semantic units. This allows for the implementation of a permissions system (see §7.3.1) to well-defined semantic objects, which is a core mechanism of the model of concealed information that is devised in chapter 7. In short, DRT is not used because it is the only capable semantic framework for the present study, but because it is well suited to the study’s purpose.

In this chapter, I introduce the aspects of DRT that are salient to the game model that is developed throughout this thesis. In §3.1, I introduce some means by which utterance semantics has been included in signalling games in the past. This discussion serves to motivate the inclusion of DRT as a general semantic component of a game theoretic model, but it also gives some indication as to the reason for choosing certain parts of DRT for incorporation. In §3.2, I introduce DRT through examples of its use and efficacy, which is followed by a fully specified, formal definition of the DRT that is used in the thesis.

1I thank an anonymous examiner for pointing this out.
I demonstrate how DRT and game theory are incorporated in §3.4, which forms the basis of the descriptive model of improvised theatre in the following chapter. The chapter concludes in §3.5.

3.1 · Game theoretic approaches to interpretation problems

In this section, I provide brief overviews of some approaches that have been proposed for dealing with problems of interpretation in game theoretic models. This is by no means an exhaustive catalogue, but is meant instead to give the reader an idea of the approaches that already exist. In particular, I discuss three major theories of pragmatic interpretation: Parikh’s games of partial information (Parikh 2006, 2000); Franke’s Iterated Best Response model (Franke 2009, 2011); and Benz’s error model (Benz 2012a,b). I do not intend to utilise or extend any of the models that are presented in this section. Instead, I use this exposition to identify the ways that current approaches to game theoretic linguistics are limited, particularly with respect to their handling of semantics. By drawing attention to these limitations, I motivate the introduction of DRT as a semantic theory for game theory, and also highlight the advantages of its incorporation.

3.1.1 · Games of partial information

Parikh (2000, 1992, 2007, 2006, 1991) proposed one of the earlier applications of game theory to interpretation problems of natural language. His model uses what he calls games of partial information. A game of partial information is simply a signalling game of incomplete information with signals of an expressive natural language. Partial is Parikh’s preferred term because the Receiver, upon observing the Sender’s strategy, has some information about the state of the world. Parikh’s model relies on the inclusion of information regarding the situation rendering certain interpretations more probable than others. For Parikh, the situation is the sum total of any information that surrounds the production of an utterance. It might include the time of day, the place where the utterance is produced, what the producer is wearing, and so on (Parikh 2000). The basic idea of games of partial information is that agents prefer to be understood, and strategies that lead to misunderstandings are generally avoided by interlocutors. Utterances in games of partial information are costly, and as such, a Sender will tend to prefer utterances that require less effort to produce. It is assumed that the more complex an utterance is, the more specific it is, and so ambiguity negatively correlates with cost. If a particular interpretation is more probable in a given situation, then the Sender will produce a less costly but more ambiguous utterance that relies more heavily on the situation to be understood correctly. If instead the interpretation is improbable, the Sender will be forced to produce a more costly utterance to ensure that the interpretation is selected. The solutions to any interpretation problem rely, then, on equilibria that (1) maximise the likelihood that a particular, felicitous interpretation will be selected, and (2) minimise the cost of the utterance that brings about that interpretation.

However, Parikh disagrees with this characterisation; see Parikh (2006: 106–109).
Parikh models semantics in the same fashion that standard signalling games do: by relying on functions to associate certain utterances with situations. The model assumes that the optimal utterances for a given situation are able to be found by identifying Nash equilibria. But it appears instead that the vagueness of the definition of situations means that games of incomplete information can be defined in such a way that the equilibria are simply those solutions that agree with the analyst’s intuitions. For Parikh’s theory to be more useful, a formal definition of situation would need to be provided. This lack of a formal definition of situation means that Parikh’s approach is not tenable in the context of a description of improvised theatre.

3.1.2 · The Iterated Best Response model

A particularly strong model of interpretation, and one that produces excellent predictions about pragmatic behaviour, is the Iterated Best Response (IBR) model, proposed by Franke (2009, 2011). The IBR model relies on a system of types of players, such that each type is defined by the belief held by a player of that type. The types are ordered in levels, and the basic premises of the model are that (a) “a level-(k + 1) player believes that she is playing against a level-k opponent” (Franke 2009: 54); and (b) each successive level nests another recursion of the belief that a player is rational. These two facts conspire to yield the following picture. Consider two players, i and j, whose level is denoted as a subscript integer: \( i_k \) stands for player i if she is of level k. If \( i_0 \), then she holds a belief that is consistent with a player who is utterly irrational, and she believes that her adversary \( j_0 \) has the same belief. Since a level \( k + 1 \) player believes she is facing a level k adversary, \( i_1 \) believes that her adversary is \( j_0 \). And therefore, \( i_1 \) believes that \( j_0 \) believes that \( i_1 \) is in fact \( i_0 \). Likewise, \( i_2 \) believes that she faces \( j_1 \), and so \( i_2 \) believes that she faces a rational adversary. But \( i_2 \) also believes that \( j_1 \) believes that \( i_0 \) is in fact \( i_0 \), and so \( j_1 \) will act as if \( i_0 \) is effectively irrational.

The IBR thus represents the limitations of players’ guesses regarding the cognitive capacity of their adversaries. Such limitations can be harnessed to explain “natural shortcomings of pragmatic reasoning and the development of pragmatic reasoning competence” (Franke 2009: 58). For instance, let \( S_k \) represent a Sender of level k, but also the set of pure strategies that will be played by a Sender of that level, and let the same hold for \( R_k \). The base upon which the IBR model is built are literal strategies. \( S_0 \) is a set of all truthful strategies for a given world state. \( R_0 \) is a set of literal semantic interpretations, which result from updating her prior belief with the semantic meaning of \( S \)’s message. According to Franke (2009: 55), “a level-zero Receiver considers the semantic meaning of an observed message, but does not take his opponent’s strategy into account” (my emphasis). This final point is important, for the IBR model’s aim is to properly capture each player’s expectation of her opponent’s cognitive capacity, which will lead her to anticipate which strategy her opponent will select. And so, the IBR model predicts that \( S_1 \) will act as if the Receiver, who she believes to be of level zero, will interpret her utterance literally, and she will design her utterance so that its literal, semantic interpretation is conducive to her preferences.
Likewise, the IBR model predicts that \( R_1 \) will anticipate that the Sender will send a truthful message, and will update her priors with this in mind. Each player thus attempts to select linguistic strategies that conform to the mental capacity of her opponent.

Players can update what level they believe their adversaries to be. Suppose that \( S_{k+1} \) believes that \( R_k \). Suppose that \( R_k \) produces an utterance that suggests instead that \( R_{k+1} \). \( S_{k+1} \) updates her belief so that she becomes \( S_{k+2} \). This process occurs iteratively. As the level of the types becomes higher, the apparatus necessarily becomes more complex, but this basic logic remains intact. \( S_{k+1} \)'s belief is such that she plays \( R_k \), and so her best strategies are simply those that best respond to \( R_k \). \( R_{k+1} \) will adopt any posterior belief that is consistent with \( S_k \).

This overview represents the foundation of the most basic form of Franke’s model. It also captures something of the elegance of the model, which is able to remain very simple in its basic mechanisms while building an apparatus that can predict deficiencies in the agents’ abilities to correctly judge the cognitive capacity of their opponents, which leads to a robust model that can predict both successful and erroneous pragmatic reasoning. While my model is decidedly less elegant and less generalised, I hope to emulate to some extent the capacities of Franke’s model through the use of heuristics, and I particularly hope to adhere to its principle of avoiding the proliferation of formal devices as much as is practicable.

Franke’s semantic model uses a semantic denotation function model of utterances to propositions. As I have discussed, this is inadequate for the purposes of this thesis.

### 3.1.3 Interpretation under error

Lastly for this section, I present Benz’s game theoretic error model (Benz 2012a,b). Benz’s papers aim to account for interlocutors’ pragmatic errors by incorporating the information theoretic study of communication, and particularly the study of noisy channels. Benz (2012a: 110) justifies the study of errors (the author’s emphasis):

> Speakers commit errors. These errors are expected by both Sender and Receiver, and this expectation is part of the common ground. Error models are an attempt to model the effects of such commonly expected errors in pragmatics. Errors may originate from multiple sources; hence, the representation of errors in error models is as general as possible.

In Benz’s model, the signalling game is altered somewhat so that \( R \) has two choices. Either she performs an action as per usual in signalling games, or, if she is unsure about the state of the world after observing \( S \)'s utterance, she may perform a special action, which involves sending a clarification request. This triggers a second iteration of the game. In the second iteration of the game, no state of the world is set, and no errors may occur, since \( R \)'s clarification request gives \( S \) enough information to know in what way \( S \)'s utterance was deficient. \( S \) produces an utterance that is unambiguous, \( R \) performs an action, and the game terminates. The signalling game that is proposed is augmented by its marriage with
an information theoretic model of communication. Briefly, the model takes the following form, according to Shannon & Weaver (1998):

1. A stochastic process, such as an experiment, outputs a certain result; the result is called the message.

2. A Sender observes this information, and encodes it as a signal. A signal must be appropriate to the channel that the information will be transmitted across. For instance, if the channel is a copper wire, then the signal must be an electrical signal that can travel across that wire. Note that the message and the signal are by definition different; and the signal is supposed to represent the message in some meaningful way.

3. The channel is treated as a function that accepts a signal and outputs a received signal. Noise is an inevitable and irreducible property of any given channel. It introduces completely random imperfections into the signal. For instance, with noiseless channels, the derivation of the received signal is completely predictable for any given signal. As a channel’s noise increases, the unreliability of the derivation also increases.

4. A Receiver decodes the received signal. Since channels are in practice never noiseless (Cover & Thomas 2012), the output of the Receiver is strictly speaking evidence of the message. It follows that as a channel’s noise increases, the reliability of the evidence for a particular message diminishes.

According to Benz (2012a,b), the information theoretic model of communication may be incorporated with the general game theoretic model by altering the granularity of a signalling game so that the Sender makes two decisions. First, having observed the world, she selects a formula from a set of formulae from a second-order logical language. Note that it does not appear that Benz is claiming that formulae are directly used by humans as part of cognition; instead, the formula is a device of the model that represents the Sender’s intended message. Second, she selects a natural language sentence that she sends to the Receiver, which is the analogue of the signal in the information theoretic model. Noise sets potentially introduce error into either or both of these selections; Benz stipulates that noise is not a property of the channel, but is an element instead of players’ cognition. Observe that, since noise is inevitably present in a channel, the Sender and Receiver respectively encode and decode the signal to account for this. That is, they alter the signal so that it is resilient against the inevitable errors introduced by the channel. The Sender and the Receiver are likewise aware that errors may be committed, for the possibility of errors and their likelihood is, according to Benz, part of the common ground of any conversation. As such, the optimal strategy for a Sender is to anticipate that the Receiver will observe

---

Information theory is one of the theoretical pillars of this project, and as such I give a fuller account of noise and information theoretic communication in §7.1.
her strategies against the canvas of this knowledge, and the Receiver will in turn expect that the Sender has designed her utterances with this in mind.

Benz’s model therefore shares a conceptual similarity with Franke’s. The model that I propose is also similar to Benz’s, in that it shares the aim of expressing each player’s belief of the cognitive abilities of her opponent, but also in that it incorporates other theoretical frameworks—most notably Discourse Representation Theory and information theory—to account for interlocutor behaviour. But perhaps the most salient similarity between my model and Benz’s is that Benz’s model uses two functions to represent semantic content of an utterance. I use a similar strategy, but one that applies to DRSs directly.

3.2 · Discourse Representation Theory

While signalling games are able to represent common ground and its effects on the behaviours of interlocutors, the representation is decidedly simplistic. The common ground is often provided as informal background description. We saw the problems with this approach earlier in the discussion of games of partial information: while an informally-described situation may have the ability to predict agents’ behaviour to some extent, the prediction is often a blunt tool. Nuanced effects of a common ground that is shared by Senders are more difficult to handle with such a model. To this end, I propose that the expressive power of game theoretic linguistics may be enhanced by incorporating a more fully-realised model of common ground. There are a number of formal theories which might well fulfil this role, which include most prominently the so-called update logics (see, for instance, Groenendijk & Stokhof 1991 and Veltman 1996). While these logics excel in representing the update of information over time, they are not necessarily as apt at representing a common ground that accrues over time. Furthermore, these update logics suffer from the same drawback as the semantic denotation functions that have been discussed in this chapter and the last, in that they are not particularly sensitive to the precise constitution of utterances.

A formal approach that combines the mathematical rigour of update logic with a high fidelity mechanism for the description of natural language sentences is Discourse Representation Theory. DRT, first developed by Kamp (1981), aims at thorough, expressive models of common ground and of the means by which antecedent information introduced into the common ground can affect interpretations of subsequent utterances. It marries a very intuitive representation of common ground with a rigorous and comprehensive model theoretic apparatus, which, if required, also includes an intensional semantics. DRT has shown its aptitude in formal accounts of a number of linguistic phenomena, including for instance presupposition (van der Sandt 1992, Geurts 1999, Kamp 2001, Kamp & Roßdeutscher 1994), and attitude reports (Asher 1986, 1987, Maier 2009). The extensible approach that DRT takes has meant that a number of varieties of DRT have been developed, each of which emphasise different facets of linguistic interaction. Segmented DRT (Asher & Lascarides 2003, Lascarides & Asher 2008), for instance, is designed to handle the rhetoric relationships between larger scale discourses.
DRT’s aptitude in describing the cumulative effects of information as common ground is established is the major factor that renders it appropriate for my purposes. But its advantages exceed this ability. The model theoretic semantics of the DRT permits it to be incorporated into the possible worlds framework utilised by game theoretic models of language use. DRT utilises an intuitive method of visualisation of common ground, which allows for far simpler descriptions of common ground in narratives than is possible using a theory that lacks such provisions. But perhaps the most profoundly important property of DRT for this thesis which is chiefly concerned with problems of interpretation is that it provides a rigorous, detailed model of what precisely constitutes an interpretation. The following quote from Kamp & Reyle (2011: 326) is particularly instructive (my emphasis):

According to DRT, interpretation of an assertion one hears or reads takes the form of constructing a DRS for it. One way to think of this DRS is as a structure which the interpreter forms in his mind and which for him identifies the content of the interpreted statement.

We have seen above that in, for instance, Franke’s model, the Receiver selects an interpretation, though the precise form of the interpretation is never rigorously defined. Indeed, at the level of the model, the form of an interpretation does not matter, but it is merely important that the correct interpretation be selected for the state of the world. I show in later chapters, that the efficacy of the model can be greatly improved by the inclusion of a complete model of interpretation.

In this section, I introduce the aspects of DRT that are salient to the project at hand. I do so in a very concise (or perhaps even terse) manner; therefore, readers who are interested in further detail are encouraged to consult the sources that I used when writing this introduction: Kamp & Reyle (1993), Geurts, Beaver & Maier (2016), Asher (1986), Lascarides & Asher (2008), Kamp, van Genabith & Reyle (2011), Kamp & Reyle (2011). In §3.2.1, I informally introduce the basics of DRT, which includes the elementary form of DRSs, and three common conditions that may be applied to them. I extend these basic formalisms in §3.2.2, where I discuss the DRT implementation of propositional attitudes. In §3.2.3 and §3.2.4, I provide the formal definition of a DRS language and its syntax and semantics, respectively. I conclude the section in §3.2.5 by discussing those parts of the normal DRT apparatus that are omitted from my exposition.

3.2.1 · Basic discourse representation structures

DRT aims at a formal description of discourses, where a discourse is understood as a collection of sentences, as in a conversation. The primary problem that Kamp (1981) was concerned with was the resolution of anaphoric expressions, especially those cases in which the resolution of anaphora requires information from multiple sentences. To illustrate such problems, let us consider the following, extremely simple discourse, which consists of two sentences.

(v) Vladimir eats a banana. He enjoys it.
It is naturally clear to any competent speaker of English that the pronouns *he* and *it* refer to the nouns *Vladimir* and *banana* respectively. However, standard Montagovian-style semantics operate at the sentence level, and so the two pronouns are unbound to any referent, and the formal interpretation of the problem is undefined.

DRT solves this problem by introducing a persistent discourse that is formed iteratively throughout a conversation. To do so, DRT stipulates a number of construction algorithms which convert the syntactic structures of sentences into DRSs.\(^4\) A DRS is a pair containing a set of discourse referents and a set of conditions (Kamp & Reyle 2011). The pair is most often represented in a box notation, such that a rectangle is partitioned into two halves, where the top half is the set of discourse referents, and the bottom half is the set of conditions. For the first few examples of the following discussion, I present both the mathematical notation proper and the box notation, but for the rest of the thesis I eschew the mathematical notation, for the box notation is generally clearer and far more easily read.

Consider the first sentence of (5). The set of discourse referents includes an individual named Vladimir, and a banana. The constraints stipulate that the former eats the latter, and so the constraint in this case is a two-place predicate relation. The DRS of the first sentence, then, is the pair in (6), represented in box notation in (7).

\[ K_1 = \langle \{x, y\}, \{\text{Vladimir}(x), \text{banana}(y), \text{eats}(x, y)\} \rangle \]

\[
\begin{array}{c}
  x, y \\
  \text{Vladimir}(x) \\
  \text{banana}(y) \\
  \text{eats}(x, y)
\end{array}
\]

The first set contains two discourse referents, \(x, y\), and the second set contains the conditions that describe those referents.\(^5\) Even though it is not strictly a part of the formal apparatus, it is useful when reading DRSs to insert a tacit ‘∃’ before each discourse referent, and a tacit ‘such that’ before the set of conditions. Furthermore, since the DRS is true iff every condition holds for every discourse referent in \(w\), a tacit ‘∧’ may be read between each of the conditions. For \(K_1\), the reading generated using this strategy is as follows.

\[ \exists x \exists y \text{ such that } \text{Vladimir}(x) \land \text{banana}(y) \land \text{eats}(x, y). \]

This style of reading this renders the meaning of a DRS clearer, and it becomes especially useful in relation to complex conditions.

---

\(^4\)More correctly, the conversion algorithm takes syntactic trees that represent the structure of sentences. But the precise process of conversion is not relevant to the project of this thesis. The algorithm is detailed at length in Kamp & Reyle (1993).

\(^5\)An objection might be raised by the keen-eyed logician that *Vladimir* is not quite a predicate in the same way that *banana* is, or especially that *eats* is. Names of individuals, and the assignment of these names, are treated as predicates in DRT by a special interpretation function that operates as part of the model theoretic formal apparatus. I discuss names in §3.2.3.
Now, let us add the second sentence. The second sentence has a DRS of its own.

\[(9) \quad K_2 = \langle \{z, a\}, \{\text{enjoys}(z, a)\} \rangle \]

\[(10) \quad K_2 = \begin{array}{|c|c|}
\hline
z, a & \\
\hline
\text{enjoys}(z, a) & \\
\hline
\end{array} \]

Note that \(z\) and \(a\) refer to entities that are only defined in a predicate relation to one another, and do not have any condition that specifies that they are entities outside of this relation. This means that these referents are not bound (Kamp, van Genabith & Reyle 2011: 130). This DRS, were it encountered without an existing common ground, would be improper (Kamp, van Genabith & Reyle 2011: 146), for \(z\) and \(a\) lack conditions that allow them to pick out an individual in the possible world. But, as we have seen, it occurs following the previous sentence, and so the pronouns may most probably bind with entities referred to in the previous sentence. A DRS that contains the information conveyed by both sentences can be formed by merging (6) and (9). The merging operator of DRT is a core part of the formal apparatus. It takes two DRSs \(K, K'\) to produce a third \(K''\), such that the set of discourse referents of \(K''\) is the union of those of \(K\) and \(K'\); and likewise for the set of conditions. But simply merging these two DRSs does not reflect the reference of the pronouns to the nouns in the previous sentence. To this end, the merge option equates the referents from \(K_2\) and \(K_1\) if they both refer to a single entity in the universe of the discourse (see Kamp, van Genabith & Reyle 2011: 143–4). It is important to note that an equation of referents becomes one of the conditions of the DRS. The representation of the discourse that follows the second sentence becomes \(K_3\).

\[(11) \quad K_3 = \langle \{x, y, z, a\}, \{x = z, y = a, \text{Vladimir}(x), \text{banana}(y), \text{eats}(x, y), \text{likes}(z, a)\} \rangle \]

\[(12) \quad K_3 = \begin{array}{|c|c|c|}
\hline
x, y, z, a & \\
\hline
x = z & y = a & \\
\hline
\text{Vladimir}(x) & \\
\hline
\text{banana}(y) & \\
\hline
\text{eats}(x, y) & \\
\hline
\text{likes}(z, a) & \\
\hline
\end{array} \]

If a DRS is true in \(w\), then there is a set of entities in \(w\) that have the same conditions as the entities in the DRS. Thus, any given DRS may be seen as a partial model of \(w\). If a DRS is true at \(w\), then it follows that it may be incorporated into, or in DRT parlance embedded within, a hypothetical, fully specified DRS that describes \(w\) in its entirety. To put it more concretely, if there is no individual named Vladimir who eats a banana and likes it in \(w\), then (11) is not true in \(w\). On the other hand, if it is true in \(w\), it represents a small subset of the world’s complete sets of entities, and of the conditions upon these entities. It is thus a partial model of \(w\).
DRT also allows for DRSs to be embedded within the conditions of a DRS. This provides for the application of conditions to DRSs themselves. There are three so-called complex conditions: ¬ (negation); ⇒ (implication); and quantification. Negation applies in (13).

(13) Vladimir doesn’t have any bananas.

The DRS for this example is given in (14), and the box notation follows in (15).

(14) \[ K_4 = (\{x\}, \{Vladimir(x), \neg(\{y\}, \{\text{banana}(y), \text{owns}(x,y)\})\}) \].

(15) \[
\begin{array}{|c|}
\hline
x \\
\hline
Vladimir(x) \\
\hline
y \\
\hline
\neg \text{banana}(y) \\
\text{owns}(x,y) \\
\hline
\end{array}
\]

Instead of the entire DRS being negated, the scope of the negation needs to be taken in the context of what was said above about the truth of a DRS in a model at \( w \). (14) stipulates that while there is an individual \( x \) named Vladimir, there is no \( y \) such that \( y \) is a banana, and such that \( x \) owns \( y \). This is clearer if we utilise the reading strategies that I suggested were useful above. The correct interpretation of (14) is given in (16), where square brackets delimit the DRSs, and where italicised text represents the tacit components that I described above.

(16) \[
\text{[There exists an } x \text{ such that } x \text{ is (named) Vladimir; and it is not the case that [there exists a } y \text{ such that } y \text{ is a banana, and } x \text{ owns } y.]}\]

\( K_4 \) provides a good example of accessibility, a core concept of DRT, on which DRT’s ability to describe anaphora rests. Accessibility describes a property of a DRS that permits a DRS’s discourse referents and conditions to access those of another DRS. Access is defined inductively below, but informally it is most often the case that if a DRS \( K \) is embedded within another \( K' \), the latter is accessible to the former, and not vice versa (Geurts, Beaver & Maier 2016: see section 4.2 of). In \( K_4 \), for instance, the negated DRS is able to access the DRS that it is embedded in, while the reverse is not true. It is for this reason that the set of discourse referents of the embedded DRS does not contain the referent \( x \), because the embedded DRS can access the outer DRS’s set of discourse referents. The rationale for restricting the opposite direction may be unclear, and so to illustrate why this is the case, consider (17).

(17) Vladimir doesn’t have a banana. It is overripe.

The pronoun \( it \) in the second sentence is unable to bind to any noun in the discourse, since the existence of a banana has been negated. The DRS of these sentences is given in (18).
(Note that, for this DRS and those that follow, I cease to provide a standard predicate logic representation of the DRS, and I only provide the box notation.)

\[
(18) \quad K_5 = \begin{array}{|c|}
\hline
x, y \\
Vladimir(x) \\
\hline
\end{array} \\
\begin{array}{|c|}
\hline
z \\
\neg \text{banana}(z) \\
\text{owns}(x, z) \\
\text{overripe}(y) \\
\hline
\end{array}
\]

The DRSs depicts the problem: there is no referent in the set of discourse referents of the outer DRS that \( z \) could bind to, and the inner DRS is not accessible to the outer DRS. Accessibility thus predicts violations of our grammaticality judgment of the pronoun \( it \). I return to define accessibility below.

Implication is represented similarly to negation, though it is a binary operator, and so takes two DRSs as its arguments. Consider (19).

\[
(19) \quad \text{If Vladimir has a banana, then he eats it.}
\]

The first thing to notice here is that there are syntactically two sentences that constitute (19).

\[
(20) \quad \begin{align*}
\text{a. Vladimir has a banana.} \\
\text{b. He eats it.}
\end{align*}
\]

These are linked by a condition of implication as expressed by \( if \). Since these are distinct sentences, each requires its own set of discourse referents, as in \( K_6 \).

\[
(21) \quad K_6 = \begin{array}{|c|c|}
\hline
x, y & z, a \\
\hline
\text{Vladimir}(x) & z = x \\
\text{banana}(y) & a = y \\
\text{owns}(x, y) & \text{eats}(z, a) \\
\hline
\end{array}
\]

Like negation, implication is treated as a condition of the partial model of \( w \) that the outermost DRS represents. As such, if the implication does not hold in a possible world, then \( K_6 \) is untrue with respect to that possible world. The equation of \( z \) and \( x \), and of \( a \) and \( y \), occurs as a result of the definition of accessibility given below in §3.2.3, which stipulates that, for the implication relation, the antecedent DRS is accessible to the consequent DRS. This property of the accessibility relation is often used to strip away extraneous detail in the equation of variables. As such, this two-sentence discourse reduces to \( K_6' \).
Quantification is a condition that affects two DRSs, and is similar in many ways to implication. Consider the following sentence with universal quantification.

(23) Every good boy deserves fruit.

The DRS of this sentence with universal quantification is as follows.

\[
\begin{array}{c}
 x, y \\
 \text{boy}(x) \\
 \text{good}(x) \\
 \text{fruit}(y) \\
 \forall x \quad \text{deserves}(x, y)
\end{array}
\]

Such conditions are called duplex conditions. The left box restricts the duplex condition to a set of entities with certain conditions, the centre diamond establishes the scope of the quantification, and the right box states the predication that is made of the referents that are targeted by the duplex condition. It is a general fact about duplex conditions that, if \( K'(Q)K'' \), where \( Q \) is a quantifier, then \( K' \) is accessible to \( K'' \). Hence, the set of discourse referents of the right DRS is empty, and the entities in the conditions utilise those in the left DRS’s set of discourse referents. The DRT formulation of quantification is not particularly useful in cases of universal quantification, since its treatment is effectively identical to implication. (23) can alternatively, and equally well, be represented as (25).

\[
\begin{array}{c}
 x \\
 \text{boy}(x) \\
 \text{good}(x) \\
 y \\
 \text{fruit}(y) \\
 \text{deserves}(x, y)
\end{array}
\]

I do not use quantification in my analysis, for implication is without exception more appropriate for describing the data. I have introduced duplex conditions here, however, because they are referenced in the formal definitions of §3.2.3 and §3.2.4.

The examples that I have presented in this subsection illustrate the range of sentences that DRT provides for. In the following subsections, I introduce the formal, model theoretic apparatus of DRT.

---

6This sentence doubles as a mnemonic for the English names of the notes on the lines of a treble clef!
3.2.2 · Propositional attitudes

I noted above in §2.2.2.1 that actors must be mindful of their characters’ motives when devising a plot. Recall that, according to Riedl & Young (2010: 220), “the internal attributes of a character such as personality, emotion, desires, and intentions manifest themselves through the [character’s] decisions”. It is therefore useful to be able to represent the mental states of characters, and particularly to be able to do so recursively, so that the beliefs of characters regarding the mental states of other characters may be represented. Fortunately, extended forms of DRT include sophisticated provisions for propositional attitudes. As the name implies, propositional attitudes express an agent’s attitude towards a particular proposition. The most common propositional attitudes express an agent’s beliefs, desires, and intentions. The DR theoretic implementation of attitudes is fairly complicated, and so I keep the justification for any element to a minimum. Kamp & Reyle (2011) give a comprehensive introduction.

Any given attitude is represented as a pair, such that the first element is a description of the attitudinal mode, and the second is the propositional content. A set of attitudes that are all held by the same agent is called an attitude complex. We utilise three attitudinal modes in this thesis: belief, desire, and intention. Each of these attitudes takes as its complement a proposition, which is the propositional content represented as a valid DRS. Any given attitude is held by an individual, and as such we need a method of attitude ascription; I introduce a special reflexive indexical discourse referent $i$ that *ipso facto* refers to the agent that holds the attitude in question.

Let us illustrate how these attitudes operate by considering a situation where Vladimir sees a banana that he believes is real.

\[(26) \langle \text{bel}, x \mid \text{real\_banana}(x) \rangle \]

The first element is an indicator of the attitude that is represented, and the second is a DRS that represents the proposition that is the complement of the attitude. Believing that the banana is real, Vladimir desires to consume the banana, an attitude represented in (27), and he furthermore intends to do so, which is represented in (28). These two attitudes are ascribed to Vladimir, and so use the special indexical discourse referent.

\[(27) \langle \text{des}, i, x \mid \text{eat}(i, x) \rangle \]

\[(28) \langle \text{int}, i, x \mid \text{eat}(i, x) \rangle \]

\[Note that this notation is provisional, for it is underspecified. Below, I demonstrate how attitudes are ascribed to particular entities.\]
These three attitudes all share a discourse referent, namely $x$, the banana. This is important because Vladimir’s desire and intention to eat the banana arise from his belief that the object he sees is a real banana, and not a plastic object that looks like a banana. This sort of interplay of attitudes is common enough that we need to be able to allow for attitudes to be able to share referents. To this end, DRT defines two kinds of anchor: internal and external.

Internal anchors are mental objects of an agent. These objects do not necessarily have to exist in the world outside of the mental states. Since they are part of an agent’s mental state, internal anchors are closely related to propositional attitudes, and should indeed be considered “separate components of the attitudinal state as a whole, on a par with those constituents that are genuine propositional attitudes” (Kamp, van Genabith & Reyle 2011: 333). Formally, an internal anchor specifies two things. First, it assigns an index to the mental object, which is used by other attitudes regarding that object. Second, it specifies the predications that are made of the mental object. Internal anchors are represented similarly to the three attitudes we have discussed above, with the major difference being that the first element of the pair uses the anch mode, and it specifies the index that is used for the mental object in the other propositional attitudes. An example, which represents the banana mental object perceived by Vladimir, is given in (29).

$$\langle \text{anch, } x, \begin{array}{c} x \\ \text{real_banana}(x) \end{array} \rangle$$

I show how internal anchors interact with proper propositional attitudes below.

External anchors are far simpler: they are objects in the actual world that are referenced in propositional attitudes, rather than mental objects. External anchors are formally represented by pairs, such that the first element of the pair is an internal anchor’s discourse referent, and the second element is a discourse referent that is not present in the internal representation, but is part of the universe of the discourse.

Now, suppose that we are interested in an attitude that is ascribed by one agent to another. Ascription of an attitude is handled by a special predicate, $\text{att}$, which takes three arguments: the individual to whom the attitude is ascribed; the attitude complex that is attributed to the individual; and a set of external anchors for the attitude complex. For instance, ascribing the belief that the banana is real to Vladimir takes the form of (30).
The first element of att is the index for Vladimir. The second is the attitude complex, which contains an internal anchor that describes the mental object of the banana, and a belief that pertains to this mental object. The third element of att is the external anchor, which links the mental object, which is a real banana, to the object that actually exists, which is a plastic banana.

With this mechanism we can describe highly complex situations of recursive belief. Let us use two such situations as examples.

First, imagine that Vladimir sees the banana and believes that it is real, but a third party, Estragon, sees the same banana, and believes it to be plastic. The fact that the two agents observe the same banana is expressed by both external anchors pointing to the same entity, but by their very different beliefs regarding the object. These two attitude ascriptions are given in (31).

There are two attitude complexes in this DRS. The top complex describes the mental state of Vladimir, who believes that the banana is an actual banana, and the bottom complex
Chapter 3 · A more expressive semantics for games

describes Estragon’s belief that it is a false banana. Note that each attribute complex has its own internal anchor, which defines a mental object of each agent that is the object of that agent’s attitude. Observe, however, that the external anchors both point to \( x \), which is the plastic banana that actually exists. The actual existence of the banana is asserted by the predicate `plastic_banana(x)`. The inclusion of these attitude complexes in the DRS in (31) means that it is able to represent the complex situation in which Vladimir and Estragon have different beliefs about a single object.

Since belief attributions utilise DRSs to express the object of the attitude, it is easy to represent recursive attributes. For instance, imagine that \( e \) believes that \( v \) believes the banana to be plastic, and so does not ascribe the desire to eat the banana to \( v \). This is represented in (32).

\[(32)\]

The outer attitude complex describes Estragon’s belief about Vladimir’s belief. Vladimir’s perceived belief is represented by the inner attitude complex. Once again, these attitude complexes have different internal anchors and equivalent external anchors. The external anchor refers to actual objects in the DRS’s entities, even if the attitude ascription itself is recursively embedded.

Attitude complexes behave precisely as any DRS condition, and as such no special allowance needs to be made to include attitude attributions in a DRS language. That is, an attribute complex is true with respect to a possible world \( w \) iff there is an individual in \( w \) who holds the attribute ascribed to her, in addition to meeting the conditions imposed on her by the DRS. In the following two sections, I introduce the formal definitions of a DRS language, and hence also the formal definitions of the attitudinal modes that have been described above.

### 3.2.3 · Syntax

In this section, I introduce the formal syntax of a DRS language. To begin, let us define a DRS language (based on Kamp, van Genabith & Reyle 2011: 145).

**Definition 3.1**: DRS language. A DRS language \( \mathcal{L} \) is defined by a vocabulary, which includes:
Discourse Representation Theory

(a) a set $\text{Ref}$ of discourse referents, which includes the reflexive index $i$;
(b) a set $\text{Name}$ of one-place definite relation constants;
(c) a set $\Phi^n$ of $n$-ary predicate constants, which includes $\text{att}$;
(d) a set $\text{Sym} = \{=, \neg, \lor, \Rightarrow\}$ of logical symbols;
(e) a set $\text{MI} = \{\text{bel}, \text{des}, \text{int}\}$ of mode indicators;

Note that this definition deals with the assignment of names, and stipulates that they are unary predicates, which rectifies a potential problem that I raised earlier in the chapter: it is problematic to treat names as predicates in the same way that attributes of an entity are treated. It is also important to bear this fact in mind later, when I discuss how a character’s name becomes the subject of errors. A DRS language $\mathcal{L}$ has syntactic rules, which are given in definition 3.2 (based on Kamp, van Genabith & Reyle 2011: 145–146).

**Definition 3.2:** Syntax of $\mathcal{L}$. Let $\mathcal{L}$ be a DRS language. The following syntactic constraints apply:

(i) if $D \subseteq \text{Ref}$, and $C$ is a set of conditions (which are defined inductively below), then $\langle D, C \rangle$ is a DRS;

(ii) if $x, y \in \text{Ref}$, then $x = y$ is a valid condition;

(iii) if $a \in \text{Name}$, and if $y \in \text{Ref}$, then $a(y)$ is a valid condition;

(iv) if $P$ is an $n$-placed predicate in $\Phi$, and if $x_1, \ldots, x_n \in \text{Ref}$, then $P(x_1, \ldots, x_n)$ is a valid condition;

(v) if $K$ is a DRS, then $\neg K$ is a valid condition;

(vi) if $K$ and $K'$ are DRSs, then $K \lor K'$ is a valid condition;

(vii) if $K$ and $K'$ are DRSs, then $K \Rightarrow K'$ is a valid condition;

(viii) if $K$ and $K'$ are DRSs, and $x \in \text{Ref}$, then $K(\forall x)K'$ is a valid condition; and

(ix) if $x \in \text{Ref}$, $A$ is an attitude description set (ADS; defined below), and $\Xi_{\text{ext}}$ is an external anchor for $A$, then $\text{att}(x, A, \Xi_{\text{ext}})$ is a valid condition.

Conditions (ii), (iii), and (iv) are atomic conditions, and (v), (vi), (vii), (viii), and (ix) are complex conditions. Atomic conditions apply to entities in the universe, while complex conditions apply to DRSs, which have atomic conditions of their own. Note that (ix) is a complex condition because the second element of a pair in an ADS is a DRS.

The attribute description set in this definition is defined as follows (see Kamp, van Genabith & Reyle 2011: 343).

**Definition 3.3:** Attribute description set. An attribute description set (ADS) $A$ of $\mathcal{L}$ is a set of pairs, such that each pair has one of the following forms:

(i) $\langle M, K \rangle$, where $M \in \text{MI} = \{\text{bel}, \text{des}, \text{int}\}$ and where $K$ is a valid DRS of $\mathcal{L}$; or
(2) ⟨{anch, x}, K⟩, where x ∈ Ref and x ∈ D_K, and K is a valid DRS of L.

The second of these formally defines internal anchors. From this definition we can derive a formal definition of external anchors (see Kamp, van Genabith & Reyle 2011: 343).

**Definition 3.4:** External anchor. Let K and K’ be valid DRSs, let Ξ be an ADS, and let \( \Xi_{\text{int}}(A) = \{ x \mid \exists K : \langle \{\text{anch}, x\}, K' \rangle \in K \} \) be a set containing each internal anchor of A. An external anchor of A is a function \( \Xi_{\text{ext}} \) such that \( \text{Dom}(\Xi_{\text{ext}}) \subseteq \Xi_{\text{int}}(A) \).

Finally, the basic structure of DRSs having been taken care of, let us define how DRSs may be merged (see Kamp, van Genabith & Reyle 2011: 140).

**Definition 3.5:** DRS merge. Let K and K’ be valid DRSs. \( \oplus \) is a merge operator such that \( K'' = K \oplus K' = \langle D_K \cup D_{K'}, C_K \cup C_{K'} \rangle \). K’’ is a valid DRS.

The merge operator is central to the model of the accumulation of common ground information that I propose in later chapters.

The definitions above completely specify the syntax of L. The following definitions of this subsection are not necessary for the full specification of a DRS language, but they are useful in granting us a vocabulary for the more complete description of DRSs.

DRSs are sensitive to failures of binding. A failure of binding occurs when an entity is introduced into the universe of a discourse, and no conditions are attached to that entity. We saw with example (17) that such a failure of binding may occur as a result of accessibility, and we saw that this led to a felicitous description of a failure in anaphora resolution. As such, it is useful to endow DRT with a taxonomy of bound and free discourse referents.

The definition that is often given of free discourse referents, as in for instance Kamp, van Genabith & Reyle (2011: 146), unfortunately suffers from a basic ambiguity. It is specified for both DRSs and for conditions, without acknowledgement that this is the case. For this reason, I have chosen to define two functions for free variables: \( \text{fv}(\phi) \) that applies to valid conditions, and \( \text{FV}(K) \) that applies to DRSs.

**Definition 3.6:** Free and bound discourse referents. The set of free discourse referents for atomic condition \( \phi \), is returned by a function \( \text{fv}(\phi) \), which is defined as follows:

\[
\begin{align*}
(a) & \quad \text{fv}(x = y) \equiv \{x, y\}; \\
(b) & \quad \text{if } P \in \Phi^n \text{ is an } n\text{-placed predicate, } \text{fv}(P(x_1, \ldots, x_n)) \equiv \{x_1, \ldots, x_k\}.
\end{align*}
\]

The set of free discourse references for a DRS is defined by a separate function, FV:

\[
\text{FV}(\langle D_K, C_K \rangle) \equiv \bigcup_{\phi \in C_K} \text{fv}(\phi) \setminus D_K.
\]

Now we can define the function fv for complex conditions.

\[
\begin{align*}
(a) & \quad \text{fv}(\neg K) \equiv \text{FV}(K); \\
(b) & \quad \text{fv}(K \lor K') \equiv \text{FV}(K) \cup \text{FV}(K'); \\
(c) & \quad \text{fv}(K \Rightarrow K') \equiv \text{FV}(K) \cup (\text{FV}(K') \setminus D_K).
\end{align*}
\]
Discourse Representation Theory

Let $BV(K) \equiv D_K \setminus FV(K)$ be the set of bound discourse referents of $K$.

With this definition, we can define the propriety of a DRS.

**Definition 3.7: Propriety of DRSs.** A DRS $K$ is proper iff $FV(K) = \emptyset$.

Let us formally define accessibility. To do so, I define the properties of a DRS which is an immediate sub-DRS of another (based on Kamp, van Genabith & Reyle 2011: 146).

**Definition 3.8: Immediate sub-DRSs.** $K'$ is an immediate sub-DRS of $K$ if any of the following holds:

(a) $\neg K' \in C_K$;
(b) there exists a DRS $K''$ such that $K' \Rightarrow K'' \in C_K$, or $K'' \Rightarrow K' \in C_K$;
(c) there exists a DRS $K''$ such that $K' \lor K'' \in C_K$.

If $K'$ is an immediate sub-DRS of $K$, we denote this relation $K' < K$. $\preceq$ is a reflexive, transitive closure of $<$.

Now that we have defined the sub-DRS relation, we can formally define accessibility (based on Kamp, van Genabith & Reyle 2011: 147).

**Definition 3.9: Accessibility.** Let $K, K'$ be DRSs. $K$ is accessible to $K'$, denoted $K \text{ acc } K'$, iff one of the following conditions is met:

(a) $K' \preceq K$;
(b) there exists DRSs $K''$ and $K'''$, such that $K'' \Rightarrow K''' \in C_K$, and $K \text{ acc } K''$ and $K''' \text{ acc } K'$; or
(c) $K(Qx)K'$, where $Q$ is a quantifier.

Suppose that $x$ and $y$ are discourse referents, that $x \in D_K, y \in D_{K'}$, and that $K \text{ acc } K'$. In this case, $K'$ can apply conditions directly to $x$, even though $x \not\in D_{K'}$.

This definition may appear to simply repeat what has already been stated about accessibility; it does not give any intrinsic reason for accessibility’s existence. This results from the fact that accessibility arises from the model theoretic component of the model, and as such is a property that is extrapolated from the intrinsic parts of the model. Indeed, Geurts, Beaver & Maier (2016) state that a full specification of DRT may be given without stipulating accessibility, since it “follow[s] from the way the DRS language is interpreted”, and so “we are not at liberty to modify the accessibility relation, should we wish to do so, unless we simultaneously revise the truth conditions associated with the DRS language”. As such, the definition of accessibility given here is merely for convenience.

3.2.4 · Semantics

I introduce an intensional semantics for the DRS language that includes propositional attitudes. This is not the most basic semantics that can be defined for a DRS language, and indeed many applications of DRT do not require that such a semantics be utilised. However, it is clear from my introduction to game theory in the previous chapter that any valid
game theoretic model of communication should include a provision for multiple possible worlds. Furthermore, attributing attitudes to characters is a commonplace strategy for building plausible plot-worlds. And so, I opt for an intensional model as the basis for $L'$'s semantics (based on Kamp, van Genabith & Reyle 2011: 155).

**Definition 3.10:** Intensional model. $\mathcal{M}_L = \langle W, D, \mathcal{I} \rangle$ is an intensional model for $L$, such that $W$ is a nonempty set of possible worlds, $D$ is a non-empty domain of individuals, and such that $\mathcal{I}$ is an interpretation function where

(a) for names, $\mathcal{I}: \text{Name} \rightarrow \{\{d\}| d \in D\}$; and

(b) for $n$-ary relations, $\mathcal{I}: \Phi^n \rightarrow (W_\mathcal{M} \rightarrow \wp(D^n))$.

The interpretation function of the intensional model does two things. First, it gives a name to every entity in $D$. Importantly, it does so in the manner of a unary predicate, by defining a relation of members of Name onto individuals in $D$. As noted above, by condition (b) of definition 3.2, a name is a valid condition. Second, in order to deal with $n$-ary predicates, two steps are taken. (1) The set of worlds of the intensional model are mapped onto the powerset of $D^n$. $D^n$ contains every possible tuple of length $n$ that can be constituted of individuals in $D$, and the power set of $D^n$ includes every possible configuration of these tuples into sets. Therefore, $W_\mathcal{M} \rightarrow \wp(D^n)$ maps worlds to sets of $n$-tuples, whose elements are members of $D$, such that the set of tuples of individuals exists in a world $w$. Ultimately, then, this relation accepts a possible world $w \in W_\mathcal{M}$ and returns a set of $n$-tuples of individuals from $D$. (2) The set of predicates is mapped onto the set of $n$-tuples that results from the previous mapping. It thus applies a predicate $P$ to a tuple of individuals $(d_1, \ldots, d_n)$.

Since a given DRS should be viewed as a partial model of a world $w$, the truth of a DRS is determined by whether it can be felicitously embedded within a full model of $w$. The formal expression of this idea is that a DRS $K$ is true if there is a homomorphism from $K$ into $\mathcal{M}$. Another way to put is that a DRS $K$ is true iff for each $x \in D_K$, there is an entity $y \in D$ such that every condition on $x$ also holds for $y$.

**Definition 3.11:** Verifying embedding function. Let $L$ be a DRS language, and let $\mathcal{M}_L = \langle W, D, \mathcal{I} \rangle$ be an intensional model for $L$. A verifying embedding $f$ for a DRS $K$ into $\mathcal{M}$ for a (possibly empty) set of discourse referents $X \subseteq \text{Ref}$ is a homomorphism from $X$ into $\mathcal{M}$, $f : X \rightarrow D$. If this homomorphism does not exist for $X \rightarrow D$, then there is no verifying embedding. Verifying embeddings are understood as holding for all worlds.

Let $f$ and $g$ be verifying embeddings for $K$. $f$ extends $g$ with respect to a possibly empty set of discourse referents $X$, denoted $f \sqsubset_X g$, iff $\text{Dom}(f) = \text{Dom}(g) \cup X$. Since homomorphisms preserve structure, a discourse referent $x \in X$ may be associated with a $d \in D$ iff the conditions imposed on $x$ are equivalent to those on $d$.

Now that the verifying embedding function has been defined, we can define the intensional semantics for DRSs (from Kamp, van Genabith & Reyle 2011: 148–149).

**Definition 3.12:** Semantics. Let $L$ be a DRS language, and let $\mathcal{M}_L = \langle W, D, \mathcal{I} \rangle$ be an intensional model for $L$. The intensional semantics of $\mathcal{M}$ are defined by the following rules, where $f \models_\mathcal{M} w \varphi$ stands for “$f$ verifies $\varphi$ for $\mathcal{M}$ at $w$”:
3.2 · Discourse Representation Theory

\( (g, h) \models_{\mathcal{L}, w} \langle D, C \rangle \)

iff \( g \subseteq_D h \) and for all \( \varphi \in C \), \( h \models_{\mathcal{L}, w} \varphi \).

\( g \models_{\mathcal{L}, w} x = y \)

iff \( g(x) = g(y) \).

\( g \models_{\mathcal{L}, w} a(x) \) for \( a \in \text{Name} \)

iff \( \{ g(x) \} = \exists(a) \).

\( g \models_{\mathcal{L}, w} P(x_1, \ldots, x_k) \)

iff \( \{ g(x_1), \ldots, g(x_k) \} \in \exists(P) \).

\( g \models_{\mathcal{L}, w} \neg K \)

iff there does not exist an \( h \) such that \( \langle g, h \rangle \models_{\mathcal{L}, w} K \).

\( g \models_{\mathcal{L}, w} K \lor K' \)

iff there is some \( h \) such that \( \langle g, h \rangle \models_{\mathcal{L}, w} K \) or there is some \( h \) such that \( \langle g, h \rangle \models_{\mathcal{L}, w} K' \).

\( g \models_{\mathcal{L}, w} K \Rightarrow K' \)

iff for all \( m \) such that \( \langle g, m \rangle \models_{\mathcal{L}, w} K \), there exists a \( k \) such that \( \langle m, k \rangle \models_{\mathcal{L}, w} K' \).

The first of the conditions in definition 3.12 is the condition that describes the truth of a DRS. To be explicit, I give it its own definition (based on Kamp, van Genabith & Reyle 2011: 149).

**Definition 3.13: Truth.** Let \( \mathcal{L} \) be a DRS language and \( \mathcal{M}_\mathcal{L} \) be a model of \( \mathcal{L} \). A DRS is true iff there exist embeddings \( g \) and \( h \) such that \( g \subseteq_D h \), and iff for all \( \varphi \in C \), it holds that \( h \models_{\mathcal{L}, w} \varphi \).

With a definition of truth, we may define propositions. For a DRS, a proposition is a set of worlds such that the DRS is true at each world in the set.

**Definition 3.14: Propositions.** Given a proper DRS \( K \), the proposition \( [K]_{\mathcal{M}_\mathcal{L}} \) expressed by \( K \) relative to \( \mathcal{M}_\mathcal{L} \) is defined as

\( [K]_{\mathcal{M}_\mathcal{L}} \equiv \{ w | w \models_{\mathcal{L}, w} K \} \). (3.2)

I will often simply write \( [K] \), without the subscript

3.2.5 · Some omissions from the standard model

The two previous subsections outline the formal syntax and semantics of the model of DRT that I utilise in the remainder of the thesis. Other components of the formal apparatus of DRT, however, are not necessarily relevant to the aims of this thesis, nor to how DRT will be used. I conclude this exposition of DRT by pointing to some aspects of DRT that would appear highly relevant to the project at hand, and yet that I have purposefully omitted.

The first element that I omit are those definitions that are used to explicitly model changes in common ground. The most salient of these are the definitions of information states (see Kamp, van Genabith & Reyle 2011: 157), and of the context change potential (see Kamp, van Genabith & Reyle 2011: 159). Information states are the DRT implementation of what is meant by common ground throughout this thesis. This has been omitted because information states are in effect a duplication of states of the scene. States of the scene are preferred to information states simply because actors appear to be aware of states of the scene, and they furthermore use states of the scene in the construction of a plot. Since mine is a descriptive model of language use in improvised theatre, I prefer to use the aspects of
communication that appear to be used explicitly by actors. The context change potential is a formal means of relating DRSs to models, and the likelihood that a particular DRS will be felicitously embedded within a DRS of the common ground. This was omitted primarily because it relies on the definition of information states.

Information states and the context change potential are particularly suited to being implementation in game theory. The Bayesian games that we have observed so far deal well with updates in information, for they model the information states of players through their beliefs, which for the Receiver alter upon observation of a contribution to the common ground. The sequential introduction of information is furthermore central to the assumptions made in game theoretic linguistics, and so this assumption and its implementation may be harnessed to grant DRT a means of representing changes of the common ground, and especially of agents’ attitudes in that common ground. More encouragingly, these changes in game theory may affect the performance of players in games, thus generating a meaningful theory of the influence of common ground on interlocutor behaviour.

The second element that has been omitted is the part of DRT that deals with tense and aspect. The reasons are similar to above: the relative positions of DRSs in time is handled by the game theoretic component of the model, and so this particular function of the tense and aspect formalism is not required. Of course, the apparatus that encodes tense and aspect into the DRS language is not confined merely to the ordering of DRSs in time, but handles the semantics of verbs, so it would appear that a comprehensive theory of communication in game theory would include it. This is certainly the case. However, the chief reason that the apparatus of tense and aspect has been omitted is because it would compliment the model, for no commensurate gain. Any model necessarily selects a level of abstraction. As we will see, the model that I propose ignores certain salient questions about the role and cognition of the audience. This abstraction is necessary, however, since otherwise any results of the modelling would become very noisy. It is for the same reason, then, that I choose to exclude tense and aspect from the model. This does present some problems, particularly to do with information that is introduced into the scene regarding events that occur in the plot-world prior to the scene. These problems are not intractable, however, and using game theoretic formalisms I propose workarounds, when it is appropriate to do so.

3.3 · Common ground
The primary role of DRT in the thesis is to model common ground. Stalnaker (2002: 701) points to Grice (1989: 65, 275) as coining the term common ground, as the latter described certain propositions operating in the background which have “common ground status”. Stalnaker (1975: 273) gives a notional definition of common ground (emphasis is mine):

A speaker inevitably takes certain information for granted when he speaks as the common ground of the participants in the conversation. It is this information which he can use as a resource for the communication of further information, and against which he will expect his speech acts to be understood.
I have emphasised the second sentence of this description, because it gives an excellent summary of how common ground operates in a conversation in the way that I will formulate it. It is the information that is in the common ground of a conversation that interlocutors use as the foundation for further information to be added to the common ground, and it is according to information in the common ground that interlocutors interpret speech acts. This appears to be a fairly commonplace understanding of the related concept of context, which records the linguistic information that is introduced during a conversation. But the notion of common ground differs in a significant way, which Stalnaker (2002: 716) articulates well: “it is common ground that $\varphi$ in a group if all members accept (for the purpose of the conversation) that $\varphi$, and all believe that all accept that $\varphi$, and all believe that all believe that all accept that $\varphi$, etc.”.

Note the nuanced difference between belief that $\varphi$ and acceptance that $\varphi$. An agent’s belief that $\varphi$ requires a commitment to the veracity of $\varphi$. Acceptance, conversely, does not require a commitment to the veracity of the proposition, but only requires that agents recognise that it will function as if it were true in the conversation. This is to say, if a fact is in the common ground, the interpretation of an utterance must take that fact into account, since it is mutually accepted by all agents for the purpose of the conversation. This is precisely what I have in mind when I speak of common ground: the total sum of the propositions that have been accepted by parties to a conversation. As I show in §6.1.3, the process of agreement itself is one that is negotiated by the actors performing improvised theatre.

The common ground of a conversation is therefore a register of facts that are agreed upon for the purpose of that conversation by its participants. But how can this be represented? We saw in the examples above that, at the conclusion of the short discourses, we were left with a DRS that represented the common ground of the discourse. The common ground of any conversation is built in this piecewise fashion. The information that eventually constitutes the common ground is contained in DRSs that represent the utterances produced in the conversation. As we have seen, this occurs by iteratively merging utterance DRSs; and as such, the common ground of a conversation is simply another DRS.

The advantages of using DRT to model common ground are similar to those of using DRT to model utterance semantics generally. The particular DRS language that I have defined has an intensional semantics, meaning that the common ground can be bound to truth conditions that are relative to possible worlds. This relation to possible worlds naturally links the common ground with the game theoretic state of the world. Therefore, a common ground-as-DRS can be incorporated into game models easily, especially since most of the theoretical work has already been done with the introduction of intensional semantics into the game model. It also means that common ground information can include implications and quantified predications. Conditions that are associated by a logical operator can thus be included in the common ground. So, if a condition is introduced into the common ground, and that condition naturally implies another condition, this informa-
tion may be captured. I demonstrated an example of this in §5.2.1.2, where a stereotypical implication that causes a specific interpretation of an utterance is captured by DRT.

But there is one requirement of common ground that DRSs do not automatically meet. Predications that are introduced into the common ground must be compatible with those that are already in the common ground. In other words, new information should not contradict already established information. This requirement presents us with a problem both of generality and of specificity. On one hand, a definition of common ground that captures this fact must be general enough to be able to detect contradictions both in atomic and complex conditions, so that both (33) and (34) are rejected as valid common ground, since both contain internal contradictions.

Yet, the definition should be specific enough that it does not exclude DRSs that are merely improper, for these are not fit for our purposes, as we will see in chapter 5.

I propose that the optimal solution is given in definition 3.15.

**Definition 3.15**: Common ground. A DRS $K$ is a valid common ground, denoted $\Gamma$, iff $\exists K$ is nonempty.

The definition is general enough that it rejects a DRS with any contradictory constraints. The inclusion of a world in $K$’s extension implies that $K$ is true at that world, and the truth of $K$ at the world implies that all of the entities in $D_K$ and the conditions imposed on them can be matched with individuals and predicates in the model. Suppose that there are two conditions $\varphi$ and $\neg \varphi \in C_K$. $\varphi$ and $\neg \varphi$ are contradictory, since $\varphi \land \neg \varphi$ is always false. In this case, there can be no individuals that satisfy those conditions at any world. And so, $K$ is false at every $w \in W$ and $\exists K$ is empty. The definition is also specific enough that it does not discount a DRS as a common ground if it is improper. It therefore meets the requirements of common ground for the purposes of this thesis.

The properties of common ground are explored in depth in chapter 6, which is devoted to this topic.

### 3.4 · Discourse Representation Theory and game theory

In this section, I demonstrate that DRT and game theory are compatible, and specifically that the former can be used as the semantic component of the latter. I use this discussion
to propose a definition of an interpretation game, which will be the basis for all games discussed in the remainder of the thesis.

In normal signalling games, the interpretation of the utterance that is produced by the Sender is implied by the Receiver’s action. If we wish to explore the interpretation problems that face agents using natural language more closely, we might replace the Receiver’s selection of an action with the selection of an interpretation. What results is an **interpretation game**. Interpretation games take a variety of forms; see Franke (2011), Benz (2012b), for instance. However, there are two basic facts about such games that are common to most applications: (1) the presence of a semantic denotation function, a function that maps messages in the message set $\mu$ onto the set of world states; and (2) the selection of an interpretation that is not an utterance or an action. Interpretations themselves are not normally mapped to worlds by a semantic denotation function. Instead, interpretations are associated with certain worlds by aligning the players’ utilities such that if an interpretation is felicitous with respect to the world state, then it is preferred to those that are not.

This characterisation of interpretation is vague, since it does not describe what an interpretation is, and reduces the interpretive act to a simple choice between mathematical entities which are disconnected from actual language. A more fine-grained model of interpretation games requires a more linguistic definition of interpretation. DRT is able to fulfil this role. Recall the characterisation of interpretation by Kamp & Reyle (2011: 326) from the previous chapter:

> According to DRT, interpretation of an assertion one hears or reads takes the form of constructing a DRS for it. One way to think of this DRS is as a structure which the interpreter forms in his mind and which for him identifies the content of the interpreted statement.

Instead of selecting vague interpretations, then, the interpretation games that I devise involve selecting DRSs that represent the player’s mental representation of what she believes is meant by the utterances that she observes.

Linguistic signalling games were defined in the previous chapter (see definition 2.13 on page 23) as a tuple:

$$\langle N = \{S, R\}, \mathcal{W}, \rho_R, \sigma = \mathcal{W} \times \mu \times A, \text{den}, U \rangle.$$  \hspace{1cm} (3.3)

Let us discuss what is required to update each entity of this definition in order that we might yield an interpretation game. $N$ remains the same in interpretation games. $\mathcal{W}$ also persists, but note that $\mathcal{W}$’s usage becomes twofold. It functions not only as the set of world states for the game, but also the set of worlds for $\mathcal{M}_L$. The $\mathcal{W}$, in fact, acts as the most important bridging element between the game theoretic understanding of natural language games and the intensional model for a DRS language, since it provides a basis both for the felicitousness of a particular action and for the truth of a particular utterance. I develop the formal connection between the two shortly.
The Receiver’s prior belief, \( \rho_R \), takes the same form as in linguistic signalling games without DRT. It assigns a probability to each world state, and her posterior belief is an update of her prior belief based on the observed utterance. The Sender’s belief assigns a non-Bayesian probability to each interpretation that is available to the Receiver. \( \sigma \) is altered somewhat so that, instead of a strategy profile containing as its final element an action \( a \in A \), it contains a DRS from a set of DRSs. I denote a set of valid DRSs \( \kappa \). The Receiver’s strategy set becomes a set of DRSs, \( S_R = \kappa = \{K_1, \ldots, K_m\} \). So, the strategy space becomes \( \sigma = \mathcal{W} \times \mu \times \kappa \). The semantic denotation function is replaced by two functions: (1) an extension function \( [\cdot] \) that returns a DRS’s proposition, as per definition 3.14; and (2) a semantic function \( \|\cdot\| \) that accepts a DRS and returns a set of utterances that encode that DRS. The extension function is a definition that is part of the full specification of DRT. Meanwhile, the semantic function links the Sender’s strategies with the DRS language \( \mathcal{L} \). Instead of mapping natural language utterances in \( \mu \) to DRSs, which might appear more natural, it maps DRSs to natural language messages. I have chosen to define the function in this way because I believe it more correctly describes how agents make decisions, in relation to both the selection of an utterance for the Sender, and the interpretation of an utterance for the Receiver.

To illustrate this, consider the following. The Receiver views the world state \( w \in \mathcal{W} \), and decides on a message to send to the Sender. The message stands for her intended meaning, and can be represented as a DRS. Meanings cannot themselves be transmitted directly between interlocutors, and instead they must be encoded into natural language utterances to be emitted by the Sender.\(^8\) The Receiver’s task in interpreting the utterance is to deduce its intended meaning. She therefore ‘decodes’ the DRS representing the intended meaning, from the utterance. It is on the basis of the DRS that the Receiver deduces that certain world states become salient.

This approach is based on Benz’s approach to semantics as described in §3.1.3, and as such we can use a modified form of a diagram that he uses (Benz 2012a: 105) to describe the interpretive process in such a model. The diagram is presented in figure 3.1.

![Figure 3.1](image)

**Figure 3.1:** The interpretative process, based on Benz (2012a: 105). The world prompts the formation of an intended message, represented by a DRS. The message is encoded as a signal, which is interpreted as a DRS. Note that \( K \) and \( K' \) are not necessarily identical, but neither are they necessarily inequivalent.

However, unlike in Benz’s model, there is no linear progression from formulae (or from DRSs, as the case may be) to utterances, and from utterances to propositions. Both the set of utterances and the proposition are generated from the DRS. This results from the

\(^8\)This notion is central to the information theoretic conception of communication, and foreshadows the inclusion of that framework in chapter 7.
notion that any given DRS ought to be seen as a partial model of a world, and so DRSs have the flexibility of modelling both an agent’s internal mental representation of a given possible world—or part thereof—and the scope of an utterance’s representation of a given possible world. The DRS thus represents an interface between an interlocutor’s internal cognition and her use of natural language, and between these two and the actual world that the interlocutor is in. It seems appropriate to utilise DRSs produced by \( L \) to act as the formal entity that sits at the intersection of these different components of the model. A DRS, then, is associated with a set of worlds for which it is a partial model and with a set of utterances that it represents by the extension and semantics functions, respectively.

Lastly for interpretation games, the utility function remains unaltered. Aside from these alterations to the entities that are already constituents of the definition, we also need to add certain entities to the definition: the DRS language, to allow for the construction of DRSs; and its intensional model, to incorporate the semantics of the language. Putting everything together, we arrive at the following definition.

**Definition 3.16: Interpretation game.** An interpretation game is a tuple

\[
\langle N = \{S, R\}, L, \mathcal{M}_L = \langle W, \mathcal{D}, \mathcal{S} \rangle, \rho_R, \sigma = \langle W \times \mu \times \kappa, \|\cdot\|, \llbracket \cdot \rrbracket, U \rangle, \quad (3.4)
\]

where \( N \) is the set of players, \( L \) is a DRS language, \( \mathcal{M}_L \) is an intensional model of \( L \), \( \rho_R \) is the Receiver’s game theoretic belief, \( \sigma \) is the strategy space of the game, \( \|\cdot\|: \kappa \to \mu \) is a semantic function such that \( \kappa \) is a set of DRSs and \( \mu \) is a set of natural language utterances, \( \llbracket \cdot \rrbracket \) is an extension function, and \( U : \sigma \to \mathbb{R} \) is a utility function.

The order of an interpretation game is identical to that of a signalling game. I provide the following schema, however, to give a better idea of the precise steps taken by the agents in interpretation games.

1. Player 0 selects a world \( w \in W \) at random. \( S \) observes \( w \). She formulates an intended message, represented as a \( K \in \kappa \).

2. \( S \) emits \( m \in \mu \). \( m \) is an encoding of the intended message \( K \) as a natural language signal. \( m \) is usually an element of \( \|K\| \), but this is not required to be the case.

3. Based on her observation of \( m \in \mu \), \( R \) selects a \( K' \in \kappa \).

4. Utilities are distributed.

This formulation thus endows game theory with the rigorous, model theoretic semantics of DRT.

I have deliberately avoided placing any restrictions on the moves that are possible at any given turn. I have not, for instance, stated that \( S \) with an intended meaning \( K \) may only select an \( m \) such that \( m \in \|K\| \). I have likewise refrained from stipulating that if \( R \) selects an interpretation \( K' \), then she must do so with the expectation that \( w \in \llbracket K' \rrbracket \). That is, the model is agnostic on questions of the appropriateness of the player’s moves. Just as any utterance may be selected for any state of the world, even if it is totally unrelated to
that world state, any interpretation may be selected for a given utterance. Of course, this
may yield completely absurd moves. Consider (35).

(35) Lucky wears a hat.

While there is only one conceivable representation of (35), given in (36), a valid interpre-
tation of (35) in this model is, for instance, (37), or even more absurdly (38).

\[
\begin{array}{|c|}
\hline
x, y \\
\hline
\text{Lucky}(x) \\
\text{hat}(y) \\
\text{wears}(x, y) \\
\hline
\end{array}
\]

(36)

\[
\begin{array}{|c|}
\hline
x, y \\
\hline
\neg \text{Lucky}(x) \\
\text{hat}(y) \\
\text{wears}(x, y) \\
\hline
\end{array}
\]

(37)

\[
\begin{array}{|c|}
\hline
x \\
\hline
\text{Pozzo}(x) \\
\text{fat}(x) \\
\hline
\end{array}
\]

(38)

The decision to eschew any kind of stipulation on players’ behaviours has been made
for a number of reasons. It avoids implying that utterances have interpretations that are by
definition correct. I instead proffer a view of interpretation that is negotiated, so that the
meaning of any utterance is arrived at by an iterative process involving multiple agents. In
§6.1, I discuss the advantages of this approach to interpretation. For the moment, I justify
this notion by observing that a negotiated view of interpretation is found in studies of
both improvised theatre (see, for instance, Baumer & Magerko 2009, Magerko et al. 2009),
and linguistic pragmatics (see, for instance, Goffman 1967, Arundale 2006, Cooren 2000,
Silverstein 2001).

But it is also the case that not stipulating these kinds of rules follows best practice in
the formulation of game theoretic models. As noted, games are models of situations, but
do not per se prescribe how agents will behave in those situations. Instead, it is the role
of solution concepts to posit agents’ optimal strategies. Any meaningful solution concept
applied to an interpretation game will produce outcomes where the absurdities that I have
highlighted above are not present. We can tweak the definition of Bayesian equilibria
given in definition 2.18 (§2.1.3) to arrive at such a solution concept.
Definition 3.17: Interpretation equilibrium. An interpretation equilibrium in an interpretation game is a pair of strategies \( m^*(w) \) and \( K^*(m) \) and a belief \( \rho(w|m) \) satisfying the following requirements:

1. \( \rho_R(w|m) \) must exist;

2. \( K^*(m_i) \) must solve \( \max_{K \in \kappa} \hat{u}_R(w, m, K) \);

3. it must be the case that \( w \in \|K^*\| \) and \( m^*(w) \in \|K^*\| \);

4. \( m^*(w) \) must solve \( \max_{m \in \mu} \hat{u}_S(w, m, K^*(m)) \); and

5. for each \( m \in \mu \), if there exists \( w \in W \) such that \( m^*(w) = m_j \), then the Receiver’s belief must follow from Bayes’s rule and the Sender’s strategy: \( \rho_R(w|m) = \frac{Pr(m|w)Pr(w)}{Pr(m)} \).

This solution concept imposes strict requirements such that a strategy is optimal iff (a) the message is salient to the world state, and (b) the interpretation of the message is salient to that message. It selects those strategies that are optimal for players who are interested in communicating successfully. More importantly for this thesis project, however, it demonstrates how DRSs may be incorporated into a game model of linguistic interaction. It suggests that modelling interpretation selection as a selection between DRSs can operate effectively as a means of formally encoding the deliberation of a Receiver who interprets an utterance that she has observed.

3.5 · Conclusion

The primary purpose of this chapter has been to introduce Discourse Representation Theory, and to incorporate it as the semantic component of the game theoretic model that is used throughout the rest of the thesis. I motivated the inclusion of a rigorous semantic model by providing examples of other formulations of language games, and by identifying limitations in their methods of handling utterance semantics. I have suggested that DRT is able to better fulfil the role of a semantic theory for signalling games, because of its ability to represent players’ internal states, most importantly their interpretations, with a high degree of precision. Furthermore, DRT has the considerable advantage that it has a fully-realised, intensional semantics, which combines well with signalling games’ states of the world. Finally, I have provided a definition of interpretation games, which are, in the context of this thesis, games that include DRT as their semantic component. This involved reworking the typical definitions of signalling equilibria so that they operate with the inclusion of DRT.

The class of games that has been defined in this chapter, interpretation games, gives a detailed picture of the effects of interpretation, and of uncertainty when selecting an interpretation, on the decisions that are made by both the Sender and the Receiver. This breaks with current models of game theoretic interpretation, wherein the Receiver selects actions that stand for interpretations, rather than interpretations themselves. By providing a full specification of the nature of an interpretation, and precisely what constitutes the differences between interpretations, we are able to better account for the decisions that are
made by interlocutors. Furthermore, defining interpretation selection such that certain inter-
pretations are more probable, and that there is no intrinsically correct interpretation of
an utterance, means that the capacity of the model to describe naturally-occurring speech
is enhanced.

Interpretation games, especially in their iterated form, are the primary contribution
of the thesis to a generalised, game theoretic linguistics. In the next chapter, I show how
they can be used to describe a complex genre of naturally-occurring speech, in improvised
theatre.
This chapter, I present a game theoretic model of longform improvised performance. §4.1 begins the process of formally defining improvised performance. This section does not deal with the game theoretic component of the model, but rather provides a formal specification of many of the aspects of improvised theatre that were introduced in §2.2.2, including plot, character, and offer–response pairs. In §4.2, I identify ways that DRT can be extended to better cope with the inclusion of the formal definition of aspects of improvised theatre. In §4.3, I follow my reworking of certain elements of DRT with the inclusion of elements of improvised theatre in the definition of interpretation games. What results is the definition of IMPROVISED PERFORMANCE GAMES. Importantly, this section also provides definitions that allow for the modelling of multiple, consecutive improvised performance games, based on the iterated games introduced in chapter 2. I conclude the chapter in §4.4 with a caveat that, while the model is almost complete, it lacks the ability to describe the
covert communication that is frequently employed in improvised theatre. The capability of the model to describe covert communication is introduced in chapter 7.

4.1 · Elements of improvised performances

The elements of improvised theatre that I include in the model are based on the parameters set by the existing literature that I presented in §2.2.2, which includes Magerko et al. (2009), Baumer & Magerko (2009), Fuller & Magerko (2010). In this section, I provide succinct, mathematical definitions of most of the elements of improvised performance that are identified by this research. §4.1.1 introduces the formal understanding of plot-worlds, especially in relation to what has been discussed so far about game theoretic states of the world, and about possible worlds and propositions. The plot, scenes, and states of the scene are defined in §4.1.2. In §4.1.3, I describe the formal properties of actors and of characters, and make a determination of how they relate to one another in the games of the rest of the thesis.

4.1.1 · Plot-worlds

The maximal element of improvised theatre is the **plot-world**. A plot-world is a possible world that the actors construct in the process of generating a plot. It is the setting for a plot, and the process of world-building is one that actors think about in explicit terms (Magerko et al. 2009). To construct a plot, actors perform actions, most saliently utterances and gestures, which I call **events** (after Pressing 1988). Since the plots of improvised theatre are generated spontaneously, the plot-worlds of improvised theatre are invented over time through the performance of multiple events in sequence.

The view that I take in this thesis is that plot-worlds conform to the description of possible worlds articulated by Stalnaker (1984). They are ways that things might have been, and they are fully specified, meaning that two worlds that differ in a single fact are different worlds. Plot-worlds, like possible worlds that we might consider viable candidates for the actual world, should exhibit internal consistency. That is, it should not be the case in any viable plot-world that two facts \( A \) and \( \neg A \) both hold. The requirement of internal consistency means that plot-worlds constrain which events may occur as part of a given plot. That is, an event will not be considered plausible if it contradicts an already-established fact, and likewise, certain events are considered more plausible than others given what has been established about the plot-world.

A particular property of plot-worlds, identified by Elam (2002: 91), is that they are “never completely stipulated”. That is, “while logical worlds are assumed to be exhausted by their state descriptions […] the possible worlds of the drama have to be ‘supplemented’ by the spectator on the basis of his knowledge and hypothesising before they are fully constituted” (Elam 2002: 91). The information about the plot-world that is revealed is a small subset of the information that is sufficient and necessary for a complete description of the plot-world. It is certainly the case in improvised theatre that information about the plot-world is limited, and that evidence can be found that suggests that the actors and
the audience do infer supplementary information; I address this topic further in chapter 5. But I disagree with Elam that plot-worlds themselves are not exhaustive. The chief reason for my disagreement is that, since this thesis views improvised performance as a genre of naturally-occurring speech, the possible worlds that underpin the genre of communication are like those that underpin any given genre of naturally-occurring speech. Any interlocutor engaged in a conversation does not possess an exhaustive knowledge of the world. Interlocutors instead behave in accordance with their limited knowledge of the world. Similarly, actors do not possess exhaustive knowledge of the plot-world of a particular performance, but instead select strategies based on what they do know. It is absurd to claim that the incompleteness of interlocutors’ epistemology implies that the possible worlds in which they communicate are not exhaustive. The same holds of plot-worlds.

Formally, a plot-world is a possible world \( w \in W \) of the intensional model. This captures the relationship between a plot-world and the facts that are established in it. In \( \text{3.2.1} \), the conditions for the truth of DRS were established such that a given DRS is true iff it can be embedded within a larger, hypothetical DRS that specifies every entity and every condition on these entities in \( w \). If a condition for an entity has been established in the common ground of the performance, then any event that entails this condition’s falsity will be considered implausible. Therefore, DRT gives us the formal mechanism to deal with the plausibility of events. If \( K \) is a DRS that describes an event, and if there is no embedding function for \( K \) at \( w \), then \( K \) is implausible at \( w \).

### 4.1.2 \cdot The plot, scenes, and states

Let us now turn to the formal description of the plot of an improvised performance. I proffer an analysis of narrative that is predicated on a tripartite system, where each part is contained within the part before it. Plots form the most general level of a narrative, the plot is partitioned into scenes; and these are further partitioned into states. A graphic depiction of this system of analysis is presented in figure 4.1 overleaf. In the following subsections, I formally define these elements of narrative. In §4.1.2.1, I define plot; in §4.1.2.2, I define scenes; and I define states of the scene in §4.1.2.3.

**4.1.2.1 \cdot The plot**

A plot is simply a series of events, such that any two events may be related to one another by their order in time. An event \( e \) may precede, follow, or occur simultaneously with another event \( e' \), though the latter is less often the case than the former two.

I take a highly restricted view of what constitutes an event in improvised theatre. As I stated above, an event is understood as an utterance or a gesture that is performed by an actor. Since there are no random events in an improvised performance besides the

---

1 It is also for this reason that I do not directly draw on the extensive literature on possible worlds in fiction; see, for instance, Pavel (1975, 1986), Eco (1979), Ryan (1985), Doležel (1976, 1979, 1988, 1998).

2 I emphasise that this definition of plot-worlds is devised on the basis of this thesis’s primary concern, namely the use of language in improvised theatre. It may therefore be unsatisfactory for a theoretical investigation of theatre in general, especially because it does not remark on the relationship between the actual world—that is, the world in which the performance takes place—, and the plot-world (see Elam 2002).
Figure 4.1: A graphic depiction of a plot. The universe of possible events is contained within the double lines. The plot-world, represented by the purple ellipse, categorises events into those that are plausible and those that are not; the latter are outside of the ellipse, and the former are within. A plot is a series of events in the plot-world, which is partitioned into scenes, then into states. Each of the grey dots represents a potential event that is not part of the plot, while the larger, yellow dots are events that occurred in this particular plot. The line between each of the yellow nodes indicates the ordering in time of the events. States are indicated by solid, blue lines around events, and scenes are represented by dashed, red lines around sets of states. Note that the geography of the graph and the space itself are unimportant, and so in this figure are generated randomly.

actors’ actions, the events in improvised theatre are exclusively the actions of the actors.\(^3\) Events must be visible: the thoughts of actors are not counted as valid events for the obvious analytic problems that this raises. This point is made specifically with respect to interpretation games. Only the Sender’s move is visible, being an utterance, while the Receiver’s interpretation is invisible. Therefore, I restrict the working definition of an event further so that it is an action that is performed by the Sender in an interpretation game, whether it is an utterance or an action.

A plot is formally defined simply as a sequence of events such that these events are members of some plot-world.

**Definition 4.1:** Plot. Let \( w \in W \) be a possible world, and let \( T \) be an integer. A plot \( E \) is a \( T \)-vector of events \((e_1, \ldots, e_T)\).

While it may seem appropriate to model plots as something akin to Markov chains (see Gamerman & Lopes 2006), recall that plots come about only through the deliberate actions of actors, rather than through a stochastic process. It is therefore more salient to assign probabilities to the actors’ actions given their preferences than to state that a plot is simply

\(^3\)This of course ignores the unlikely cases where the venue of the performance loses power, or burns down, or similar other interfering events. One assumes that, in the event of such interference, the show would not go on.
the result of a stochastic process. I develop this idea further in §4.2. It is also important to observe that the plot is a somewhat artificial construct that is available in its entirety for observation and analysis only after the performance has concluded. In actual fact the plot is not the level at which a performance’s plot is devised and executed. Instead, it is at the level of individual offer–response pairs (see §2.2.2.2) that the actors construct a plot. Preferences over the plot should be seen as long-term goals of actors, and, while an actor might have a particular ending of a plot in mind as she performs it, that the intervening events will be determined is not possible in improvised theatre.

4.1.2.2 · Scenes

A plot is partitioned into a number of discrete scenes. Scenes are generally restricted to a certain small number of characters, and to a single location.

Definition 4.2: Scene. Let \( E \) be a plot, and let \( \psi \) be a partition of \( E \). A scene \( Q \) is a cell in \( \psi \).

Scenes are fairly unremarkable at the level of the mathematical analysis. They do, however, provide actors with opportunities for interesting strategies, which I discuss in §5.3.

4.1.2.3 · States of the scene

There is one further level of granularity of our analysis of plot: scene-states, states of the scene, or simply states (I use these terms interchangeably). Scene-states are perhaps the most important level of analysis of the narrative, because preferences over and beliefs regarding the state of the scene exert the greatest effect on the local deliberation of an actor. This claim is supported by evidence collected by Baumer & Magerko (2010) who found that the actions of actors in improvised theatre suggest that the actors are aware that a scene is composed of numerous states, and that it is according to this mutually understood composition that actors select their strategies.

A state of the scene’s most relevant information includes the following: (1) the location of the scene; (2) those existents who are involved directly (instead of being, for example, off stage); (3) the relationship between the existents. A state is a particular configuration of this information. More formally, a state is defined by a set of predications of the existents of the scene. Mathematically, a state can be identified with a given extension, which is determined by the common ground.

Definition 4.3: State. Let \( E \) be a plot, let \( T \) be an integer, and let \( 1 \leq t < T \). A state \( q \) is a proper subset of \( E \), such that for all \( e^t \in q \), the extension of the common ground \([\Gamma^t]\) is equivalent.

Definition 4.3 relies on the notion that the number of models that satisfy the DRS negatively correlate with the complexity of the DRS. Basing the definition of states on the extension of the common ground, rather than on the precise constitution of the common ground itself, avoids the sensitivity that would be inherent if the latter were used. Consider the sentences in (39) of a hypothetical performance of improvised theatre.
1. A: (pulls a trout out of the cupboard). I am holding a trout.
2. B: I know.
3. A: A trout is in my hands.

While there is no meaningful shift in the state of the information beyond B’s first utterance in line 2—the audience understands that A holds a trout, and that B knows this—, the common ground is updated, because each sentence is parsed and included in the common ground DRS. Conversely, the number of possible worlds that may plausibly be candidates for the possible world does not differ when A’s and B’s respective second utterances in lines 3 and 4 are produced. The state, then, should not shift, for no new information is added to the plot-world. Likewise, if an offer is made and subsequently rejected, meaning that the information does not enter the common ground, for common ground propositions must be agreed upon, then the state will not shift, since the extension of the common ground remains unaltered.

In particular, this definition of states of the scene is also sensitive to attribute ascriptions, for ascriptions are DRS conditions. This means that, as a character’s propositional attitudes shift, the state also shifts. This definition acts as a replacement for DRT’s information states (see Kamp & Reyle 2011: 161). It performs basically the same function, but in a simpler manner that is more appropriate for the task at hand. Furthermore, this definition is consistent with the discussions of the narrative moves that I conducted in chapter 2 (see §2.2), and which I resume in §4.2.3 of this chapter.

Kinds of offer and response. The narrative functions of offers and responses that are identified by Baumer & Magerko (2010), and that I introduced in §2.2.2.2, provide a good starting point for a quantification of states and state shifting with their categories of offers and responses in o–r pairs. The category of an offer is determined by whether it proposes to shift the state: we have state-maintaining moves (SMMs), and state-shifting moves (SSMs). On the other hand, responses are characterised by their function relative to the offer. There are two axes which interest us. The first axis describes whether a response accepts what is posited for inclusion in the plot-world. An affirmative response accepts what is posited in the offer into the plot’s common ground, while a negative response does not. For instance, if a predication \( P(x) \) is made in the offer, the affirmative response accepts that \( P(x) \), while a negative response states that \( \neg P(x) \). The second axis describes whether the response conforms to the function of the offer with respect to the state of the scene. If the response to an SMM maintains the state, for instance, then it is an augmentation; otherwise, it is a redirection. Table 4.1 overleaf summarises the four response types according to these axes.

The primary way that I will characterise a given o–r pair is by its effect on the state of the scene. This results from the kinds of offer and response that constitute an o–r pair.
\textbf{4.1 - Elements of improvised performances}

<table>
<thead>
<tr>
<th>conforms</th>
<th>accepts what is posited</th>
<th>rejects what is posited</th>
</tr>
</thead>
<tbody>
<tr>
<td>affirmative augmentation</td>
<td>affirmative redirection</td>
<td>negative augmentation</td>
</tr>
<tr>
<td>negative augmentation</td>
<td>negative redirection</td>
<td></td>
</tr>
</tbody>
</table>

\textbf{Table 4.1:} A summary of the possible response types.

<table>
<thead>
<tr>
<th>offer type</th>
<th>response type</th>
<th>shifts state?</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSM</td>
<td>affirmative augmentation</td>
<td>yes</td>
</tr>
<tr>
<td></td>
<td>negative augmentation</td>
<td>yes</td>
</tr>
<tr>
<td></td>
<td>negative redirection</td>
<td>no</td>
</tr>
<tr>
<td>SMM</td>
<td>affirmative augmentation</td>
<td>no</td>
</tr>
<tr>
<td></td>
<td>affirmative redirection</td>
<td>yes</td>
</tr>
<tr>
<td></td>
<td>negative redirection</td>
<td>yes</td>
</tr>
</tbody>
</table>

\textbf{Table 4.2:} How offer and response types combine to operate on the state. Recall that negative augmentations are impossible as responses to an SMM, and that affirmative redirections are impossible as responses to an SSM.

Table 4.2 presents each possible combination of offer and response kinds, and their relative effects on the state of the scene.

I provide a formal definition of o–r pair types in §4.2.3 below, since to do so requires a more fully-realised account of common ground, of the way that new information may be admitted into the common ground, and of how new information affects the common ground.

\textbf{4.1.3 - Actors and characters}

Unlike plot, a description of character is fraught. For an indication of the long history of debate about the nature of character, see the review by Chatman (1980: 107–138). I take as a starting point Baumer’s and Magerko’s work on improvised theatre, in which a character is viewed as the sum total of her traits (Baumer & Magerko 2009: 144), which include “relationships, goals, history, physical and mental attributes” (Magerko et al. 2009: 124). Actors attempt to ensure that traits remain, like those of individuals in the actual world, consistent throughout the performance. For instance, if a character is assigned a certain name in a performance, then actors attempt to call that character by the same name throughout the plot. This applies to all traits equally. I discuss the difficulties that the requirement of consistency adds in §4.2.1 below.

If we treat a character as the sum total of her traits, we can best represent a character formally using an attribute–value matrix (AVM) (see Copestake 2002 for a comprehensive overview of feature structures, of which AVMs are a kind). An AVM is a collection of pairs, such that each pair is constituted of a named attribute and a value for that attribute.
A key feature of AVMs is that they may be recursively embedded as values of attributes in other AVMs. AVMs are defined formally in definition 4.4, based on Blackburn (1993).

**Definition 4.4:** Attribute-value matrix. Let $\text{Att}$ be a finite set of attributes, let $\text{Val}$ be a finite set of values, and let $\text{Att} \rightarrow \text{Val}$ be a function. An attribute–value matrix is defined inductively:

1. if $\langle a, v \rangle \in \text{Att} \rightarrow \text{Val}$, then $\langle a, v \rangle$ is a valid AVM;
2. if $\langle a, v \rangle$ and $\langle a', v' \rangle$ are valid AVMs, and if $a \neq a'$, then $\langle \langle a, v \rangle, \langle a', v' \rangle \rangle$ is a valid AVM; and
3. if $\langle a, v \rangle$ is a valid AVM, then $\langle a', \langle a, v \rangle \rangle$ is a valid AVM iff $\langle a', \langle a, v \rangle \rangle \in \text{Att} \rightarrow (\text{Att} \rightarrow \text{Val})$.

AVMs are used often in linguistic semantics (see Pustejovsky 1991, 1995), and in syntax (see Gazdar 1985). As a simple example, the grammatical characteristics of the English pronoun *she* can be represented as follows.

\[
\begin{bmatrix}
\text{she} \\
\text{AGREEMENT} \\
\text{CASE}
\end{bmatrix}
\begin{bmatrix}
\text{gender} & \text{feminine} \\
\text{person} & \text{3rd} \\
\text{number} & \text{singular} \\
\text{nominative}
\end{bmatrix}
\]

I argue that characters may be represented and distinguished from one another by a finite set of traits, represented by the attributes and their values in an AVM. There is a single aspect of a character that may not be represented in this way, namely a specification of which actor plays a character. Treating this aspect of a character as an attribute does not suit the model, for the simple reason that a character’s traits should be verisimilar. In other words, since a character’s traits are supposed to be elements of the plot-world, it is inappropriate to model the metanarrative fact that an actor is playing the character in an improvised performance as a part of the plot-world. So, suppose that actor A plays a character named Pozzo, who speaks with a silly French accent. While his character’s name and accent are traits, the actor’s identity is instead included as an index or name of the AVM.

\[
\begin{bmatrix}
\text{A} \\
\text{name} & \text{Pozzo} \\
\text{accent} & \text{silly French}
\end{bmatrix}
\]

We will find throughout the data that are presented in subsequent chapters that traits, like any given fact regarding the possible world, are assigned only through the deliberate actions of actors. The requirement of deliberate action is important in distinguishing traits of the actor from traits of the character. Only the latter should be included, since traits of
the actor are strictly absent from the plot-world. For instance, a character’s accent is a trait of that character, but only if the actor playing that character makes an effort to speak with that accent. Or, if actor B asserts that A’s character’s name is Pozzo, then this is included as a trait of the character that A portrays.

The notion of deliberate action becomes central to a game theoretic understanding of improvised theatre, for it forms the chief criterion for what is included in the DRSs that describe improvised performance. Let us turn to these now.

4.2 A DRT for narratives

In this section, I wish to propose a means of representing the common ground established in the course of a performance of improvised theatre. My core claim, and one that motivates the use of DRT to describe improvised performance, is that any narrative in improvised theatre may be analysed as an iteratively constructed discourse. The means of constructing a discourse are precisely analogous to the means by which a discourse is constructed by interlocutors in talk-in-interaction of other genres, and, as I show in chapters 5, 6, and 7, many of the same norms of conversation apply to the communication in improvised theatre. However, unlike the discourses that are usually captured by DRT, narrative discourses include information that is specific to improvised theatre, such as the traits of the characters, the location of the scene, and so on. Therefore, to devise a formally valid theory of narrative discourse representation requires the stipulation of the kinds of information that may be included in a narrative DRS.

In standard DRT, the division is simple and very clear: information that may be derived from the syntactic structure of a sentence is included in a DRS, to the exclusion of all else. Under this view, only the explicitly linguistically encoded aspects of the utterance are treated as subsequent parts of the common ground; all components of implicit meaning are functions of the production of utterances rather than parts of the utterance’s semantics. That is, implicit meaning is inferred by the Receiver of an utterance. These kinds of inference rely as much on aspects of the situation and on player beliefs as they do on the semantics of a particular utterance.

The aspects of improvised theatre that would ideally be captured by a DRS for narratives, however, are more extensive and include the following: (1) the constitution of the utterances themselves; (2) traits of the characters; (3) facts about the plot-world and its existents that are communicated by gesture or by mime. The common aspect is that each requires actors to expend effort. As in other genres of naturally-occurring speech, actors must deliberately produce utterances. But unlike other genres, the traits of a character, and indeed the composition of the plot-world, are also determined solely by the deliberate actions of actors. It is most often the case that the narrative of any given scene requires physical objects that are not immediately available to the actor, such as a knife, an aeroplane control joystick, or a cat. And so the actors must mime the use of or interaction with these objects. The inclusion of these mimed objects is thus necessary for a sufficient model of narrative in improvised theatre, and a DRS should be able to account for them.
A DRS specified according to this criterion is able to capture and include far more specific and comprehensive depictions of the narrative common ground than a DRS in standard DRT.

Given such a broad chief criterion, one might well question how it is possible to set an interpretation of a mimed action that is definite enough to be able to posit that that interpretation be included in the common ground. I suggest that a means of ensuring that the information that enters a narrative common ground is of a similar quality is to stipulate that only the information that is an exact description of extralinguistic events be considered for inclusion in the common ground. For instance, if a player mimes holding a knife, then the fact that she holds a knife, and that the knife exists, should be included in the narrative common ground. If a player likewise indicates a direction, this information should be included in the narrative common ground. Note that in both of these examples, I have suggested that it is only a description of the physical action that is included. What the action is supposed to convey, instead, is omitted. So, if a player shrugs to communicate that she is uncertain of the answer to a question, only the shrug should be included in the DRS, and not what the shrug is meant to communicate. This stipulation is made with the same justification as the exclusion of any utterance meaning that is not explicit. Meanings that are intended to be conveyed by a gesture are treated as implied meanings of that gesture, for interpretations of gesture require the Receiver to infer the meaning based on common ground information.

I formally define DRT for narratives in this section, insofar as it differs from standard DRT. In §4.2.1, I provide a description of the key differences between standard DRT and DRT for narratives. I introduce a special predicate in §4.2.2, has_attr, that allows for the proposal of character traits using utterances. has_attr is introduced to link attributes ascribed to characters to the narrative common ground by providing a means of including attributes as a condition (i.e., a predicate) in a DRS. Since the truth of a DRS relies on it being embedded in a world, the truth of a DRS in a narrative relies on there being a plot-world that contains the characters with certain traits. has_attr allows for this to be realised. Finally, in §4.2.3, I describe how DRT for narratives handles offer–response pairs.

4.2.1 · What is different about DRT for narratives

The formal apparatus of DRT for narratives generally conforms to that of standard DRT. There are four major differences: the set of possible worlds; the constitution of \( \mathcal{W} \); the strategy-based criterion for inclusion of information in DRSs; and assertion of existence for on-stage characters. These are discussed in turn below.

The set of possible worlds \( \mathcal{W} \). If we are concerned with natural language sentences in naturalistic settings, the set of possible worlds of the intensional model of DRT is populated by candidates for the actual world. Generally these candidates differ from the actual world in only certain salient facts that are of concern. In DRT for narratives, the worlds in the intensional model are plot-worlds, which may differ greatly from candidates for the actual
world. Despite this, the same requirements of consistency, and the same definition of truth conditional verification apply to common ground in plot-worlds as to common ground in narratives.

While this may seem like an obvious point in the abstract, consider for instance a scene where zoo animals speak English to one another, but cannot understand the English that is spoken by the human characters in the scene. There is an internally consistent set of truths that are verifiable at a world in $W_{gr}$. And if some event violates this truth—for instance, if an animal interacts directly with a human—, then either an error has been committed, or else the facts of the interaction between animals and humans are more complicated than at first established. A clearer example still is the practice of an actor addressing the audience directly. It is practically always the case that only certain characters are allowed to address the audience, which is a distinction that becomes embedded in the truth conditions of the plot-world. Indeed, it is often the case that lines that are addressed to the audience have the restriction that they cannot be heard by the other characters in the scene, and so the beliefs and preferences of the other actors are unaffected by this kind of utterance. This all suggests that, while these characteristics are typical of the plot-worlds of improvised theatre, they are governed by the ‘natural laws’ of that plot-world. Plot-worlds therefore deserve to be treated analogously, and not identically, to candidates for the actual world.

The constitution of $D$. Perhaps the most radical difference between standard DRT and DRT for narratives is that the $D$ is constituted of the existents of the plot-world. The category of existents covers all animate and inanimate entities in the plot-world, and also includes other things such as places or conceptual entities. As such, an existent, for the purposes of this thesis, is anything that may be indexed. This does not differ greatly from other types of possible world, except in one key fact: a major class of existent is character, which is specified in the formal model as attribute-value matrices. And so, $D$ includes a set of AVMs that is indexed by names by the interpretation function $\mathcal{I}$ of the model. Characters are treated as ‘real’ individuals, which means that they are ‘fully’ specified: a character has a complete history, consistent physical traits, realistic desires, and so on. But characters are realised progressively, and are invented or discovered throughout the execution of a plot. In the same way that a DRS is only a partial model of a world $w$, what is known about a character is treated as a partial specification of the full AVM that defines that character.

Any fact that is introduced about the character must therefore be verifiable against the known facts of the character, as DRSs must be verifiable with the plot-world. What I have proposed here, then, gives us a formal definition of consistency of character.

**Definition 4.5**: Character consistency. Let $\mathfrak{M}_L = \langle W, D, \mathcal{I} \rangle$ be a model of language $L$ such that $W$ is a set of plot-worlds, and such that there is a set $AVM \subset D$ of character AVMs, and let $E$ be a plot with common ground $\Gamma_E$. A character $A \in AVM$ is consistent iff there is a verifiable embedding $f$ such that for every $e^I \in E$, $f \models_{\mathfrak{M}_L w} \Gamma^I$.
This definition states that the common ground throughout the performance should be verifiable in a particular world of the model. Recall that, for a DRS $K$ to be verifiable, there must be entities in $D$ that correspond to entities in $D_K$. And therefore, there must be entities in the common ground that correspond to entities in the model’s set of individuals. Furthermore, specifying this across all events in the plot means that entities must preserve their correspondence for the entire length of the plot. Therefore, the descriptive game theoretic model that I develop in this thesis has a formal definition of consistency, which the decision theoretic model of Magerko et al. (2009) lacks.

The strategic criterion. Since the physical resources that are available to the actors tend to be insufficient for the demands of their performance, they often have to use mime to convey facts about the scene. In chapter 5, I describe scenes in which mime is used to establish a scene in a range of locations, such as in an office, or in the cockpit of an aeroplane. Furthermore, miming can be used to introduce various entities into the scene. It is necessary that a model of improvised theatre be able to properly account for the inclusion of facts about the plot-world that are not able to be derived by the utterances spoken during an improvised performance alone.

As discussed in the introduction to this section, I include simplistic descriptions of the actions as natural language sentences that are subsequently incorporated into the common ground DRS. This is based on the simple intuition that physical actions may be interpreted, though perhaps not to the same degree as natural language utterances. In chapter 2, we examined how a Receiver’s strategy can be interpreted by the Sender as an indirect signal of which state of the world the Receiver reckons is the true state. Similarly, it is conceivable that, by performing a certain action at a certain time, with a certain common ground supporting that action, an action can be used by the Sender to signal the true state of the world to the Receiver. Since the model that is formulated here is game theoretic, I call this approach a strategy-based approach. Anything that is used as a Sender strategy is able to be inserted into a DRT for narratives. As I have indicated, this includes actions that are not natural language utterances. If $A$ picks up a cat, $x$, then the description of the action would be a two-place predicate,

$$\text{picks_up}(A, x).$$

Similarly to the description of utterances by giving their sentence form, these predicates should describe the activity precisely, and should not describe the information that is inferred by any Receivers of that message.

Existence of on-stage characters. A somewhat unexpected consequence of the strategic criterion is that simply being on stage qualifies as a special kind of existential assertion. Actors must expend effort to be on stage when a scene begins, or to enter the stage partway through a scene. Such actions therefore qualify for inclusion in a DRS.
But this raises the question: what is being communicated? The presence of an actor on the stage predicates very basically of her character that (a) the character exists; and (b) the character is present in the scene. And so the actor’s action of being on the stage communicates both of these facts. These statements are implemented easily in DRT, by simply including an actor in the universe of the common ground’s DRS. This does not impact on the propriety of the DRS. I therefore use such a method when a character first enters the scene.

4.2.2 · Character traits and Discourse Representation Structures

While the definition of character consistency provided above in definition 4.5 allows the model to take into account the traits of characters, it does not grant us the means to deal with the way that ascriptions of traits arise. As I will show in the data that will be presented later in the thesis, a character’s traits are almost always predicated through an utterance. Consider names, for instance. It is most often the case that actor A will simply call B by a certain name if a name has not been established, which asserts that B’s character has an attribute name, and provides a value for that attribute. If A calls B Bob, then the AVM for B’s character becomes the AVM in (43).

\[
\begin{array}{c}
B \\
\text{name} & \text{Bob}
\end{array}
\]

Standard DRT is meant to deal with names and predicates, and so an interpretation function is built into the intensional DRS model that was presented in the previous chapter.

But what standard DRT is not fit to deal with are strategies such as the aspects of communication that are not strictly part of the utterance-as-string-of-words, such as use of accents, the effects on interpretation that arise from the physical characteristics of a character, and so on.

A simple strategy to overcome this limitation, and one that I will employ more extensively in subsequent chapters, is to include traits as predications of the character, in the same way that a predication might be included were it explicitly stated. So an accent, for instance, is included as an attribute of that character. For this purpose, I introduce a predicate that is used specifically for situations where an attribute is applied to a character: has_attr, a trivalent predication function that maps the universe of a DRS to a set of attributes. A mapping of an entity in the DRS’s universe to an attribute implies that the entity possesses that attribute. So, for instance, if character A speaks with a Cockney accent, this is represented in the DRS in (44).

\[
\begin{array}{c}
\text{A} \\
\text{has_attr(A, ACCENT, COCKNEY)}
\end{array}
\]

Or, for instance, if a character walks with a limp, it is predicated of the character that has_attr(A, GAIT, LIMP).
In these examples, the attributions are made of the characters by the actors who are playing those characters, but this is not necessarily true of all instances of has_attr. Instead, an actor may predicate of other actors’ characters that they have a certain attribute.

The formal definition of has_attr is provided as follows.

**Definition 4.6:** Special predicate has_attr. Let $L$ be a DRS language, let $\alpha$ be a set of characters, and let $M_L = \langle W, D, I \rangle$ be a model of $L$, such that $AVM \subseteq D_M$ is a set of valid AVMs. has_attr $\in \Phi$ is a three-place predicate has_attr$(A, a, v)$, where $A \in \alpha$, $a$ is an attribute of $A$, and $v$ is a value of $a$.

When an actor ascribes a trait to the character of her fellow actor, this can be encoded in the DRS that represents that utterance, therefore it can be accounted for according to the truth conditions of the narrative DRT model.

### 4.2.3 · Offer–response pairs and DRSs

I have already established, following Baumer & Magerko (2009, 2010), that the basic unit of analysis of improvised theatre is the offer–response pair. While I concur with the position espoused by Baumer and Magerko, I do so for different reasons, since I offer an analysis that is linguistic in nature rather than purely cognitive. The offer–response pair as the basic unit of analysis works for a linguistic account of improvised theatre because it places common ground at the centre of the structure of an improvised performance. Making offers and responding to those offers are both events that come about only through the production of utterances, and their subsequent interpretation. O–r pairs reflect the fundamental structure of signalling games, which are not only the most basic unit of game theoretic linguistic analysis, but perhaps the most basic unit of the analysis of communication that is possible. Furthermore, since any plot is a common ground, the most basic level of establishing common ground in a situation involving multiple interlocutors is the production and interpretation of a single utterance.

I propose that the key mechanism that bridges DRS and o–r pairs is the use of the merge operation. To illustrate this, let us consider a highly abstract communicative scenario involving two agents, $i$ and $j$, which has the following order:

1. A common ground $\Gamma'$ exists and has extension $[\Gamma']$.
2. In turn 1, $i$ produces an utterance $m$ with a DRS $K_m$, such that $m$ introduces new information; that is $[K_m] \neq [\Gamma']$.
3. In turn 2, $j$ interprets $m$, and then selects whether to accept what $m$ proposes. She does so by signalling her acceptance or rejection. If $j$ accepts $m$, then the common ground is updated so that the common ground after the discussion becomes $\Gamma^1 = K_m \cup \Gamma^0$; if not, $\Gamma^1 = \Gamma^0$. $j$ may also partially accept $K_m$, such that some subset of the conditions of $K_m$ (and the entities that those conditions apply to) is merged with $\Gamma^0$. The relevant alterations to the extension of the common ground are made.
This is, in the abstract, how o–r pairs behave in improvised theatre. i produces an offer, and j subsequently responds to that offer. j’s response may accept or reject information that is proposed by i’s offer. The various types of o–r pairs that are described above (see table 4.1 in particular) must be described, and should be understood, in relation to a common ground.

**Definition 4.7: Offer types.** Let E be a plot, and let Γt be a common ground at turn t. An utterance at turn t is a state-maintaining offer iff
\[ J_{\Gamma t} K_m^t \subseteq J_{\Gamma t} K_{m-1} \cup J_{\Gamma t} \Gamma_{t-1}, \] (4.1)
and is a state-shifting move iff
\[ J_{\Gamma t} K_m^t \not\subseteq J_{\Gamma t} K_{m-1} \cup J_{\Gamma t} \Gamma_{t-1}. \] (4.2)

The logic of these definitions is identical to my discussion of state above: a state shifts if the extension of the common ground is altered. From this definition, it is a simple matter to define the various response types.

**Definition 4.8: Response types.** Let E be a plot, let Q be a set of states, let m_{t-1} be an offer and m_t be a response to m_{t-1}, and let Γt be a common ground at turn t. m_t is affirmative with respect to m_{t-1} if
\[ \Pr(K_{m_{t-1}} | m_t) > \Pr(\neg K_{m_{t-1}} | m_t), \] and negative if
\[ \Pr(\neg K | m_{t-1}) > \Pr(K_{m_{t-1}} | m_t). \] If \[ \Pr(K_{m_{t-1}} | m_t) = \Pr(\neg K_{m_{t-1}} | m_t), \] the polarity of the response is ambiguous. The conformity of the response is defined for an SMM relative to state q_0 ≠ q_1 as follows:

- m_t is an augmentative response iff
\[ J_{\Gamma t} K_m^t \subseteq J_{\Gamma t} K_m^t \cup J_{\Gamma t} \Gamma_{t-1}, \] and if e_{t+1} ∈ q_0; and
- m_t is a redirection response iff
\[ J_{\Gamma t} K_m^t \not\subseteq J_{\Gamma t} K_m^t \cup J_{\Gamma t} \Gamma_{t-1}, \] and if e_{t+1} ∈ q_1.

If m_{t-1} is an offer that shifts from q_0 to q_1,
- m_t is an augmentative response iff
\[ J_{\Gamma t} K_m^t \subseteq J_{\Gamma t} K_m^t \cup J_{\Gamma t} \Gamma_{t-1}, \] and if e_{t+1} ∈ q_1; and
- m_t is a redirection response iff
\[ J_{\Gamma t} K_m^t \not\subseteq J_{\Gamma t} K_m^t \cup J_{\Gamma t} \Gamma_{t-1}, \] and if there exists some e_{t+1} ∈ q_2 such that q_2 ≠ q_1.

This definition gives a full specification of possible types of offer–response pair.

### 4.3 · Improvised performance games and their iterated forms

The formal representation of facets of improvised theatre has now been developed. In this section, I turn to the further development of interpretation games, which were introduced in the previous chapter, so that they are rooted more strongly in a description of improvised theatre. The resultant type of game, called **improvised performance games (IPGs)**, takes what has been described so far in this chapter, and attempts to distill it into a basic, extensible game model. In §4.3.1, I formally introduce the kind of game, and discuss the differences between the entities that are present in the formal definition of an IPG, and their analogues in previously introduced games. In §4.3.2, I formally include iterated games into the game theoretic model of improvised theatre. This allows for the game theoretic description of entire scenes, and entire plots.
4.3.1 · Improvised performance games

Improvised performance games are a type of interpretation game (see Dekker & van Rooij 2000), where the elements of improvised theatre that have been defined so far in this chapter are included as components of the game. An improvised performance game follows the same basic structure as an interpretation game, but with some amendments:

(i) a state of the scene \( q \in Q \) is selected;
(ii) the Sender \( S \), who is always a member of the set of actors, observes the state of the world, and produces an utterance \( m \) in the set of utterances \( \mu \), which has semantics \( K_m \) (that is, \( m \in \|K_m\| \));
(iii) the audience and all actors (except for \( S \)) select an interpretation \( K \in \kappa \).

The final stage of the game is the most complicated. Realistically, the interpretations are all selected simultaneously by every player who makes an interpretation. However, at the level of the game theoretic analysis, the interpretations are made in a sequence. This is preferable, because the selection of the state of the world occurs prior to \( S \)'s utterance, which in turn occurs prior to any interpretation event. We can model \( i \)'s lack of knowledge of the interpretation selected by \( j \) be excluding \( j \)'s interpretation from \( i \)'s private history.

Improvised performance games are formally defined in definition 4.9.

Definition 4.9: Improvised performance game. An improvised performance game (IPG) is a tuple
\[
\langle N, \mathcal{L}, \mathcal{M}_L = \langle W, \mathcal{D}, \mathcal{I} \rangle, Q \subseteq E, \rho_S, \sigma, \|\cdot\|, U \rangle
\]

such that

- \( N = \alpha \cup \{P\} \), where \( \alpha \) is the set of actors, and \( P \) is the audience;
- \( \sigma = Q \times \mu \times \kappa \times \ldots \times \kappa \) is the strategy space;
- \( \|\cdot\| \) maps world–strategy profile pairs to vectors. These vectors have a length that is equivalent to the cardinality of the union of the set of actors and the set containing the audience; and
- there is a set \( \text{AVM} \subseteq \mathcal{D}_W \) such that each element of \( \text{AVM} \) is a character AVM.

This definition is complicated, and the entities that are included require more explanation and justification, which I undertake in the following paragraphs.

The players. While it should be the case that actors are included as players in an improvised performance, it is perhaps unclear why the audience should be included as a player. The audience has a very limited strategy set, which includes only interpretation selection moves. Audience feedback, such as laughter, is not included, because these actions are often not rational decisions.
The audience is included as a player because their interpretation enforces the requirement of verisimilitude that appears to carry with it a significant degree of gravity for the players. For the audience, any given utterance should be plausibly attributable to a character in the plot-world. Conversely, the primary purpose of communication for the actors of an improvised performance is the successful communication of information regarding the plot of that performance. So, the audience is included to provide an agent for whom the communication of the actors must appear verisimilar. I justify this choice more thoroughly in chapter 7, for the formal means of dealing with concealment of information relies on the presence of an agent from whom the information is concealed.

States of the scene. Nature’s move probabilistically selects states of the scene, rather than states of the world or worlds themselves. This decision is informed primarily by four factors.

First, it adheres more closely to empirical findings of the behaviour of actors. Baumer & Magerko (2009, 2010) claim that, according to the data that they collected, actors’ behaviour indicates that they structure plots according to states and to scenes. It is appropriate that the model treat states of the scene likewise as units of the plot structure.

Second, the progression of certain states of the scene during a plot implies a specific plot-world. A single event may exist within multiple plot-worlds, provided that it is plausible within that plot-world. Each event in a performance discounts the possibility of all those plot-worlds in which that event is implausible. As the partial model of the plot-world becomes more comprehensive by subsequent utterances, the number of plot-worlds that the model will be able to felicitously be embedded within decreases. Put differently, the more we know about a plot-world, the fewer reasonable candidates for that plot-world there are. And since scene-states are sets of events, the same narrowing applies to them. Only certain scene-states are included in a given plot-world, for the events that make up the scene-states may likewise exist only in a given plot-world. Therefore, states of the scene, when taken in aggregate over a larger period of plays of a game, can act as a valid replacement for selection of worlds.

The third factor follows from the second. It is implausible that an actor has an exhaustive specification of the plot-world in mind when she performs the first action of a scene. A complete specification would engender a complete understanding of the entire history of the plot-world, and a complete specification of the attributes of every existent in the plot-world including of those that are not present in the scene. Indeed, an actor’s possessing of a specification to this degree of completeness is implausible at any stage of a performance. It is more justifiable to claim that an actor has a partial model of a particular state of the scene, which indicates a number of different candidate plot-worlds.

Finally, selection of the state of a scene in an IPG can represent the Sender’s preferences for the trajectory of the plot in that game. Consider that, in Vladimir’s Bomb Dilemma in §2.1.3, Estragon was interested in signalling the correct world to Vladimir; Estragon preferred that Vladimir behave as if a certain world was the correct world. Likewise, when
an actor has a preferred plot trajectory, she wishes to communicate that preference to her fellows. In an IPG, then, the selection of the world by Nature is the formal implementation of this preference. In iterated IPGs, this device can model an actor’s changing preferences over time, for at each period that she is the Sender, a different world may be selected by Nature.

For these reasons, states of the scene are preferred to plot-worlds for Nature’s selection.

The strategy space. The definition of the strategy space is somewhat complicated, but it needs to be so because the size of the set of actors is changeable. Note definition 4.9 specifies the order of turns, by stipulating that the Cartesian product that defines a member of $\sigma$ must contain the selection of a state, a message, and interpretation by the remaining players. The repetition of the latter occurs $|\alpha \cup \{P\}|-1$ times: the ‘$-1$’ term ensures that the Sender, who is a member of $\alpha$, is not counted in this operation, since she produces $m$. Note that, in keeping with what was posited above, the order of agents does not matter. As I show below, the lack of knowledge of players of iterated IPGs means that a Receiver is only able to view her own interpretation, and she must infer the interpretations that are selected by her fellows. All Receivers operate under these epistemic conditions.

Utility. The assignment of a utility to the audience is motivated by similar factors to its inclusion in the set of players. If the audience is granted a utility, and if the audience is interested primarily in the verisimilitude of the performance, then this will be reflected in the utility that they receive. I show in chapter 7 that the audience’s utility can be mobilised to increase the strength of the game theoretic model of improvised theatre.

The definition of improvised performance games that I have introduced provides a foundation for the game theoretic description of improvised performance. In the next subsection, I extend the model so that IPGs become the stage game of an iterated game model.

4.3.2 Plots as iterated games

The model that I have sketched so far is able to account for the smallest part of an improvised performance, where an actor produces an utterance and the utterance is interpreted by another actor. However, with such a model, we are unable to describe even offer–response pairs, which involve a sequence of two such games, let alone a plot that is constituted of a large number of o–r pairs.

To solve this, and to endow the model with the capacity to describe larger-scale narrative structure, I introduce the formal apparatus of iterated games into the model. This introduction addresses four major topics: the order of turns in §4.3.2.1; the form that histories take in iterated improvised performance games in §4.3.2.2; the measure of utility that is used in iterated improvised performance games in §4.3.2.3; and proposing a solution to
the problem of conversations being constituted of multiple Senders and multiple Receivers in §4.3.2.4.

Before we proceed, let me address a potential criticism of the approach taken here. It may be argued that what I define in this subsection is not precisely an iterated game, but is instead analogous to an iterated game. The point at which my model seemingly departs from iterated games is that at each period of an iterated game, the strategy set of the stage game must be exactly equivalent. This may appear to be a suboptimal modelling decision for a model of linguistic behaviour, since the linguistic strategies that may be selected in a discourse are highly contingent on those that precede them. For instance, if a trait is ascribed to a character at the beginning of a scene, and if the actors wish the characters to be verisimilar, then any utterance that contradicts this ascribed trait will be greatly dispreferred by the actors—perhaps to the extent that such strategies will be deleted from the strategy sets entirely, thereby rendering the strategy sets different at different periods of the game.

The model that I propose deals with this problem in a different fashion. The strategy set for the actors is, at every period of the iterated game, an exhaustive set of the natural language utterances that are available to actors. Instead of deleting utterances, established information adjusts the probabilities of the members of the strategy set. If an utterance contradicts an established fact, then its probability will be lower than that of an utterance that is consistent with this fact. Note that this solution is consistent with the game theoretic definitions of strategy spaces and strategy sets—which may contain strategies with zero probabilities—and of beliefs—which is the primary means by which certain strategies become less probable given a certain context.

This solution provides a more robust, general model than one in which certain strategies are deleted. Deletion of strategies presents problems that a probabilistic approach does not. For instance, the threshold at which a strategy might be deleted is unclear, whereas strategies with zero probabilities can be identified using the information established in a scene, and Bayes’s equation. A more practical difficulty with deletion is that it becomes difficult to handle the errors that actors might commit. If utterances that contradict prior information are deleted, then there is no way that an actor could produce an utterance that contradicts what is already established; but, as I discuss in §5.4, these errors are committed by actors in the dataset, and therefore must be accounted for. A probabilistic approach is, therefore, more appropriate than outright deletion of strategies, and such an approach underpins the definitions that follow in this subsection.

4.3.2.1 · Order of turns

Infinitely iterated games are used to model improvised theatre because the length of scenes is not determined prior to their execution. In Harold performances, the actors that are off stage usually terminate a scene when they feel it appropriate to do so. In the data that I

---

4I thank an anonymous examiner for bringing this to my attention.
have collected, this is performed by an actor running across the stage with a hand raised above her head, or by interrupting the scene with the introduction of another scene.

The order of turns for each period of the iterated game is the same as a normal improvised performance game. But there is one amendment to the order posited above that I would like to make: prior to the first period of the game, Nature makes a one-off selection of a plot-world. I refer to this selection as occurring in period $t = 0$. This formal device encodes the fact that the first actor to speak in a scene generally sets the location and premise of a scene. For instance, in §5.2.1.2 I examine a scene in which the first utterance of the scene sets the scene in a Mexican restaurant, establishes a character as a Mexican cook in the restaurant, and posits a premise that patronage of the restaurant has been reduced. These are facts about the plot-world that are established by the first utterance, and they form the basis of the rest of the plot; indeed, none of these facts change once they are established early in the scene. And so the actor who utters the first line of a scene has the special position of determining important information about the scene. This is formally implemented by setting a $w$ before the scene commences, and stipulating that the selection of $w$ is revealed only to the Speaker at $t = 1$. I formally specify this in definition 4.10 in the next subsubsection.

4.3.2.2 · Histories

Recall from §2.1.4.2 that, according to Aumann, Maschler & Stearns (1995) each player is informed of her history at the conclusion of each period. Histories are members of the strategy space, and a history that a player is informed of may not correspond to the actual strategies that are played. I would like to preserve this notion in spirit, but it needs to be adjusted to suit IPGs.

In improvised theatre, and indeed in any linguistic interaction, the only element that is available to be observed by every interlocutor are the utterances, mimes, and gestures that are produced in that interaction. The histories of agents, then, ought to include the correct message, for it is absurd to claim that while an agent may observe the utterance that is produced in an interaction, she is not informed of it as part of her history. However, there are differences in the provision of information about states and interpretation. While the Sender of an utterance cannot observe the interpretation that the Receiver selects—for interpretation selection is a process that occurs in the Receiver’s mind—, the Receiver certainly should have the interpretation that she makes in her history. Likewise, the state of the world should be included in the Sender’s history, for it represents her preference for the plot’s trajectory, but it should not be included in the Receiver’s history.

And so I propose that only utterances should be included in every player’s history, and that the moves that the player cannot observe are determined probabilistically. For the Sender, an estimate of the Receiver’s interpretation is included in her history; and for the Receiver, an estimate of the $q$ selected by Nature is present. Here, I depart from the formulation of Aumann, Maschler & Stearns in that I do not propose that these parts of the players’ histories be determined stochastically. Instead, it is on the basis of the players’
respective beliefs that the contents of their histories are determined. Let me define this situation formally.

**Definition 4.10: Histories in iterated IPGs.** Let $G^T$ be an iterated improvised performance game with horizon $T$. At the conclusion of each period, $S$ is informed of her history, which is found by solving $\eta_S = (q, m, \max \delta_S(\kappa))$, and $R$ is informed of her history, which is given by $\eta_R = (\max \delta_R(Q), m, K)$. At period $t = 1$, the Sender’s history includes the world that was selected by Nature, such that $\eta_S^0 = (w)$, and the Receiver’s history $\eta_R^0 = \emptyset$.

This definition of histories in IPGs attempts to capture the information that is available to the different players of an IPG. It thus formally encodes actors’ epistemic states into the game model, and it provides us with a basis for the definition of utility in improvised theatre.

**4.3.2.3 · Utility**

The definition of a utility in iterated games is a function of the history of a play of the game onto $\mathbb{R}^n$, where $n$ is the number of players. This implies that a utility requires a complete history, and that utilities cannot be assigned prior to the end of a game. That is, players are not made aware of their utilities until the conclusion of the iterated game. However, since each player is informed of her private history at the conclusion of each period, each player may estimate the utility. At the conclusion of each game, then, each player calculates a **provisional expected utility**, which she uses to inform her moves throughout the rest of the performance. The provisional expected utility represents a player’s assessment of her action in the context of the other actions played during that turn, and therefore it provides players with a formal mechanism of continuous, iterated appraisal of their success or failure. I show in chapters 5, 6, and 7 that such a mechanism is necessary and sufficient for a game theoretic model of improvised theatre, for it allows for shifts of preferences as the common ground alters.

**Definition 4.11: Provisional expected utility.** Let $i$ be a player, $G^T$ be an iterated game with horizon $T$, and let $\eta_i^t$ be $i$’s history at period $t$. $i$’s **PROVISIONAL EXPECTED UTILITY** at $t$ for history $\eta_i^t$ is found by evaluating

$$\hat{u}_i(\eta_i^t) \equiv \sum_{\tau = 1}^{t} \sum_{w \in W} \Pr(w|\eta_i^\tau) \times u(\eta_i^\tau). \quad (4.4)$$

The definition states that at a given period $t$, the provisional expected utility for a particular history is the sum of the expected utilities for each period up to and including $t$. But note that the definition of expected utilities is altered somewhat, so that the probability of the plot-world, given the history, acts as a weight on the utility. Similarly then to expected utilities, felicitous actions in more probable worlds yield greater utilities than those in less likely worlds.

The probability of the plot-world has been selected as the weight factor rather than the probability of any given state for a number of reasons.
(1) The state of the scene shifts very frequently, and so metrics based solely on the state of the scene will tend to be less reliably consistent. There is good evidence that it is in fact the consistency of the plot-world that actors are most concerned with; I detail such evidence and discuss implications of this fact in chapter 5. Recall that it is a basic fact of game theory that a given action is felicitous only in certain worlds. This is also true of plot-worlds, where only certain actions may be plausible according to the facts that have been established about the plot-world. This notion may be extended to histories, too: only certain histories are felicitous in a given plot-world. This means that the plot-world is a better analogy to worlds in standard games of incomplete information.

(2) The procession of states provides evidence for the plot-world, the inverse of which is not true. As such, the cumulative effect of the procession of states of the scene is taken more into account if we use the plot-world as the primary weighting factor. Observe that this definition is the formal expression of the notion discussed above: a series of states of the scene discount certain potential plot-worlds as implausible candidates for the true plot-world, and simultaneously promote others as more likely candidates. This definition includes this notion as a valid, powerful inferential tool for improvising actors.

It is useful to have this system of constant evaluation of an actor’s performance because it can help account for why unsuccessful strategies tend not to be reproduced repeatedly throughout a plot. If an actor has no means of reflecting on her decisions at each juncture, then she is in effect irrational, for she cannot respond to her adversaries’ actions with the intention to bring out her preferred outcomes. To be clear, the provisional expected utility is entirely divorced from the reality of the situation, for players all lack significant information. The utility that ends up being assigned at the end of the game may be utterly different from the provisional expected utility that is calculated.

4.3.2.4 · The Sender and Receiver as roles, and empty utterances

The final issue that arises with iterated improvised performance games is that the roles of the Sender and the Receiver cannot remain statically attached to a single player. A single actor does not spend the entire performance speaking, and so it is inappropriate to model a situation where a single agent retains the role of the Sender throughout the entire performance. To solve this problem, I propose the simplifying assumption that interlocutors in improvised theatre cannot produce adjacent utterances, but instead must take turns to produce and interpret utterances. Most often we are interested in situations with two speakers, for in the dataset this is the number of actors that generally occupy the stage at any given time. For two players $i$ and $j$ of an iterated IPG, $i$ is the Sender at every odd-numbered period, and $j$ is the Sender at every even-numbered period.

This solution produces good results in most cases, but there are cases where an actor $i$ makes $j$ an offer, and $j$ subsequently fails to respond to the offer. For instance, $i$ may
pause, indicating that an offer has been made, but $j$ does not recognise it as such, and so
$i$ has to carry on herself. This may be a mistaken construal of $i$’s intention, but it also
may be an intentional strategy of $j$ to indicate that she cannot think of any response that
she believes will be felicitous in that given situation. Instead of stipulating exceptions
to the assumption of alternating player roles, we can solve this problem by including a
special empty utterance in the Sender’s strategy set. The empty utterance is a deliberate
strategy of silence, and may be thought of as an ‘utterance with no words’. It is in this
respect analogous to the empty string of computer science, and hence I use $\varepsilon$ to denote it;
see, for instance, Sipser (2006) for the use of $\varepsilon$ to denote the empty string.

$\varepsilon$ is distinct from not making a strategy selection, which is illegal in game theory, in
that it may be interpreted. As I have indicated, if an actor is silent when she is made an
offer, she appears to most often mean to indicate that she does not have a valid response.
There are other possible functions of a strategic silence: it could indicate that the actor who
employs it urges the actor making the offer to continue on a plot trajectory, or it may be
used to characterise a role in a way that is consistent with her traits, for instance. Common
to all of these functions is the notion that the selection of silence is a deliberate one,
and is performed with the same degree of rationality as any other strategy. The interpretation
of $\varepsilon$ is precisely equivalent to the interpretation of any other utterance, for $\varepsilon$ is simply a
member of $\mu$.

4.3.2.5 · Conclusion

With these adjustments, we are able to model an improvised performance in its entirety
as an iterated game, and I do so in detail in chapter 8, after some refinements are made in
chapters 5, 6, and 7. In general we assume that the game is infinitely iterated, in the sense
established above where $T$ is indeterminate. This decision is made so that the players must
behave as if the game would continue forever, even though in reality scenes always end.
Furthermore, since the end of the scene comes about primarily through the intervention
of actors who are not taking part in the scene, it is best to model scenes in improvised
performances in such a fashion.

4.4 · Conclusion and a caveat about concealment and utilities

This chapter has introduced a game theoretic model of improvised theatre, which concen-
trates most notably on actors’ use of language. The formal definition of a game theoretic
model of improvisation was commenced in §4.1, where definitions were given for the vari-
ous facets of improvised performance. In §4.2, adjustments were made to standard DRT in
order that it better represent the speech of improvised theatre. In particular, the separation
of actors and their characters, so that only the latter are in the universe of discourse, per-
mits DRSs to be constructed that are more accurate depictions of the speech of improvised
theatre. These definitions led to the core of the chapter, namely the definition of IPGs in
§4.3. These games, which are modified interpretation games, incorporate the definitions
of entities in improvised theatre, and also the adjusted DRT that was introduced in this
chapter. This chapter also defined the means by which entire plots might be formulated as an iterated series of IPGs, thus completing an almost-complete formal specification of a game theoretic model of improvised theatre.

A caveat applies to the model that I have sketched here. While this model captures most aspects of improvised performance, and may predict more simplistic games, it will be unable to formally account for the more complicated actions and motives that are commonly performed and attributable to actors. The major deficiency in the model is that it cannot account for covert communication that the actors use to plan and to discuss the narrative structure. Such communication is mentioned repeatedly in the computer scientific literature on improvised theatre that I have discussed throughout this chapter and the last (see, for instance, Magerko et al. 2009, Baumer & Magerko 2009, 2010, Riedl & Young 2010, Fuller & Magerko 2010, Bruce et al. 2000). Any valid model of the use of language in improvised theatre must account for this important component of the communication that is typical of the genre. I have omitted it from the model described in this chapter because the formal apparatus for concealing information within natural language utterances is highly complex, and requires a great deal of background, and a great deal more illumination of the operation of the elements of the formal model that have been sketched in this chapter.

A consequence of this lack of concealed information is that, as it stands, there is no satisfactory account of utility in improvised theatre. While I have demonstrated the use of a Lewis utility function in IPGs and their iterated forms, this is an imperfect measure of utility. What actors consider good improvisation appears to involve more than just the optimally efficient transfer of information. For improvised performances should entertain an audience in the broadest sense. The plot may be comedic, or it may be tragic and thought-provoking, but in any case it must be interesting. The problem of assigning a utility based on these considerations becomes immediately apparent. If we take interest as a measure, clearly the biases of the analyst emerge. If we take a more data-based approach, and listen to, for instance, the intensity measured in decibels of the audience’s laughter, we get a sense of the success of a joke, but this does not necessarily tell us anything about those scenes that are more serious. The actors’ reactions do not tend to reveal a great deal about the success of the plot, either, since they are always trying to preserve the verisimilitude of the plot.

I delay finding a precise measure of utility until chapter 7, since it is in chapter 7 that I provide a formal apparatus for covert communication, and so endow the model with the capacity to deal with it. Instead, for the next two chapters, the measure of utility is simply an ordering of preferences over the plot, which is arbitrarily determined by what the actor considers the best direction for the plot at any given time. The utility of a strategy is determined by its efficacy in performing the narrative function that it appears to be aimed at. This is a perfectly valid measure of utility, since utility is merely an ordering of preferences; but I will posit a more satisfying definition in chapter 7.
Chapter 5

Plot-world generation and consistency

He himself was a believer; he affirmed the miracle of translation—the near-sacred moment in which the miniature artefacts of the layout no longer merely represented Earth but became Earth. And he and the others, joined together in the fusion of doll-inhabitation by means of the Can-D, were transported outside of time and local space. Many of the colonists were as yet unbelievers; to them the layouts were merely symbols of a world which none of them could any longer experience. But, one by one, the unbelievers came around.

Philip K. Dick, *The Three Stigmata of Palmer Eldritch*

The previous chapters have devised a formal framework for a rigorous analysis of improvised theatre. In this and the following two chapters, I examine how this framework applies to data of improvised theatre, and explore the ways that the framework might be improved. I do so by focussing on specific aspects of improvised performance. In chapter 6, I look at how the common ground of a performance is built through negotiation, and in chapter 7, I analyse the methods that actors use to conceal information in their utterances. In this chapter, I examine strategies that actors use to create and sustain plot-worlds.

Plot-worlds are a good place to begin an empirical analysis, for they are the strongest element that directly affects any decision that the actor might make. As I have suggested in the previous chapter, plot-worlds act to constrain what is possible by requiring internal consistency. This means that when a fact \( A \) is established, \( \neg A \) may not be simply introduced, for it would violate the internal laws of the plot-world. But plot-worlds also have the opposite effect. Since a given plot-world is the setting for any given scene, and since the laws of the plot-world are not constrained by those of the actual world, the possible scenes that actors can perform may differ greatly from those in the actual world. For instance, characters may be animals that speak English, which is patently impossible in the actual world.
Any given plot-world is treated as a Stalnakerian possible world (see chapter 2), for it appears that this is how both actors and audience members interact with the worlds that the actors build. That is, a plot-world has the following characteristics: (a) independence from the actual world, meaning that the plot-world is fully specified and not merely a nonlocal state of the actual world, regardless of how closely the plot-world resembles the actual world; (b) consistency, insofar as the plot-world has a complete set of laws that are noncontradictory; and (c) causal relation, insofar as the audience’s feedback to the actors has some effect on their actions, and thereby on the constitution of the plot-world. For these reasons, it appears that a plot-world possesses the same ontological status as any possible world that is able to be included within a proposition.

However, there is something problematic in stating that a plot-world has the same ontological status as the actual world, since there seems to be something fundamentally unreal about plot-worlds that is not necessarily attributed to other possible worlds. To tease out this difference, let us suppose that you and I, inhabitants of the actual world, wager on the outcome of a football game on the weekend.¹ There are three possible worlds that are salient to the bet: the possible world in which the team that I bet on wins, the possible world in which the team that you bet on wins, and the possible world in which there is a drawn game.² These possible worlds are not true with respect to the known facts of the actual world at the time that the bet is made, in the sense that they are merely contingencies, yet to be made manifest. In this respect, they are like plot-worlds before a performance: possibilities yet to be realised. But there is an extremely important difference: the agents in the three worlds that are salient to the wager are consistent with the actual world—the teams we bet on will be the teams that play, and either you or I will be around to collect the winnings of the wager—, and the contingencies are viable continuations of the timeline of the world—that is, the event of the wager precedes the event of the football match. So, the football match occurs in the same possible world in which the wager is made.

On the other hand, plot-worlds are neither consistent with the actual world, nor a continuation of it. For instance, the characters in plot-worlds do not exist independently of the actors who play them, the traits that the characters exhibit are chosen by actors. Even more strangely, the decisions that characters apparently make are in fact made by actors, who do not strictly exist in the plot-world! This discontinuity, which is a fundamental property of plot-worlds, means that they are ontologically distinct from possible worlds that might be considered as potential candidates for the actual world.

Yet, plot-worlds are treated by actors and audience alike as fully-realised possible worlds. Despite the serious inconsistencies that plot-worlds exhibit when compared with the actual world, such as those that I have pointed out with respect to characters and ac-

¹Vladimir and Estragon are not involved in this example because they are invented characters. You and I are (supposedly) real.
²If we are pedantic, there are of course the possible worlds in which some cataclysmic event causes the cancellation of the match. But let us not be pedantic. We shall assume that these are not salient for the present example.
tors, actors continue to endow plot-worlds with the properties of internal continuity and internal consistency that are generally ascribed to Stalnakerian possible worlds.

This chapter deals with how actors create this continuity and this consistency, and the optimal strategies for doing so. This chapter also sees the first application of the formal apparatuses described in the previous chapter. In §5.1, I describe the dataset that is used, and I introduce the Conversation Analysis-style method of transcription that I use for the presentation of data. In §5.2, I examine the methods that actors use to introduce information in the plot-world. The analysis deals with the employment of such methods at various points of the scene. §5.3 continues this discussion by analysing how information is accessed after it is established. Specifically, I explore how plot-worlds that are constructed in a previous scene are accessed so that the existents of the previously constructed plot-world may be reused. In §5.4, I describe how this process of resumption of a plot-world can fail, and how actors tend to behave when it does. Finally, in §5.5, I propose a number of heuristics that describe how actors expect their fellows to behave. This approach is taken instead of an approach proposing a single solution concept for action in improvised theatre. The heuristics act similarly to the Gricean maxims as employed by, for instance, Benz & van Rooij (2007): that is, as formal constraints on interlocutor behaviour, but not strictly as solution concepts. The chapter concludes in §5.6.

5.1 · The data used in this study

For the discussions in this and the three following chapters, I use a dataset of recordings of an improvised performance which took place in November 2014, in Melbourne, Australia. The theatre troupe that was recorded was selected because they were, at the time of recording, the only group in Australia that performed Harold longform.

Four performances were captured, each lasting 25–30 minutes; around two hours of footage were recorded. Each performance had its own group of actors. When presenting data, I designate each scene by its usual Harold denotation. The numbering for a scene is ‘p-s’, where p is the ordinal number of the performance, and s is the scene, as was discussed in §2.2.2.4. The location of the performances was the back room of a pub, and therefore the actors had limited access to physical props: the only props that the actors could use were a number of barstools that sat beside the performance area. During the performances, a pianist improvised accompaniment, and while the music certainly had some effect on the decisions that the actors made, I do not include music in this formulation so as to keep the model manageable.

Representing improvised theatrical performance is problematic, since factors such as the position of the actors, their gaze, use of gestures, and so on, are often important to understanding the action of a performance. Since I am primarily concerned with actors’ linguistic strategies, I use a particular style of markup used in Conversation Analysis, which is used commonly for linguistic interactions (henceforth CA; see Hutchby & Wooffitt 2008 for an overview). Datum 1 overleaf gives an example from the beginning of a scene, which is a typical example of an excerpt of a scene.
Datum 1 Example of a CA-style transcript, from scene 4-B1.

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>that MIND of yours (0.8) that mind (0.4) that brain (0.8)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>hhh (1.9) it’s beautiful (2.9) wish i could hhhhh (1.0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>take it out of that skull and just (1.3) just lightly (1.1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gently &gt;just&lt; (1.3) caress you’re just so smart (1.3)</td>
</tr>
<tr>
<td>2</td>
<td>B</td>
<td>i’m just a truth-teller</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>[what:]</td>
</tr>
<tr>
<td>4</td>
<td>B</td>
<td>[stephen]</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>(1.3)</td>
</tr>
<tr>
<td>6</td>
<td>A</td>
<td>wo:w</td>
</tr>
</tbody>
</table>

As we can see, a CA transcript is presented in the same fashion that a theatre script generally is, except that there are three columns: the first gives a line number used to index the utterance, the second gives the speaker of the utterance, and the third gives the utterance itself. Characters of the performance are represented with capital letters in a sans-serif typeface (e.g., A, B, etc.); the audience is always denoted P. This typographical convention is present not only in the transcript, but also extends to references to actors in the body of the text. Note that, even if a single index is used in two scenes, one should not assume that they refer to the same character. It will be made clear when they do.

A number of symbols are used in CA transcripts. The markup scheme that is used in this thesis is based on that used by Hutchby & Wooffitt (2008), and is summarised in table 5.1 overleaf.

5.2 Information creation in a plot-world

Let us begin with a discussion of how actors introduce information about the plot-world. Consider for the moment the problem that the actor is faced with at the beginning of a scene. If the performance is a purely improvised one, none of the usual elements which form the basis of a scene-state (existents, previous actions, and so on; see §4.1) have been established, so the plot has no discernible trajectory. The major problem for the game theoretically rational agent that arises from this type of scenario is that there is no indication of what an optimal strategy might be, for this situation is, in essence, one in which no actor has access to the true state of the plot-world. The set of available actions is virtually unconstrained, and so an actor has no prompt to tell her which possible world is that of the plot-world, and no information to lead her to form hypotheses about the preferences of her fellow actors. In linguistic terms, the problem becomes especially salient: no common ground exists within the plot-world according to which actors may select their interpretations. This problem is unique to the beginnings of scenes, for the space of possibilities narrows as more entities are introduced, or as the location of the scene becomes more completely realised.

3P is selected as the denotation of the audience based on the Italian pubblico (audience). This is preserved as acknowledgement of the beginnings of this thesis project, which was a game theoretic description of the commedia dell’arte, a genre of improvised theatre that was popular in late-Renaissance Italy.
symbol | meaning
-------|---------------
(x)    | a pause of x seconds; for instance, (1.8) indicates a pause of 1.8 seconds
(.)   | an extremely short but perceptible pause
((action)) | a description of an extralinguistic action performed by an actor
( )   | blank space between parentheses indicates that what is said is audible enough to understand who uttered it, but not clear enough to understand what was said
(word) | a word in parentheses indicates that what was said is unclear, and the word is the transcriber’s best approximation of what was said
underline | primary stress on a syllable
CAPITALS | emphasis on a syllable that is greater than primary stress
:      | lengthening of the sound preceding the symbol
square brackets | vertical adjacency of two strings of lines in square brackets indicates that they are produced simultaneously, as in lines 3 and 4 of datum 1
=      | indicates that there is no pause between one line and another; may also represent an interruption
>word< | the word is said more rapidly than normal
<word> | the word is said more slowly than normal
°word° | the word is said more quietly than normal
xxxx | laughter of the audience as a whole; capitals indicate louder laughter
heh hah | close transcription of the laughter of an individual member of the audience
↑ and ↓ | noticeable upwards and downwards inflections
£      | indicates that the speaker is smiling, or laughing softly while saying the word that follows the symbol
#      | applause
::     | indicates omitted lines in the datum

Table 5.1: Summary of transcription symbols used in this thesis.

By far the commonest method of overcoming this challenge is for actors to assert the verisimilitude of the plot-world. That is, the actors behave as if the world is fully realised: that it has a complete history; that there are internal laws which govern the behaviour of the entities that exist within it; that the characters themselves have comprehensive histories and motivations, and have chosen actions that have led them to the situation that they find themselves in when the scene begins; and so on. Consider datum 2 overleaf, which exhibits a typical means of establishing the setting of a scene. What is interesting in this datum is that the aeroplane is not indicated directly, and instead the occupations of both characters are able to be inferred by their actions. There is no literal description in declarative sentences of what inhabits the scene. Instead, the audience is left to deduce the traits of the characters, given the performance of the characters’ actions.

This naturalistic quality of speech in improvised theatre is a deliberate device to assert the plausibility of the plot-world. That is, two people acquainted with one another, and who are in the same physical space, do not state their occupation or the contents of that physical space, for without prompt these things are mutually taken to be in the common ground. Each character in datum 2 would know that the other is a pilot, and that they
Datum 2 The beginning of 2-C1. The characters establish the setting of the scene, an aeroplane cockpit, without literally describing it.

1 A: (mimes taking hold of what appears to be a steering wheel) ssssshhhhhhhhhh ((i.e., the sound of an aeroplane))
2 ((B enters, mimes pushing some buttons on the vehicle’s console))
3 (4.8)
4 A: flight seven three seven (0.1) off the ground (0.6) successful takeoff

are piloting an aircraft at that moment, and so this fact would not normally be explicitly stated in the actual world. A’s utterance in line 4 is conceivably an appropriate sentence for a pilot to utter, and taken together with his miming gesture, the audience and the other actor infer the character’s occupation.

From observation of the data, the plausibility of any information that is introduced into the plot-world appears to be of great concern to actors in improvised theatre. I claim that this tendency arises from the fact that the genre of improvised theatre is distinguished amongst genres of naturally occurring speech in that spectators are consciously aware that they are viewing fictional events. That is, the audience expects plot-worlds to differ from the actual world, and so they are willing to accept the truth of certain assertions, even if these assertions contradict what the actors or the audience believe about the actual world. Actors can with some legitimacy make any claim about the plot world (within reason, as I show in §5.4).

In this section, I offer analyses of two linguistic strategies that actors use when creating a world: presupposition in §5.2.1; and sociolinguistic variation in §5.2.2, which includes the use of accent, and word choice and grammar.

5.2.1 · Presupposition

A very subtle means of asserting the world’s persistence—that the world exists not merely for the duration of the scene, but has a complete history—is the deliberate use of presupposition triggers. Presupposition is used prolifically by actors, but it is in the current circumstance of establishing a plot-world that it is used in the most surprising and counterintuitive ways.

Presuppositions arise from utterances when knowledge of some fact that is necessary for the felicitous interpretation of the utterance is taken for granted, or assumed to be common knowledge (Beaver & Geurts 2014, Geurts 1999). For instance, the sentences in (45) presuppose that there is a banana that is available to Vladimir so that he may eat it, and they do so in different ways.

(45) a. Vladimir eats the banana.
   b. It is the banana that Vladimir eats.
Knowledge of the existence of the banana appears to be taken for granted by the individuals who produce these sentences, and this is expressed by the use of the definite article. Sentences such as (45) are called presupposition triggers (Beaver & Geurts 2014), a term which entails that presuppositions are inherent to the utterance. Each of the sentences in (45) is a presupposition trigger, for each requires a fact to be known or assumed, namely the existence of the banana, to felicitously interpret the sentence.

Interlocutors accommodate presuppositions that are introduced in the course of a conversation. Lewis (1979: 393) provides the following broad definition of accommodation, called the rule of accommodation for presupposition:

If at time $t$ something is said that requires presupposition $P$ to be acceptable, and if $P$ is not presupposed just before $t$, then – ceteris paribus and within certain limits – presupposition $P$ comes into existence at $t$.

However, what Lewis meant by ‘certain limits’ is vague, and so are the mechanisms which allow for accommodation, and the rules which govern it. To fill this gap, von Fintel (2008: 163) proposes that presuppositions are encoded into the semantics and syntax of sentences, and he posits a ‘bridging’ principle which “turns semantic presuppositions into pre-requirements imposed on the context to be updated by the sentence.” The bridging principle stipulates that a proposition as presupposition may only update a context, or a “body of knowledge”, if it assigns a truth value to all possible worlds compatible with that context. This notion of presupposition adequately accounts for occasions in which what an interlocutor takes for granted precludes the possibility of accommodation. Such instances are called presupposition failure (see Strawson 1964, van Eijck 1994). Suppose, for instance, that the hearer of either of the sentences in (45) knows that there is no banana in Vladimir’s possession. In this case, the knowledge of the hearer interferes with what would be a simple case of accommodation if the hearer did not have access to this knowledge.

Vander Sandt (1992) proposes a means of dealing with presupposition in DRT. The idea is simple: presupposition triggers are anaphoric expressions, and so bind to entities in the discourse in precisely the same way that pronouns do. Consider a DRS $K$. A presupposition that predicates $P$ of an entity $x$ is felicitous iff $x \in C_K$. Otherwise, it fails.

Presuppositions thus give actors ways of asserting the existence of existents in the plot-world. Since it relies on entities and conditions existing within a DRS, and since that DRS’s truth relies on these entities and conditions holding in a possible world, a presupposition gives the appearance that the utterance is part of an existing context. At the beginning of a scene, this may not be the case: no facts have yet been established of the plot-world. Nevertheless, actors generally wish to assert the persistence of the plot-world. Therefore, presupposition is used by actors to give the plot-world a keener sense of persistence.

The problem is this: since presuppositions rely on the resolution of anaphora, and since the actors and audience iteratively discover facts about the plot-world together, there is, in many cases, no way for the anaphoric expression to bind to an entity in the DRS. From
these kinds of situations, we would normally expect that the presuppositions fail. Yet, they are accommodated. How is this achieved? In the subsections that follow, I examine two ways that actors trigger presuppositions and presupposition accommodation in seemingly difficult environments. In §5.2.1.1, I suggest that actors use their characters’ world-knowledge; and in §5.2.1.2, I describe how actors use framing effects. In §5.2.1.3, I compare these devices.

5.2.1.1 · Exploiting characters’ world-knowledge

It is common for actors to select utterances that trigger presuppositions that the characters can supposedly accommodate with their knowledge of the plot-world, but that the audience cannot. This gives the impression that the characters have some encyclopaedic knowledge of the plot-world that the audience does not. The very existence of this knowledge suggests that the plot-world’s existence prefigures the scene. The audience, then, is likely to treat the presupposed fact as just another fact that they become aware of regarding the plot-world.

To illustrate this, let us consider the beginning of 1-A1, which constitutes datum 3. The ‘object’ that is referenced in line 6 is unidentified: A’s miming does not provide sufficient information to give a fixed identity to the object, but is merely the action of unpacking something. We learn only later in the scene that the object is a typewriter.

<table>
<thead>
<tr>
<th>Datum 3 Beginning of 1-A1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1  A: well they fired me</td>
</tr>
<tr>
<td>2  B: why’d you show up to work then dad</td>
</tr>
<tr>
<td>3  P: xxx</td>
</tr>
<tr>
<td>4  B: why’d you show up to work then dad</td>
</tr>
<tr>
<td>5  P: [x ]</td>
</tr>
<tr>
<td>6  A: [i’ve got (.) i’ve got nowhere else to go do i↓ (4.2) ((begins unpacking an object)) been here for (1.5) forty years (1.8) this is all i know</td>
</tr>
</tbody>
</table>

If we consider only line 1 in isolation, the DRS in (46) is generated.

$\text{(46)} \quad K_1 = \begin{bmatrix} i, x \\ \text{fired}(x, i) \end{bmatrix}$

Recall from §3.2.2 that $i$ is a special, deictic referent that binds to the producer of the utterance, and therefore that $i$ is automatically bound, and the anaphoric expression $\text{me}$ is resolved. Also recall that, per definition 3.7 in §3.2.3, a DRS is proper if it does not have any free variables. So, (46) is improper, for there are no definite referents in the DRS’s universe to which the pronoun $\text{they}$ may resolve. The utterance should fail to be interpreted. Yet, there is nothing intuitively wrong with the sentence, and it appears that it is interpreted without problem by the actors and the audience.

I propose that two phenomena operate in concert to allow this sentence to be felicitously interpreted. The first is that, since the entity which resolves the anaphora must
have the ability to fire A, the domain is restricted to employers. Typically, an individual has only one employer, so the term fire places a selectional restriction on the pronoun they.\footnote{Recent work by Magidor (2009, 2013, 2017) explores the relationship between selectional restriction and presupposition in greater depth.} Note that the use of the perfective aspect strengthens this selectional restriction. Since this is the first utterance of the scene, and more importantly the first established fact of the plot-world, the verb indicates that the action of firing has taken place prior to the commencement of the scene.

The second is more specific to improvised theatre. Since no objection is raised by B, it appears that the resolving entity for they is known to both characters. That is, we are able to claim, on the basis of the evidence the actors provide, that there is some common ground DRS which is proper, and so does resolve the anaphora. That is, the bound entity has been established in the common ground prior to the actual scene.

Used together, these linguistic strategies suggest that the plot-world itself exists prior to the conversation that the audience views, and therefore that there are events that have occurred in the plot-world that the audience is unaware of. This means that the audience can recognise this fact as one among the many facts of the plot-world of which they are unaware. They are provided with evidence that the two players are able to resolve the anaphora, and they therefore ‘learn’ of the existence of an entity to which the anaphoric expression is able to bind. It appears that this results in the audience accommodating an expression which, given the information to which it has access, it should not be able to accommodate.

This observation implies that the actors appear to intentionally try to bring about this accommodation, through various means. Most saliently, neither challenges the assertion of the fact, and both act as if the anaphora resolves.

The second utterance of the datum exhibits a similar case. B’s use of the verb to turn up—which in Australian English is felicitous only in situations in which the subject arrives unexpectedly—and of the noun work without a determiner presuppose that (a) A and B share (or shared) a workplace, and (b) both characters are located in that workplace at the time when the utterance is produced. And like the example above, the stereotypical uniqueness of a person’s workplace allows the audience to accommodate the presuppositions that are triggered by this utterance. We can predict, based on this fact, that, if B were uncertain whether the audience would be able to accommodate this fact, he might have used the possessive, and produced something akin to (47).

(47) Why’d you turn up to our work, then, dad?

In this case, the shared nature of the workplace would have been more explicitly conveyed, but perhaps unnecessarily so. Besides, the use of our is very stilted: both characters should be presumed to know that they share a workplace, and so using our would sound odd.

This exemplifies a way of asserting that the plot-world is persistent, and, perhaps to a lesser extent, consistent. It involves the actions of (at least) two actors: (1) an actor pro-
duces an utterance which triggers some presupposition; and (2) her fellow does nothing to contradict the presupposition: she does not assert that there is not an entity in the apparent common ground DRS that can resolve the anaphora. Since the resultant common ground arises only through the cooperative action of multiple actors, the common ground is co-constituted. This is an important concept to bear in mind, for it informs the discussions in chapter 6. Sequences of this kind are quite common in the data, especially at the beginning of scenes.

5.2.1.2 · Filling in gaps using prototypical frames

In datum 3 above, the domain of the entities that are able to felicitously resolve the anaphoric expressions is somewhat constrained, and it is the relative poverty of possible resolutions that allows for their existence to be accommodated by the audience. Instances abound in improvised theatre, meanwhile, where the utterance by itself does not provide the audience with the sufficient information to accommodate it, and so recourse is made to certain prototypical frames. A frame here is understood in the sense proffered by Fillmore (1976, 1977, 1976), as “a system of categories structured in accordance with some motivating context”, where “the motivating context is some body of understandings, some pattern of practices, or some history of social institutions, against which we find intelligible the creation of a particular category” (Fillmore 1982: 119). A frame is a component of cognition that relates encyclopaedic knowledge to a specific context, and describes the way that certain entities and events are perceived to be related to that context. In specific frames, certain resolutions for anaphoric expressions become more salient, for frames enforce selectional restrictions on entities that are permitted to resolve certain anaphoric expressions within that frame. And so actors can use prototypical frames to prompt certain resolutions of an anaphora.

To see this kind of device in action, let us consider datum 4.

**Datum 4** Beginning of 2-A1.

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A:</td>
<td>((mimes moving a frying pan on a stove, speaks in a stereotypically Mexican accent)) eh compadre: (...) ah they not coming in any more</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>(0.6)</td>
</tr>
<tr>
<td>3</td>
<td>P:</td>
<td>XXXxx</td>
</tr>
<tr>
<td>4</td>
<td>B:</td>
<td>((speaks in a stereotypically Mexican accent)) people that come here because they say you make uh the good sauce and you make uh the bullshit sauce they don’t COME</td>
</tr>
</tbody>
</table>

The DRS of the utterance in line 1 is as follows, with the caveat that, as motivated in §3.2.5, I do not provide a formal description of tense and aspect.

\[
\begin{array}{c}
  \exists x \\
  \neg \text{enter}(x)
\end{array}
\]
5.2 · Information creation in a plot-world

Like line 1 of datum 3, the utterance contains an anaphoric expression that is unbound: in the present case, it is *they*. The DRS in (48) is improper. And, as in datum 3, the utterance is still interpretable by the audience, since it relies on the resolution of anaphora that is possible only if the Receiver of the utterance presumes that information precedes the commencement of the scene.

But unlike in datum 3, more information is used by the audience to accommodate the presupposition than merely the presumed, prior knowledge of the characters. The actions of A, especially the mimed movement of a frying pan on a stove, suggest that the characters are in a kitchen. This fact alone does not suggest what kind of kitchen the characters are in. It may be that the kitchen is in a private household, for instance. In the production of the utterance, A’s use of a stereotypically Mexican accent, and the use of the hispanism *compadre*, further constrains the possible locations of the scene, by suggesting that it is more probable that the action takes place in a Mexican establishment. With these pieces of information, the setting of the scene is most probably a Mexican restaurant, and in fact it is the use of *they* and the verb *coming in* that makes this the most plausible resolution given the selectional restrictions of the restaurant frame.

The provisional establishment\(^5\) of the scene’s location as a restaurant sets up an interpretation of *they* as patrons of the restaurant. This follows from the prototypical frame of a restaurant: if people come in to a restaurant, then they are most likely to be patrons of that restaurant. It is certainly possible that an individual who enters the restaurant be an employee of the restaurant, but since restaurants stereotypically have fewer employees than patrons, it is more probable that the individual is a patron. It therefore appears that the Receivers of this utterance use a very simple entailment to bind the anaphora to a nonspecific referent: that is, it is the case that if \(x\) enters a restaurant and \(x\) does not work at the restaurant, then \(x\) is a patron of the restaurant. This entailment can be incorporated explicitly into the common ground, as in (49).

\[
\begin{array}{c}
\neg \text{enter}(x) \\
\text{enter}(x) & \Rightarrow & \text{patron}(x)
\end{array}
\]

Thus, the anaphoric expression may be resolved.

5.2.1.3 · Sufficient information

There are two key differences between the use of *they* in datum 4 and the use of *they* in datum 3. While both exhibit selectional restrictions imposed by the respective frames, the

\(^5\)It is important to note that this setting is not established only by A’s utterance. While it is most probable that the scene takes place in a Mexican restaurant, it is by no means certain. B’s response appears to confirm that she believes that this is the case, however, as she refers to people entering the location to try A’s formerly “good sauce”. Such co-constitution of the scene and of common ground is the topic of chapter 6.
means of specifying entities that resolve the anaphoric expression differs. This difference is primarily located in the specificity of the respective referents.

Consider datum 4. In order to interpret A’s use of they, B selects for its binding entity a class of individuals, that is the class of patrons of the restaurant, without having specific members of that class in mind. In other words, they refers to the patrons of the restaurant, which is a nonspecific group of people with an undetermined size. Meanwhile, in datum 3, there is enough evidence to suggest that B is aware of the precise group of people that A’s use of they resolves to. That is, the set of individuals that A has in mind is a specific group of individuals, which is the group of A’s employers. Since B shares the workplace, these are known to him. The group of A’s employers may be large and complicated (as in any large organisation’s management bureaucracy), but there is a finite number of individuals who are A’s employers. Meanwhile, any individual may be a patron of the Mexican restaurant if they dine there.

These differences in the deployment of linguistic devices that force presupposition accommodation demonstrate well the degree to which actors are aware of the informational needs of their fellow actors, and of the audience. In datum 3, A does not provide any common ground information, because he correctly believes that the semantics of the verb to fire provides Receivers of the utterance with enough information to felicitously bind they to a set of individuals. A in datum 4 is aware that the utterance itself does not give enough information for that instance of they to be bound, and so he provides more information through mime and the use of an accent. This additional information grants Receivers of the utterance sufficient evidence for an inference to bind the pronoun to a class of individuals.

The need for sufficient information to accommodate presuppositions is particularly marked at the beginnings of scenes, as is the case in data 3 and 4. The common ground at the beginning of a scene is equal to ∅; there is no information that can be used by Receivers of the initial utterance when interpreting the utterance. In some scenes, such as the establishment of the setting in an aircraft cockpit in datum 2, the interpretation of the initial utterance is fairly simple, since it does not rely solely on specific information of the existents and locations in the plot-world. The use of presuppositions that are unable to be accommodated without specific knowledge of the plot-world in data 3 and 4 necessitates that the actors control the information that is available to the Receivers of the utterance more carefully. The utility of such devices appears to derive chiefly from the fact that they cause the plot-world to appear more fully-realised, and more self-contained than it might otherwise. Both uses of they refer to existents of the plot-world that are absent from the scene, which gives the impression that the plot-world extends beyond merely the local situation that is presented to the audience.

5.2.2 · Attribute ascription through sociolinguistic variation

A second linguistic device that actors employ in order to establish a plot-world, and especially in order to easily establish the traits of their characters, is to use certain soci-
5.2 · Information creation in a plot-world

Sociolinguistic features of language that are stereotypically associated with classes of people. Examples of features that are frequently employed are accent, word choice, and grammatical variation. These together will be termed sociolinguistic variation for the purposes of this study. We have already encountered an example of a device which conforms to this description. In datum 4, the stereotypically Mexican accent while speaking English was used to convey that the restaurant is a Mexican restaurant.

In §5.2.2.1, I examine the use of accent; in §5.2.2.2, I examine word choice and grammatical variation. The incorporation of these sociolinguistic devices into a DR theoretic model of improvised theatre is discussed in §5.2.2.3.

5.2.2.1 · Accent

There are two major types of accent that the actors in the data set tend to select. These are, unsurprisingly given that the performance took place in Melbourne, Australian accents and non-Australian accents. Some actors had accents other than English (including other Englishes than Australian English), and so what I refer to is the use of accent such that an actor alters her native accent to achieve some effect, not merely an instance of a particular accent per se. For instance, that one of the actors has a US accent is not of interest to this discussion, but the scenes where that actor exhibits a more rural US accent are.

Non-Australian English accents. The non-Australian English accents that are most often employed are US accents, and Mexican and British English accents each occur in one scene. Accents are produced in such a way that certain features are most distinct. Namely, those features that are most idiomatic of another accent for a speaker of Australian English. For instance, rhoticity and prosody are the most distinct features of the US accents that the actors employ. In particular, a number of scenes feature Southern US accents, which are marked by their slowness, extremity of pitch variation, and diphthongs that reduce to the schwa (ə). These are all stereotypical features of Southern US dialects (Thomas 2004). When using British English accents, it is most common for actors to use either accents that converge to Received Pronunciation (RP) and that are stereotypically associated with upper-class, posh individuals (Roach 2004) or accents approximating Cockney English, which is usually associated with working-class individuals (Giles & Sassoon 1983). Other British English accents, such as Scottish accents, Northern English accents, and Western English accents, are absent from the dataset. For accents close to RP, the features that stand out most are deliberateness of speech which manifests itself in, for instance, the use of /t/ and /d/, instead of the flapped [ɾ], in a conspicuous lack of h-dropping, and in the use of the vowel /ɑː/ in words like dance, instead of the /æ/ common in certain dialects of Australian English (Cruttenden 2014). These are contrasted with those features of Cockney English most obvious to speakers of Australian English: th-fronting, where /θ/ becomes /f/ in all

---

6. Recent work by Burnett (2017) also deals with sociolinguistic variation and its role in identity formation using game theoretic linguistics.

7. Note that accent here does not mean prosodic stress, but instead refers to the features of a subgroup’s phonetic realisation of the phonemes of the language.
environments, and /ð/ becomes /v/ in most environments; h-dropping; and t-glottalisation, where the glottal stop (ʔ) is used in place of /t/ in various positions, and in place of /p/ and /k/ in word-medial positions (Sivertsen 1960), though the latter feature is more likely to be used in the production of more obviously stereotypical accents. Similarly, in the instance of Mexican English, the use of dental consonants is far more widespread than Australian English, and the /r/ phoneme is realised as a trilled [r] or a tap [ɾ] rather than standard Australian English postvelar approximate [ɹ]. For instance, in line 1 of datum 4 above, reproduced as (50), the /r/ in both compadre and more is realised as [ɾ].

\[ ((\text{mimes moving a frying pan on a stove, speaks in a stereotypically Mexican accent})) \text{eh compadre: (.) ah they not coming in any more}) \]

Non-Australian English accents serve as shortcuts to a particular characterisation. They do so by predicking certain stereotypical characteristics of the characters. Why such devices might be preferred is clear when one considers that characterisation in this way is extremely efficient: without having to explicitly predicate anything, details of the character’s education, class, and details of the location may be communicated between the actors and to the audience. Since the actors are presumably interested most strongly in only the most general characterisations being communicated, they use the most typical, noticeable features of those accents. The cases detailed above are only a few features of the dialects that have been mentioned, but they are the most obvious ones to speakers of Australian English, that is, the audience.

*Australian English accents.* The use of Australian accents is much more nuanced, since the audience is expected to have a far greater cultural sensitivity to slight changes in Australian accents. For instance, the production of prelateral /e/ as /æ/, despite being a relatively common phonological phenomenon in accents across Australia, is “popularly and frequently identified as being typical of only speakers from Melbourne and Victoria” by speakers of Australian English (Loakes, Hajek & Fletcher 2009: 2). Because this is associated with the stereotypical Melburnian, it would be used for a character who is a stereotypical Melburnian.

Thus, the information that is predicated by Australian accents is, for the most part, more fine-grained, and the stereotypes that are accessed by accents more localised. The information that is intended to be conveyed to the audience of (presumably mostly) Australian English speakers is far richer than is possible with other accents. For this reason, in datum 3, A’s accent when producing his first utterance communicates to the audience not only that the character is working class, but also that he is middle-aged, for his accent exhibits features that are stereotypical of middle-aged anglo-Australians. The features that most clearly indicates that A’s character is middle-aged are: (a) a use of a combination of creakiness, breathiness, and roughness, which are characteristics that generally increase in intensity with the age of the speaker (see Verdonck-de Leeuw & Mahieu 2004); and
(b) the use of a Broad Australian accent, which is discussed below. These features are effective in communicating the age of the character precisely because they are not present in the normal accent of the actor, who has spoken before this point in the performance.

The feature of Australian accents that appears to be of great importance to the actors is the **Broadness continuum**, by which accents range “from the most overtly local form (Broad) through to that which bears some resemblance to Received Pronunciation of British English (Cultivated)” (Cox & Palethorpe 2007: 341). At the centre of this continuum is the General accent, which is the accent that most of the actors possess.

The style of the Australian accents that are used most by actors are generally those of the extremes of the broadness continuum. Broad accents used by the actors generally exhibit a characteristic of such accents, noted by Cox & Palethorpe (2007: 345), that sees the first element of diphthongs vary from their production in General accents. The most obvious of these are retraction of the first element of the /ɑe/ vowel (as in, for instance, *pie*), and fronting and raising of the first element of /æɔ/ (as in, for instance, *how*). Both of these occur in A’s speech in scene 1-A1 (as in datum 3). In line 6 of datum 3, reproduced as (51), the instances of the pronoun *I* are all produced [ɑe].

\[
\text{(51)} \quad \text{A: [i'v\text{e}] got (.)} \text{ i've got nowhere else to go do i↓ (4.2) ((begins}\ \text{unpacking an unidentified object)) been here for (1.5) forty}\ \text{yea}\text{r}\text{s (1.8) this is all i know}
\]

According to the data, Cultivated Australian accents are interestingly dispreferred in favour of accents that are closer to RP. I propose that the reason for this preference is simply that in most cases, the upper-class character of a Cultivated accent is better emphasised to Australian English speakers through a more outrageously upper-class accent. Since RP accents are perceived as more upper-class than Cultivated Australian accents, RP accents are clearer indicators for this trait, and therefore are preferred.

**Accents to subvert expectation.** From these observations, it becomes apparent that there is broadly a difference in the narratological function of accent selection. In my dataset, non-Australian accents are used to define characters who possess a characteristic that is stereotypically associated with people who speak with that accent. An accent typical of the American ‘wild West’ is used when the scene is set in the ‘wild West’, and when the characters in the scene act as characters do in Western genre films. In datum 5 overleaf, a Southern US accent is used by naïve, uneducated characters. They are puzzled by their pregnancy that has miraculously occurred “without a university education”.

Meanwhile, Australian accents that are Broad are more often selected when the actor wants to subvert expectations about the character. Since Broad accents are used in the actual world by working class individuals who are stereotypically uneducated and unsophisticated, using a Broad accent for a character who is erudite and cultured is a device that is frequently used for comic effect. For instance, in a scene in which a janitor is clandestinely an artist (scene 4-B2), the actor who plays the janitor uses characteristics typical of
Chapter 5 - Plot-world generation and consistency

Broad accents to emphasise the juxtaposition between her presumed class and her artistic activity. Indeed, it seems that the more extreme on the Broadness continuum the accent is, the more subversive will the actions be.

Accents as strategies. Accents are used as a means of quickly establishing certain attributes of a particular character, or efficiently setting the scene in a particular location. Most often, accents are deliberately selected by the actors to indicate social categories of characters. This follows social practices in naturally-occurring conversations, in which speakers shift their accents to signal their standing with respect to their interlocutors (Giles & St Clair 1979). For instance, Bourhis & Giles (1977) note that Welsh speakers are more likely to emphasise their Welsh accent when they are speaking with an interlocutor using RP than with fellow Welsh speakers. Actors’ use of accent demonstrates an awareness of this facet of accent use. Implicit in accent selection, then, is the actor’s expectation that her fellows and the audience will have the requisite knowledge to properly associate particular accents with particular stereotypical attributes. An actor’s anticipation of the audience’s knowledge helps explain why Australian English accents are used in a more complex fashion than non-Australian accents, such as subversion of expectation in my dataset. Since the performance occurs in Australia, the audience is expected to be made up chiefly of Australian English speakers, and therefore the majority of the audience is assumed to have knowledge of the accents that are associated with particular groups of speakers in Australia. Small changes in a character’s accent allow the actor to effectively communicate nuanced differences in that character’s attributes.

Game theoretically, the use of accents as part of an utterance strategy is generally advantageous in that it is an extremely clear method of proposing a trait of a character, and of setting the scene in a particular location. That is, the conditional probability of a character being of a certain nationality is greatly increased if the character speaks with an accent typically associated with people of that nation. This means that the risk of misinterpretation is significantly reduced. The chief way that these kinds of devices are
linked to strategies in the game theoretic model is through the introduction of hermeneutic principles, called *heuristics*, discussed in §5.5.

5.2.2.2 · Word choice and grammatical variation

Word choice and grammatical variation is another means of efficiently ascribing certain attributes to a character, or for setting the location of the scene. Like accent selection this is a way of conveying stereotypical information about the character. In cases where the actor selects an Australian accent, certain phrases are used to reinforce the class connotations of the accent that has been chosen. For instance, the use of the phrase *old girl* when referring to the character’s car in line 18 of datum 6, is typical of an older, working class Australian, and so it supports the characterisation given to the character by the accent that is selected.

Datum 6 End of i-A1.

14 B: i make myself a coffee (2.1) i go back down (1.0) and i check that the car hasn't overheated
15 P: xx
16 A: and then i turn it off for a bit
17 P: XXXxx
18 A: cause the old girl can only run for a certain amount of time before she'll conk out
19 B: XXx

Likewise, in scene 3-A1, the characters are both uneducated Southern US English speakers, and as such nonstandard English grammar is employed. Example (52) is the first sentence of the scene.

(52)   A: i got mary-lou pregnant and i don't know how the hell I DONE it

The fact that the actor does not normally speak a Southern US dialect means that he deliberately selected a verb form that is nonstandard for him.

The same trends are found in word choice. Word choice in the case of non-Australian English accents is generally constrained to the more stereotypical parts of the vocabulary of the speaker community that uses a particular accent. For instance, the use of *compadre* in line 1 of datum 4 reinforces that the accent that A uses is Mexican, despite the fact that it is not an especially unambiguous Mexican English accent. On the other hand, the data suggest that the actors are willing to use culturally specific Australian vocabulary because they are more confident that the audience will be able to map that use to a particular Australian subculture.

As with all devices that actors employ, word choice is used to convey some information about the character, the environment, or the plot of the scene. We assume that actors of improvised theatre make these choices rationally and, to sometimes differing degrees, purposefully. Word choice and grammatical variation becomes important later in the study,
particularly when we begin discussions of interpretation resolution and of covert communication, in chapters 6 and 7, respectively.

5.2.2.3  · Sociolinguistic variation and truth conditions

These examples of the strategic use of devices employing sociolinguistic features give a general idea of how and why such devices are used, and to what end they are deployed. These devices are a means of providing a rapid and efficient characterisation to characters in improvised theatre, by leveraging the general cultural knowledge of other actors and the audience. But strategies that rely on sociolinguistic variation are difficult to capture using DRT, for DRSs are built using an algorithm on the syntactic tree of the sentence. It seems therefore that these devices, particularly accent, are unable to be captured in standard DRT, which is problematic given that actors use such devices frequently.

The extensive use of sociolinguistic variation in improvised theatre justifies the introduction of the has_attr predicate in the previous chapter. As has been shown, a character’s attributes arise in precisely the same way that any other information is introduced into the plot-world. An actor makes an offer of the inclusion of that attribute in the specification of the character, and another actor produces a response that either accepts or rejects the inclusion of this information. Ascription of attributes to characters appears to be rejected less frequently than other kinds of information, for this kind of rejection would be far more likely to break the verisimilitude of the performance.

So, attribute ascription strategies tend to be taken up without disagreement, and usually with tacit acceptance. The actors, in other words, simply behave as if the ascription holds, and do not need to respond with explicit acceptance of the ascription.

5.2.3  · Creating a plot-world

In this section, I have looked at two of the major devices that actors use to create plot-worlds. The first was the use of presupposition, which is chiefly used to create a sense that the plot-world’s existence is not bound by the scene, but instead that it exists prior to the scene. This permits the accommodation of presuppositions that would otherwise be problematic. The second device that was examined was the use of sociolinguistic strategies.

As we have seen, creating a plot-world is an intensive process that involves a range of devices. For this reason, actors may prefer to bypass this process by using a plot-world that has already been established. I describe the devices associated with this in the following section.

5.3  · Resuming a plot-world

The alternative to beginning a scene and creating a completely novel world is to resume a plot that has already been established in a possible world. I call this the resumption of a plot-world. The fundamental difference between scenes that establish plot-worlds and those that resume plot-worlds is that the devices of the latter revolve around the clear signalling of which of the previously established plot-worlds is the setting of the current scene. As I
5.3 · Resuming a plot-world

discuss in this section, the devices that actors utilise differ between these two methods of
initiating scenes to such a degree that it is necessary to provide a brief description of the
means of resuming a plot-world. This is particularly important for the present study, for
scene resumptions are especially common practice in Harold longform. It is an integral
element of the aesthetic of the form that connections be established between scenes and
that these connections be exploited (Halpern, Close & Johnson 1994: 86), and this aesthetic
is promoted by the structure of the Harold.

In this section, I describe common devices that are used to resume a scene. The section
is divided into two parts, each of which deals with a particular narrative strategy. In
§5.3.1, I describe situations in which the state of the scene is shifted when the new scene
is introduced, and the strategies that are used therein. In §5.3.2, I describe situations in
which resumption of the scene is performed in such a way as to subvert the expectations
that are built up by the prior scene.

5.3.1 · Shifting state between scenes

Between scenes, the commonest shifts in the state of the scene have to do with changing
the location of the scene, or introducing new existents into the plot-world. Both of these
actions are performed with the existing knowledge of the scene that is being resumed.
This is interesting to our analysis because of a problem that actors in this situation must
navigate: the initial speaker of the scene must not only signal (a) that this scene resumes
another; (b) which of the previous scenes is being resumed; and (c) what has changed
between the previous scene that is being resumed and the current scene. The value of
submitting this situation to a game theoretic analysis is that felicitous strategies in these
cases often involve some element of optimisation, in that agents must formulate a strategy
that simultaneously achieves these goals, but does so as unambiguously as possible. The
optimisation problem is made more explicit below in §5.5.2.

Let us consider data 7 and 8, both overleaf. These data are taken from two separate
scenes in a single performance; datum 7 is from 3-B1, and datum 8 is from 3-C1. In datum
7, it is established earlier in the scene that the characters are watching a horror film. B
suggests that the horror film they are watching may not be a fictional account of some
horrific event, but may be a ’sick home video’, which makes A afraid. In datum 8, B gives
very clear indications that he is the same character of scene 3-B1. He does so by utilising
the same high-pitched, American-accented voice in 3-C1 as in 3-B1, which is particularly
marked when uttering the name assigned to A, Emily, which is not given to any other
character in the performance. The information here is both necessary and sufficient to
indicate to the audience that this scene is a continuation of the previous scene. It is nec-
essary because the resumption of the plot-world of 3-B1 might be unclear without such
an obvious characterisation of A, and it is sufficient because the audience does not require
any further information to make this inference correctly.

Meanwhile, B indicates in what way the scene has shifted. Once more, B attempts
to give only that information which is necessary and sufficient to communicate that the
Datum 7 Middle of 3-B1.

1 A: i watch this stuff for entertainment coz (1.1) i like the feeling but i don't really want to be this scared (1.3) 
2 B: ((speaks with a high-pitched, American accent)) well then why do you WATCH it? (1.0) emily l- ((heavy breathing; 1.8)) that's (0.4) the whole point of the horror ↑genre 
3 P: hah hah 
4 B: (honey) the whole point is to scare you 
5 A: to be a little bit scared in a safe environment (.) with a friend= 
6 B: =NO (0.7) NO it's to make you think at any point (0.6) someone could come in 
7 P: xXxxx 
8 B: and STAB you to death just like [that girl] 
9 A: [STAB me? ] 
10 B: or ME (0.1) or maybe I'm the killer (0.4) maybe i'm the killer [emily ] 
11 P: [XXXXXX] 
12 B: maybe i've been the killer all along (0.5) ((tears off a face-mask)) HOW-URRGHH 
13 P: XXXXXXXXXXXXXX 
14 B: how FUN IS THIS? [PLOT TWIST ] 
15 A: [WHAT THE FUCK?] 
16 B: I'M [GOING TO KILL you ] EMILY 
17 A: [YOU'RE THE KILLER? ] 
18 P: Xxxxx 
19 A: THAT'S A REAL KNIIFE 
20 B: YEAH i got it when i was cutting cheese earlier

Datum 8 Beginning of 3-C1. The actors and their indices are consistent with datum 7.

1 B: ((speaks with a high-pitched, American accent)) <emily> 
2 P: xxx 
3 B: where are you emily? 
4 P: Xxxx 
5 B: trying to hide (0.8) in this forest?
scene has shifted, and which respects have changed. In this instance, B’s utterance in line 5 of datum 8 indicates that the location of the scene has changed, through his use of the noun phrase this forest. Lines 1 and 3, conversely, indicate what in the scene has persisted from the previous scene.

**A predicate for location.** At this point, we do not have a formal method of dealing with the location of scenes, and so I introduce one now. To represent the location of a scene, I use a predicate, be_at, which has a valency of \( n + 1 \), where \( n \) is the number of characters who are at the location of the scene, and where the final place of the predicate is a description of the location. be_at has the restriction that a given character may only be predicated of being at a single location. That is, given character \( x \) and locations \( y, z \), it is not possible that \( \text{be}_\text{at}(x, y) \land \text{be}_\text{at}(x, z) \).

This method of using predicates, which is used also with has_attr, is an efficient way of including facts about the scene in the common ground. It does not require the establishment of a formal specification of such information with direct relation to the scene, but it instead uses the truth conditions that apply to DRT generally, and does so by applying predications to the characters. Therefore, the normal truth conditions of DRSs apply to a DRS \( K \) such that \( \text{be}_\text{at} \in C_K \). That is, such a DRS is true iff there are individuals in \( \mathcal{T} \) that meet the conditions applied to the entities specified in the DRS. As such, a DRS that includes a be_at statement is true in a plot-world iff there are characters who have the traits of those characters, and are at the location specified by the predication.

**5.3.2 · Subverting expectations**

As I set out above, one of the chief game theoretic reasons that actors employ strategies of scene resumption is that it is a means of accessing characters, locations, and the common ground that are all established. Actors are able to ‘latch on’ to the common ground of a previous scene, which allows them to constrain plausible plot-worlds. This is extremely efficient, as it allows the actors to use expectations that have already been established, without having to build them themselves.

An alternative use for the resumption of a plot-world is to subvert the expectations of the audience. The subversion of expectations results from certain trajectories that are established throughout the antecedent scenes, and a subsequent redirection of those trajectories; or, as is often the case, a recasting of those expectations.

An excellent demonstration of this is in the series of scenes that includes 1-A1, 1-B1, and 1-C3. The salient parts of each are reproduced in data 9, 10, and 11, respectively. Data 9, 10, and 11 all take place in a single plot-world, which involves a single family: A is the father, C is the mother, and B is the son. Additionally, the car, which is mentioned in each of the data, is the same existent in each datum.

The car is discussed in scene 1-A1. Recall from datum 3, which is part of 1-A1, that A has lost his job. Scene 1-A1 is resumed in 1-B1, when A utters (53) during a conversation about either turning the heating on or C having a hip operation.
Datum 9 End of 1-A1. The indices are consistent with datum 3, since this is a continuation of that scene.

1 A: d’you think every time you came home and disappointed me in some way i thought fuck it let’s have another kid?
2 P: XXXXXXXXXXX[xxx ]
3 A: [it’s not] how it works, james (2.5) you try and fix what you’ve got (1.0) to make it the best that it can be (0.7) you don’t replace it with a model that (0.7) takes ten minutes to warm up (0.6) that’s got no [gou:;l] to it
4 B: [to- to ] boot (. ) ten minutes to boot
5 P: [xxxx ]
6 A: [i don’t know]
7 (0.7)
8 B: warm up’s what you do with engines (1.5) or what you used to have to do with engines they’re pretty good now
9 P: [XX ]
10 B: [you use those] (0.4) little key clickers as you approach the car (. ) you don’t even the- to- it’s contactless start↑
11 (0.3)
12 A: i(’d) get up a half hour earlier to turn on the car
13 P: XXXxxxx
14 B: i make myself a coffee (2.1) i go back down (1.0) and i check that the car hasn’t overheated
15 P: xx
16 A: and then i turn it off for a bit
17 P: XXXxx
18 A: cause the old girl can only run for a certain amount of time before she’ll conk out
19 B: XXx

Datum 10 Middle of 1-B1. The indices are consistent with datum 9, for at an earlier point of the scene, an indication is given that A in this scene is identical to A in that scene (see below). Note that B is absent from this and the following scene; B is the son from 1-A1, and since he is still relevant to the common ground of this scene, I avoid using the index that denotes him.

1 A: at least we can still warm the car up (. ) there are some things that (. ) no matter how long we wait we’re not gonna get back
2 (0.9)
3 C: d’ya kno- (0.4) d’ya know ↑what (0.7) d’ya know what we should ↑do: (1.5) we should go an- we should just go to the car go:;v (1.1) ((raises eyebrows, speaks more quietly)) pop the heating on (. ) like we used to
4 (0.6)
5 A: i’ll take a while to heat up
### Datum 11

Beginning of 1-C3. The actors and their indices are consistent with datum 10.

<table>
<thead>
<tr>
<th>Turn</th>
<th>Dialogue</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>C: d’ya know what graham</td>
</tr>
<tr>
<td>2</td>
<td>P: hah heh hah hah</td>
</tr>
<tr>
<td>3</td>
<td>(3.3)</td>
</tr>
<tr>
<td>4</td>
<td>A: what’s that maureen my love?</td>
</tr>
<tr>
<td>5</td>
<td>(1.2)</td>
</tr>
<tr>
<td>6</td>
<td>C: well now the kids are all gone</td>
</tr>
<tr>
<td>7</td>
<td>D: VVVVVVVVVVVVVVVVVVVVVVVVVVVV ((the sound of heating))</td>
</tr>
<tr>
<td>8</td>
<td>P: heh heheh heh</td>
</tr>
<tr>
<td>9</td>
<td>C: why don’t we (0.5) you know (0.8) just</td>
</tr>
<tr>
<td>10</td>
<td>(0.7)</td>
</tr>
<tr>
<td>11</td>
<td>A: wind up the windows and asphyxiate?</td>
</tr>
<tr>
<td>12</td>
<td>P: XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX</td>
</tr>
</tbody>
</table>

(53) A: well i lost my job as well (0.5) so we couldn’t afford to do both

While it is not the first line, and in fact occurs only at the sixth turn, this conforms to what is stated earlier about resuming the common ground, in that it clearly marks out a particular, earlier scene, and provides some evidence that the 1-B1 is a resumption of that plot-world. The means of establishing that 1-C3 is a resumption of 1-B1, and therefore 1-A1, is simpler, as C calls the name attributed to A in the same, elderly, female\(^8\) voice:

(54) C: graham?

The scenes are thus resumed.

The ‘warming up’ of the car is mentioned in lines 12–18 of datum 9 by A. This idea is mentioned again by A in line 1 of datum 10, but is immediately transformed by C to a discussion of using the car’s internal heating, rather than heating the engine, to represent intimacy between A and C. In the final scene, the car’s heating is once again mentioned, and it appears as if C is going to suggest that the characters follow the expected course of action. However, A uses the heating, and the fact that the car’s engine must be ignited for the heating to be operational, to suggest that the characters asphyxiate themselves.

This series of scenes is a perspicuous case of the cumulative effect of discourse on interpretation. The notion of warming the car up is first introduced in 1-A1 by A as a means of emphasising his character’s nostalgia. It is mentioned again in 1-B1 to the same end, by suggesting that it is something that the middle-aged characters can rely on. C offers a reinterpretation of the notion, to include overtones of sexual activity; an offer that is subsequently taken up by A. The interpretation of suggestions to turn on the heating or to heat up the car are constrained to be euphemistic for sex. This constraint is reinforced a couple of times in 1-B1 (in parts not included in the transcript above), and so the actors’ and the audience’s expectations are led to conform to this premise. The constraint is dramat-

---

\(^8\)The actor in this instance is a male.
ically overturned in line 11 of datum 11, and apparently to great effect, given the loudness and duration of the audience’s laughter that follows.

This is a simple example of cumulative effects of discourse on constraining interpretation. The subversion of expectation works precisely because of this constraining effect, and it is only by cumulatively building the common ground so that certain interpretations are more plausible that such strategies may be deployed. Furthermore, as implied by data 9, 10, and 11, the actions of the actors in this series of scenes suggest that actors are aware of this effect, and use it to their advantage. For evidence of this claim, we can examine, for instance, a small subversion of expectation in A’s utterance of (55) from lines 3 and 5 in datum 10.

(55)

C: d’ya kno- (0.4) d’ya know ↑what (0.7) d’ya know what
we should ↑do: (1.5) we should go an- we should just
go to the car no: (1.1) ((raises eyebrows, speaks more quietly)) pop the heating on (.) like we used to
(0.6)

A: i’ll take a while to heat up

The warming up of the car having already been established as a topic of conversation in datum 9, and so A’s use of the first person I’ll, rather than the expected third person it’ll is another subversion of expectation.

We have seen that the resumption of a plot-world and the common ground established therein is often employed as an efficient way to dispense with the difficulty of establishing a plot-world, so that a plot trajectory can be developed without the burden of this task. As further details of the plot are offered and taken up, the available trajectories are constrained and so too are the expectations of the trajectories that might proceed from given point. A’s utterance subverts the expected trajectories of the audience and presumably also of the actors. In this instance, the strategy appears to be particularly effective. I devote chapter 6 to a discussion of this issue, in which I provide a probabilistic model that captures this basic device.

5.4 · Errors of consistency within and across scenes

As we have seen, the major undertaking that faces an actor in the early moments of a scene is the establishment of a plot-world. Establishing a plot-world involves either creating an entirely new plot-world or resuming a plot-world that has been constructed previously. While these methods differ somewhat in the approaches that suit them, they both require that actors ensure that the plot-world be consistent, for the group of actors involved in a performance and the audience both expect that this will be the case. This expectation makes subversion of expectation possible for instance, but it also imposes some severe constraints on the actors.

It is worth considering consistency in some depth, for it appears to be extremely important to both actors and the audience. We find that when actors introduce a fact that
5.4 Errors of consistency within and across scenes

Errors of consistency within and across scenes

Contradicts an already-established fact; they will generally attempt to repair the error. They usually do this by providing an explanation for how the inconsistency introduced might be consistent with the facts already known. The audience, likewise, makes indications that consistency is important to its members. The collective cringe of an audience, as well as its knowing laugh following an act of repair, are readily audible in the data when an actor makes an error that violates the consistency of the plot-world. This appears below in datum 12, in line 8.

Errors are of interest as an object of game theoretic and linguistic enquiry. Rational agents pursue their preferred outcomes. For whatever reason, an agent may perform an action that does not conform to her preferences, and in her attempts to repair the error, we are able to glimpse both her preferred outcome, and how her first action was erroneous. This argument also suggests as a corollary that actors have to be constantly aware that errors may occur, and so must be cautious to deploy strategies that minimise error.

An extremely obvious and relatively common example of violating the consistency of a world is calling a character by a name other than the name established for that character. This is a particularly egregious error when, as often happens, an actor will mistake a name that she herself has assigned to that character. The following is an example of precisely this situation.

**Datum 12 Middle of 1-A2.**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A: cheryl i love that you've learnt (. ) i love that you've changed</td>
</tr>
<tr>
<td>2</td>
<td>B: it's gonna be a very confused child</td>
</tr>
<tr>
<td>3</td>
<td>A: that's how you grow</td>
</tr>
<tr>
<td>4</td>
<td>B: right</td>
</tr>
<tr>
<td>5</td>
<td>A: you grow THROUGH confusion</td>
</tr>
<tr>
<td>6</td>
<td>B: yes</td>
</tr>
<tr>
<td>7</td>
<td>A: you gotta keep on evolving (. ) keep on changing (0.8) CATCHing every curveball jessica</td>
</tr>
<tr>
<td>8</td>
<td>P: hahhh ((uncomfortable laughter from an audience member))</td>
</tr>
</tbody>
</table>

Two names, Cheryl and Jessica, are given to the same character, by the same actor, in one scene. This may be treated as a double attribution of the character.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(56)</td>
<td>has_attr(B, NAME, CHERYL)</td>
</tr>
<tr>
<td></td>
<td>has_attr(B, NAME, JESSICA)</td>
</tr>
</tbody>
</table>

I stated above that in general actors try to remedy errors that are made. In scene 1-A2, part of which constitutes datum 12, A leaves the mistake unresolved. In datum 13 overleaf, we see a very similar error being committed. However, B is able to deftly repair

---

9 Recall from §3.1.3 that this was also the foundation of Benz’s error model (Benz 2012a,b).
Datum 13  Beginning of 3-C1.

1  A:  i don't know jerome (. i mean (0.2) i get that hook on the inside of the car so i can hang up (0.3) fluffy dice but i don't understand why you put a hook on the outside of the car=

2  B:  =birds

3  (1.1)

4  P:  XXXxx

5  B:  you think i'm gonna let my (2.2) daughter be swooped by a bird while she's drivin?

6  P:  XXXxx[xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx]

7  B:  [you think that, sally? (0.3) you think i'm gonna be that irresponsible as a parent

8  P:  ha hah hah hah

9  A:  um (2.2) i don't think they're that dangerous (0.8) i mean (0.9) i'll be in a car (0.3) i don't think they'll (0.5) like (0.7) swoop through the (0.3) windshield?

10  P:  XXX

11  A:  i [mean]

12  B:  [jess]ica let me tell you a story

13  A:  i hope=

14  B:  =ok?

15  (1.1)

16  A:  yeah::?=

17  B:  =alfred hitchcock made a film once

18  P:  Xxx#x#x

19  A:  and it w- it was a WARNING if anything wasn't it sally

20  P:  hh hh

21  B:  jessica-sally

22  P:  XXXXXX#x#x#x

:  

34  A:  and that's why i can't take showers either

35  (0.3)

36  B:  well yea- bird could get in the shower couldn't it?

37  P:  xxx

38  A:  "well"

39  B:  couldn't it sally (. jessica

40  P:  xxxXXXXx
Errors of consistency within and across scenes

the error. It is acknowledged by B, insofar as he makes an attempt to resolve it. To do so,
he incorporates the erroneous fact into the plot-world, by positing that *Sally* and *Jessica*
are two constituents of a hyphenated first name. Note that this strategy is successful only
because it is plausibly the case that a character has a hyphenated first name. What is
interesting about how plausibility is achieved is that B appeals to what is plausible in the
actual world—namely that some people have hyphenated first name—and provides further
grounding for the assertion by the casual use of S.J. to refer to A later in the scene. Thus,
the potential of a double name attribution, and incompatibility of the world, is avoided.
This strategy conforms to notions of plausibility that we have already encountered. In
establishing a scene, actors must work to ensure that the facts they introduce into the
common ground of the scene are plausible according to the plot-world as it is understood
up until that moment.

The following sequence of DRSs demonstrates how this repair occurs. Line 7 intro-
duces a name for A’s character, represented in (57).

\[
(x \text{ has_attr}(x, \text{name}, \text{Sally}))
\]

In line 12, B perhaps accidentally introduces a second name for the character, represented
in (58).

\[
(x \text{ has_attr}(x, \text{name}, \text{Sally})) \cup (x \text{ has_attr}(x, \text{name}, \text{Jessica}))
\]

\[
(x \text{ has_attr}(x, \text{name}, \text{Sally})) \cup (x \text{ has_attr}(x, \text{name}, \text{Jessica}))
\]

A does not object, and so this apparent contradiction is accepted into the plot’s common
ground. This allows for B to later assert in line 21 that A’s character has a hyphenated
name, as in (59).

\[
(x \text{ has_attr}(x, \text{name}, \text{Sally-Jessica}))
\]

This is permissible because both of these names have been used, and because of conven-
tions of naming in the actual world that the audience inhabits. In the model, this is treated
as an update of information that negates facts that are previously established.
Note that B switches the order of the constituent names. I have predicated that Sally-Jessica be the canonical order, for, as I noted above, S.J. is used later in the scene. I claim that the order of the names is switched to further ascribe nervousness and agitation to the character, rather than it being the commission of another error. This is suggested by the short pause in line 39, which makes the character appear nervous. Therefore, the analysis is not affected greatly by this.

What is revealing about this example, and the similar naming error in datum 12 above, is that neither of these errors is challenged, and, therefore, the effect of the repair is incorporated into the common ground of the plot. Failure to challenge may seem to be a secondary error on the part of the other actor, but in fact it reveals the means by which actors navigate errors in plot-world consistency.

Suppose that A in datum 13 challenges B by uttering something akin to (61) after B has uttered the incorrect name.

(61) That’s not my name! My name is Sally.

This strategy risks dispelling the scene’s verisimilitude, for it disputes a metanarrative move, namely the ascription of an attribute to a character. It therefore seems that actors prefer to allow the inconsistencies to enter into the scene, in order that they might be remedied later in the scene in a fashion that is consistent with the plot-world. A successful instance of this is the introduction of the hyphenated name in datum 13, while an unsuccessful instance of this strategy, where the mistake is not corrected, is in datum 12 above.

Instances of error therefore teach us a great deal about the preferences of actors in general. The most important lesson of the errors is that verisimilar repair strategies are preferred to those that break the verisimilitude of the scene.

5.5 · Heuristics for world building

The examples that we have examined in this chapter, as well as others in the data, suggest that certain devices are in general preferred by actors. A major contention of this thesis is that actors are game theoretically rational, and they thereby select actions on the basis of their overall (perceived) benefit.\[^{10}\] It appears, then, that there are certain kinds of moves which yield greater utility on average than others in improvised theatre. This section

\[^{10}\]I develop a fuller account of utility in improvised theatre in chapter 7.
makes some claims as to what types of strategies these are, and the consequences of these observations for the game theoretic model that was proposed in the previous chapter.

To identify the linguistic strategies that actors prefer, I outline a number of **heuristics** that actors appear to use when devising and construing strategies. These heuristics are devised from observing the data, and they attempt to outline what an actor involved in improvised performance can reasonably expect will inform the choices of her fellows. This expectation also flows the other way: an actor performs her action knowing that the heuristics detailed in this section form a lens through which actions are most likely to be interpreted. The heuristics are not meant to be taken as prescriptions, even if their statements generally appear to suggest as much. They instead detail optimal ways of producing an utterance *given some desired characteristic of that utterance*. The heuristics always refer to a Sender’s strategy, and as such the term *optimal strategy* refers to an utterance that fulfils that description. Receiver strategies are also informed by the heuristics, for the reason that I have articulated: if agents assume that their fellows will attempt to use optimal strategies, then describing such strategies allows us to identify which strategies agents will expect to be deployed. The heuristics, together with the formal apparatuses of game theory, DRT, and, as is introduced in chapter 7, information theoretic steganography, constitute the full specification of the formal descriptive model of improvised theatre. Each heuristic defines the optimality of a strategy with respect to a specific function of that strategy.

To illustrate the function of a strategy, let us consider the selection of an utterance in Vladimir’s Bomb Dilemma, the game discussed in §2.1.3. To recapitulate, Vladimir wishes to disarm an explosive device. Estragon, being an explosives expert, sends Vladimir a signal that indicates which of two wires to cut to do so. Estragon’s utterance must be *optimally unambiguous* and *optimally truthful* to communicate which wire will disarm the device, and the utilities of this game are aligned to make an optimally unambiguous, truthful utterance the best strategy. The solution is obvious: an utterance that is optimal with respect to ambiguity and truthfulness for Estragon is simply telling Vladimir which wire to cut. It is important to observe that this strategy is optimal only because the game’s utilities favour an optimally unambiguous and truthful utterance. If Estragon is evil and wishes the device to detonate, an optimally truthful utterance is not the preferred strategy. However, the preferred strategy will still be optimally unambiguous, for lying to Vladimir and telling him the incorrect wire to disarm the device will succeed as a strategy only if it is optimally unambiguous.

Heuristics also work to guide interpretations. Optimality with respect to a particular function is defined by stipulating certain properties of a strategy. So if a Receiver notices that an utterance has a certain property, she may infer that the Sender intends for that utterance to be optimal with respect to a particular function, especially if that property

---

11Neither of these requirements is included in the heuristics that I define below, and they are used for illustrative purposes only. They are not heuristics in their own right, for the selection of unambiguous and truthful strategies occurs according to the standard formulation of preference in game theory.
is generally associated with that property. Therefore, the Receiver is able to estimate the Sender’s motive in producing that particular utterance.

The chief advantage of including a specification of a number of rigorous heuristics is that the cumulative effects and interaction of a number of heuristics may be well captured. As is suggested by the consideration of Estragon’s strategies above, an agent may be interested in different functions of her utterances in different situations. We have seen in the data presented in this chapter that actors alter their strategies to suit the situations that they find themselves in. Strategies played when resuming a scene, for instance, differ from those played when an entirely new plot-world is being generated. It is therefore appropriate to have multiple heuristics that operate together, so that a strategy can be coordinated according to multiple axes of optimality.

The definition of multiple heuristics allows us to capture these axes. But it also allows for a description of the means by which heuristics interact to constrain an actor’s strategy selection. For instance, Estragon’s strategies that are only optimally unambiguous differ from those that are optimally unambiguous and optimally truthful. The heuristics that are specified in this section, then, define the function according to which a given strategy performed by an actor might be considered optimal, specifically with respect to generation of plot-worlds and of plots. It is important to observe throughout that no heuristic operates in isolation, each only describes the characteristics of a strategy that is optimal according to a specific function. It is moreover necessary to assume that each heuristic has an equal ability to influence the selection of a strategy, and that each strategy is considered with respect to the sum total of the heuristics that are presented in this chapter, and chapters 6 and 7.

The heuristics operate as a set of expectations mutually shared by actors in improvised theatre. The audience is not necessarily aware of these heuristics, and as such the audience acts as a naïve interpreter of utterances, which means that the audience interprets each utterance only according to the common ground of the scene that has been established, and not according to the heuristics. I leave this important notion unexplored for the present, for it is not a vital component of the model at this stage. I return to it in chapter 7.

I define five heuristics in this section. In §5.5.1, I attempt to define a heuristic for verisimilitude, but am unable to do so with the model in its present state. It is not until chapter 7 that I am able to provide a formal definition of verisimilitude. I also define compatibility, which describes a strategy’s sensitivity to the facts that have already been established in the plot-world. Likewise, in §5.5.2, I define two heuristics that pertain to the resumption of scenes: salience and distinction. Salience describes a strategy with respect to that strategy’s lack of ambiguity when signalling which of the previous scenes is being resumed. Distinction refers to a strategy’s optimality in signalling which facts about the state of the world have shifted. A heuristic that describes an utterance’s plausibility in the plot-world is introduced in §5.5.3. I conclude this section with some reflections on the use of heuristics in game theory in §5.5.4.
5.5.1 · Verisimilar worlds

The tendency of actors to assert the verisimilitude of a scene’s plot-world by asserting its persistence—that the world exists not merely for the duration of the scene—and its consistency has been documented thoroughly above. Now we turn to the game theoretic implications of the devices that are used to achieve this.

According to the data, the most common, and seemingly most successful, device that actors deploy to posit the consistency of the plot-world is presupposition. What follows is a game theoretic justification for using presupposition. Suppose that actor $i$ has a world $w$ in mind for the plot-world, and that she produces utterance $m$, which implies that a $K$ exists such that $m \in \|K\|$. As with any sufficiently complex model of game theoretic interaction, the number of worlds in $\|K\|$ may be extremely large, to the point that the likelihood of a second actor $j$ guessing the true $w$ may be very low. The strategies that are preferred by $i$ are those that ensure that $K$ is sufficiently detailed to constrain the worlds that satisfy the model to a sufficient degree that $j$ can guess $w$, or at least a plot-world that if sufficiently compatible with $w$ for $i$’s purposes.

Here we find again the problem of concealing the mechanism of the performance from the audience. In previous examples of signalling games, the best strategy was to be as explicit as possible, in selecting an utterance that unambiguously selected a single world. It appears that actors prefer strategies that could conceivably occur as part of a normal conversation in the plot-world. All of the data that have been presented in this chapter act to confirm this proposal, for none of the data reference either explicitly or implicitly that the scene is part of an improvised performance. Strategies that exhibit this property are called verisimilar, because they appear as if they are truly part of a plot-world, rather than part of a performance. This is formalised as a heuristic as follows.

**Heuristic 1: Verisimilitude.** *An action is called optimally verisimilar if it appears to believably belong to the plot-world; that is, if it does not appear to be produced by an actor pretending to be a character.*

It is imperative to understand that this heuristic, which is specific to improvised theatre, operates together with the normal Bayesian game theoretic notion of belief revision (see Battigalli & Siniscalchi 1999). That is, while actors’ preferences are expected to concur with heuristic 1, they also prefer strategies that provide clear evidence of the state of the world. Actors must find a strategy that reveals the world they have in mind to their fellow actors, but not necessarily to the audience. We have already seen that this tension is fundamental to the operation of actors in improvised theatre, and this phenomenon is well documented in the theatre studies literature (see §2.2.2.1). But we presently lack the formalisms to properly approach this tension game theoretically. Indeed, we lack at this point of the thesis even the vocabulary to describe the difference between what is ostensibly communicated to the audience, and what is communicated about the performance to the actors, let alone how this baroque system influences the deliberation of an actor in selecting a strategy, and in interpreting an action they have observed. The impacts of
Chapter 5 · Plot-world generation and consistency

...this heuristic are extremely complex, and require, amongst other things, a more advanced understanding of the interpretive act in improvised theatre. And so the construction of an apparatus to properly model communication of this type is given its own chapter: chapter 7.

We have seen that actors tend to prefer strategies that endow the plot-world with the qualities of a possible world, the two most important of these qualities being those that I have flagged: consistency and persistence. The most prominent of such strategies that have been explored in this chapter are those that design utterances that trigger presuppositions which are resolvable if and only if conditions exist in the plot-world that have not been introduced. The presupposition triggers of these utterances are, then, the targets of the strategic choice of the actors, in the same fashion that the state of the world is the target of signals in signalling games. It is probably unreasonable, however, to suggest that an actor has a fully-specified possible world in mind when she commences a scene. We should instead suppose that the actor has in mind a set of salient facts that happen to constrain candidates for the plot-world. This notion is supported by the ease with which new facts are accepted as true by the actor who has a plot-world in mind. But this is only the case if these facts are compatible with what is established of the plot-world beforehand. This is distilled into heuristic 2.

Heuristic 2: Compatibility. A strategy is optimally compatible if the facts expressed by that strategy do not contradict what is established about the plot-world prior to that strategy’s deployment.

This heuristic expresses the notion that utterances that introduce new information are allowed, provided they do not contradict a piece of information that is already established. Establishment of information refers to the assertion of that information. Otherwise, surprising speech acts would not be permissible. A formal definition of this heuristic is given in definition 5.1.

Definition 5.1: Optimal compatibility (from heuristic 2). Let $\Gamma^t$ be a common ground at time $t$. The set of optimally compatible strategies $S^*$ is the smallest set such that, for each $s^* \in S^*$, $s^* \in \|K\|$, such that $\|K\| \subseteq \|\Gamma^t\|$.

This definition states that compatible strategies are those that are expressed by DRSs whose extensions are subsets of the extension of the common ground until that point.

5.5.2 · Resuming plot-worlds

There are two key factors that appear to inform actors’ strategies regarding the resumption of a previously established plot-world. First, actors tend to use some key piece of information that signals to their fellows and the audience the identity of the world that is to be resumed.

Heuristic 3: Salience. If a Sender intends that a scene be resumed, a strategy is optimally salient with respect to that scene (called the target scene) if there is no other scene that is more probably the target scene given the strategy.
The strategies that fulfill this heuristic need not be complex. A very basic strategy is to have actors reprise their characters, and to use names that have been established in previous scenes. Both 1-B1 and 1-C3 (portions of which are transcribed in data 10 and 11, respectively) begin with Maureen calling Graham by his name in the same middle-aged voice. Indeed, throughout the data the actors tend to deploy strategies of resumption as efficiently as possible; that is, by expending as little effort as possible. This involves effecting a strategy that has a greater probability of invoking a certain common ground over all previously-established common grounds, and over entirely new common ground. The formal definition is given in definition 5.2.

Definition 5.2: Optimal salience (from heuristic 3). Let $Q_i$ be a scene, and $Q_1, \ldots, Q_{i-1}$ be the preceding scenes. Let each scene $Q_j$ have a common ground $\Gamma_{Q_j}$, and let $\Gamma_{Q_1} = \emptyset$. A scene-initial strategy for Speaker $s^*_S \in S_S$, is optimally salient with respect to a target scene $Q_j \in \{Q_1, \ldots, Q_{i-1}\}$ iff

\begin{itemize}
  \item[(1)] there is no $s' \in S_S$ such that $\Pr(\Gamma_{Q_j} | s') > \Pr(\Gamma_{Q_j} | s^*_S)$; and
  \item[(2)] there is no scene $Q'_j \in \{Q_1, \ldots, Q_{i-1}\}$ such that $\Pr(\Gamma_{Q'_j} | s^*_S) > \Pr(\Gamma_{Q_j} | s^*_S)$.
\end{itemize}

This formal expression of the heuristic states that there are two conditions for optimally salient strategies. First, there must be no strategy available to the Sender other than the optimal strategy that will yield a higher conditional probability of the target scene. Second, there must be no scene that is more probable than the target scene given the strategy that is uttered.

The second key factor that informs an actor’s selection of strategies in situations that involve the resumption of the plot-world is the need to signal what has changed between the plot-world’s current and previous instantiations. Actors do this, relying on the fact that their fellow actors and the audience will assume that everything else has remained the same.

Heuristic 4: Distinction. If a Sender intends that a scene be resumed, a strategy is optimally distinct if only the intended differences between the target scene’s common ground and the present scene’s common ground are communicated.

We have seen some examples of this heuristic’s operation. Datum 8 is a particularly interesting instance. In line 5, A utters (62), which indicates that the location of the scene has shifted from a lounge room in the previous scene to a forest.

\begin{quote}
(62)  B: trying to hide (0.8) in this forest?
\end{quote}

However, the attributes of the characters and their relationship to one another are not mentioned; B continues to be pursued by A, and the latter continues to carry a knife. These facts, because they are not mentioned, are assumed by actors and audiences to persist from one scene to another. This is codified as an assumption.

Assumption 5.1: If a scene is resumed, all facts are assumed to be identical to those in the target scene, unless this assumption is explicitly overturned.
Let us define heuristic 4 formally.

**Definition 5.3: Optimal distinction (from heuristic 4).** Let $Q_0$ and $Q_1$ be two scenes, such that $Q_0 < Q_1$, with common grounds $\Gamma_{Q_0}$ and $\Gamma_{Q_1}$, such that initially $\Gamma_{Q_1} = \emptyset$. A strategy $m^*$ is optimally distinct iff the following conditions are met:

(i) after observation of $m^*$, $\Gamma_{Q_1}$ becomes $\emptyset \cup \Gamma_{Q_0} \cup K$ such that $m^* \in \|K\|$; and

(ii) $\llbracket \Gamma_{Q_1} \rrbracket \subset \llbracket \Gamma_{Q_0} \rrbracket$.

This definition warrants further explanation. The heuristic only applies to situations which take place at the beginning of a scene, and in which the Sender wishes to resume an earlier scene. This is expressed by the common ground of $Q_0$ being contrasted with that of $Q_1$, which is empty. When $m^*$ is produced, two things occur. First, the DRS that expresses $m^*$ is merged into $\Gamma_{Q_1}$. Second, $\Gamma_{Q_0}$ is also merged into the common ground, because $m^*$ resumes a scene. Finally, the extension of the common ground at the commencement of the second scene must be a proper subset of the extension of the common ground at the end of the first scene. It must be a proper subset, that is, the two cannot be equivalent, because the common ground of the second scene contains more information than the common ground of the first scene, and therefore there are fewer viable candidates for the plot-world.

I have chosen to avoid enforcing a stipulation on the quantity of information that should be included in the optimal strategy. This choice is motivated largely by the fact that the quantity of information is governed by the players’ preferences over the trajectories that the plot might take, and is a separate issue.

### 5.5.3 · Plausibility

The final heuristic that I propose has to do with the plausibility of facts that are introduced. I have noted that actors are able to introduce facts about the possible world with a great degree of confidence that the audience will accommodate those facts as true for the plot-world, even if they are incompatible with the actual world. However, it appears that the actors will tend to provide more elaborate explanations of one of their strategies if it seems implausible.

While a specific event might be described as possible in a given plot-world, I prefer the term plausibility, for it gives us leeway that the term possibility does not. To illustrate this point, consider the following two scenarios that appear in the data. In scene 3-B2 which constitutes datum 16 in chapter 6 below, the animals in the zoo speak amongst themselves about the human characters, who are in fact the attractions in the zoo. In lines 20–22 of datum 7, A notes that B has a knife, and B explains that he got the knife while cutting cheese earlier. The less probable of these two scenarios—that is, birds speaking—does not require, or elicit, any explanation from the actors, while the mundanely possible but odd fact that a character has a knife while watching a movie does. That birds speak is a given fact in the particular plot-world, because it is dissimilar enough from the actual world for the audience to accept bird speech as a fact in the fictitious plot-world. It appears that the
audience views this fact about the plot-world as plausible *in itself*, since it is not possible in the actual world. However, carrying a knife while watching a movie with a friend is possible in the actual world—in the sense that people in the actual world have knives in their houses—but it is slightly implausible. Since the situation in datum 7 is recognisable as being possible in the actual world, the actor is compelled to explain the implausible fact in the scene; that is, how and why his character has the knife.

*Plausibility*, it seems, is measured against the facts that are introduced in the world, *and* the known facts of the actual world. It is simply expressed in the following heuristic.

**Heuristic 5: Plausibility.** A strategy is optimally plausible in the plot-world if, according to the information that is already known about the plot-world, the plot-world has a high probability of being an element of the extension of the strategy’s DRS.

As with heuristic 4, there is an additional assumption that underlies this heuristic.

**Assumption 5.2:** If some aspect of the plot-world is unknown, then it is assumed that the aspect is identical in the actual world.

The counterintuitive claim about plausibility and explanation above is a consequence of this assumption. If a particular fact in the plot-world is utterly dissimilar to a corresponding fact in the actual world, then the frame involving the fact in the plot-world is assumed by the audience to also be utterly dissimilar to the frame involving the fact in the actual world. And therefore, the audience does not expect that anything about the frame will resemble the actual world. However, if a particular fact in a plot-world is sufficiently similar to a corresponding fact in the actual world that it does not erase the frame involving that fact, then the facts constituting the frame are expected to be identical in the actual world. In other words, slightly odd facts in a familiar frame are more likely to raise questions of plausibility than those that are in a completely foreign frame.

But how can we express this formally, without introducing a large array of formalisms to deal with plausibility? While it may not be entirely satisfactory, a way of doing so that fits the purpose of this project is simply to use Bayesian probabilities. Suppose two players, *i* and *j*, are engaged in a conversation, and suppose that the state of the world has been more-or-less established at a prior point in the conversation. That is, both players have received some evidence of the state of the world *w*, and also each has evidence for what the other believes the most probable possible world is. Now suppose that *i* wishes to introduce a fact that is plausible in *w*. This is simple. She selects an utterance that expresses this fact, and, knowing that *j* believes that the fact is probable, she can be confident that it will be accepted without resistance. Now suppose that the opposite is true as regards the fact that *i* wishes to introduce; that is, that the fact is implausible in *w*. Implausibility of a fact is simply the condition that the fact has a low probability *given what is known about the world*. In the case of an implausible fact, then, *i* has to expend more effort to ensure that the conditional probability of the fact given *w* is increased such that *j* is more likely to accommodate the introduction of that fact.
I do not stipulate precise values for the *high* and *low* probabilities that I describe above, for it depends greatly on how similar to the actual world the scene is. Therefore, I prefer to stipulate that this is a problem of maximising the probability. In effect, *i* has to select a strategy that maximises the probability of *j* accommodating that fact. And to do so, she must maximise the conditional probability that the DRS representing her utterance $K^µ$ is consistent with the common ground’s DRS $K^Q$. A formal definition that captures this is definition 5.4.

**Definition 5.4**: Optimal distinction (from heuristic 5). Let $m$ be an utterance, let $κ$ be a set of DRSs such that for each $K ∈ κ$, it holds that $m ∈ [K]$. Furthermore, let $Γ_Q$ be the common ground of scene $Q$, and let $w$ be the plot-world. A strategy $m$ is optimally plausible in $w$ iff arg max$_{K ∈ κ}$ $Pr(w ∈ [K]|Γ_Q)$.

This definition states that an optimally distinct strategy is one whose DRS representation is the most probable amongst possibilities of like kind given the common ground of an established scene $Q$.

### 5.5.4 · Heuristics in interpretation

I noted at the beginning of this section that these heuristics should not be viewed as prescriptions, but instead as generalised principles of how actors behave in order to optimally achieve certain functions of their utterances, based on how actors tend to behave in the data. Since the heuristics are based on observations of actual behaviour, I propose that they capture not only what appears to be the best practice for producing utterances with certain qualities, but also how actors expect their fellows will act to bring about a certain linguistic function, and how certain actions are generally interpreted by the other actors with respect to their linguistic function. One of the most basic aspects of the establishment of successful models in game theory is the modelling of a player’s predictions of how her fellows will act. A player’s expectations based on her beliefs are extremely important, as it is through this lens that the player interprets the actions of her fellows, and that the player selects her strategy. The heuristics that I have identified are expectations of this kind, specific to improvised theatre.

The optimality of the heuristics becomes most important in chapter 7, where the information that actors attempt to convey becomes mathematically complex. We return to them then.

### 5.6 · Conclusion

To conclude this chapter, I summarise what has been discussed about how actors generally approach the task of creating a plot-world.

When actors begin a scene with a plot-world that has not already been established, they tend to assert that the plot-world exists beyond its instantiation as a scene. It is generally asserted that the plot-world (1) has a persistent history and existents; (2) is consistent; and (3) is independent from the actual world. Two common means of asserting these are as follows. First, players produce utterances that trigger presuppositions that are not able to be
resolved using only the information about the plot-world that is available to the audience. The intended effect of this is to suggest that the characters have the requisite information to resolve the presuppositions felicitously, and that this information is simply part of the world knowledge that the audience lacks. Second, players attribute sociolinguistic traits to characters in order to convey information about the plot-world, such as the location of the scene, information about the characters who inhabit the scene, and so on. The information that is introduced by this mechanism deals mostly with stereotypes or typical characteristics of possessors of those sociolinguistic traits.

When actors begin a scene with a plot-world that has already been established, they tend to clearly identify which of the previous plot-worlds they are resuming, and to announce what differs between the previous and current instances. They achieve the former generally by assuming a character they played in the previous scene, and by using a characteristic trait (such as an accent) of that character. The latter is usually performed early on in the scene through mentioning the fact that has changed. All other facts, if they are not mentioned, are assumed to remain static, and the actors act as if they are.

Actors always attempt to ensure that the plot-world is internally consistent. They are careful to avoid introducing any facts that appear to contradict a fact that has already been introduced into the plot-world. I noted that the commonest example of a break in consistency is to call a character by an incorrect name. If a break in consistency does occur, successful actors will attempt to incorporate the break into the logic of the plot-world.

In this chapter, we have examined the strategies that individual actors use to construct plot-worlds. This gives us an understanding of how actors typically introduce information into an empty common ground, or how they utilise information that has already been established, and how they signal their intentions and preferences to the audience, and to each other. In the next chapter, we take this idea further when we turn to an exploration of the strategies that actors use to control the overall common ground of a performance, and how these strategies impact specifically on questions of interpretation. This exploration leads to the claim that the overall common ground is in fact constituted of so-called partial grounds that are shared between any set containing at least two players. The following chapter, then, deals with means of simultaneously controlling multiple common grounds.
Chapter 6

Resolving interpretations with divergent common grounds

Most controversies have arisen from men’s failure to explain their own mind, or to interpret the mind of someone else. For really, when they contradict one another most energetically they either have the same thoughts or they are thinking of different things, so that what each thinks are errors and absurdities in the other are not.

Benedict Spinoza, Ethics

In the previous chapter, we examined the various ways in which actors build common ground, and the strategic functions that world-building involves. In this chapter, I discuss something that was repeatedly touched upon in the previous chapter, though not fully explored in any great depth: interpretation. Interpretation problems are the basic problems that are dealt with in game theoretic treatments of language (see, for instance, Benz 2007, Franke 2011, Asher & Williams 2006). As we have seen so far, signalling games are chiefly concerned with two interrelated phenomena. First, signalling games are models in which a Receiver selects an interpretation that responds to what she infers is the most probable candidate for the possible world. This inference is based on the observation of an utterance that is produced by the Sender, the design of which is the second phenomenon that signalling games model. It is these two phenomena taken together that constitute the game theoretic understanding of interpretation.

I would like to explore in greater depth how we might deal with the inclusion of a common ground in this game theoretic notion of interpretation, and in particular how this applies to improvised theatre. My major contention in this chapter is that, once one introduces a notion of common ground, one must necessarily revise the basic view one takes of interaction. Negotiation becomes central to the understanding of how meaning is conveyed and taken up in conversation. We have already seen that negotiation is implicitly present in discussions of both communication in improvised theatre—for actors must negotiate to successfully generate and execute a plot—, and in game theory—for players
in a cooperative game must adapt their behaviour to suit their beliefs about their fellows, and so negotiate.

But I take a more radical position in this chapter. I claim that, since the interpretation of an utterance depends on the common ground, and since information in the common ground may be disputed or rejected, meaning thereby that common grounds are negotiated, the meanings of *utterances themselves* arise only through a process of negotiation. There are a few ways this statement may be read, and so to be clear, I contend that, if the interpretation of an utterance may only arise as a result of negotiation, it follows that the *utterance itself* only has meaning if a meaning is agreed upon.

This chapter examines negotiation, and the effect it has on the design and interpretation of utterances. Part of negotiation is making concessions, and so I am also concerned with the effect that disagreement has on interpretation, especially with the effects of divergences in perceptions of the common ground on utterance design and interpretation. The discussions that I undertake here underpin the formal model of covert communication that forms the basis of communication in improvised theatre, which is the topic of chapter 7. The chapter proceeds as follows. In §6.1, I introduce a formal apparatus for the inclusion of common ground in interpreted signalling games. In §6.1.1, I describe some existing approaches to this problem. In §6.1.2, I discuss the formal properties of a common ground that exists as an entity of a game’s description, and I discuss some important advantages and consequences of this formulation. In §6.1.3, I describe how this characterisation of common ground leads to a formal notion of negotiation becoming more prominent than it normally is in signalling games. I present a number of data that support this conception of common ground and negotiation in §6.2. This section is divided by the ‘targets’ of the interpretation: §6.2.1 deals with the negotiation of the interpretation of parts of utterances, or of utterances-as-offers themselves; and §6.2.2 deals with the negotiation of the interpretation of players’ intentions over plot trajectories. Finally, as was the case in the previous chapter, §6.3 proposes some heuristics, and devises formal definitions of these heuristics, that describe the lens through which actors view the actions of their fellows. Three heuristics are proposed: a heuristic for optimal markedness of a certain kind of interpretation in §6.3.1; a heuristic for optimal use of strategies of repetition in §6.3.2; and a heuristic for the various kinds of response strategies in §6.3.3. The chapter concludes in §6.4.

6.1 · Context and negotiation

We have seen already, in §3.2, that in DRT anaphora resolution requires an entity to be present in the common ground that has corresponding conditions to those placed on the anaphoric expression. For instance, the entity $x$ which is able to resolve *him* in sentence (63) must be, according to the semantic constraints on the pronoun, a salient and unique male, and it must be different from the entity that resolves *he*, or else the reflexive pronoun would be used. The grammatical case of the pronouns does not matter for the present example.
(63) Vladimir sees Pozzo. He loves him.

The first sentence of the two provides an entity which satisfies the semantic requirements of *him*, and so naturally the anaphora is resolved. The same process is true, *mutatis mutandis*, of the resolution of *Vladimir* and *he*. The DRS of this small common ground, derived after merging the DRSs of the individual sentences, is as follows.

\[
K_1 = \begin{array}{c}
  x, y \\
  \text{Vladimir}(x) \\
  \text{Pozzo}(y) \\
  \text{loves}(x, y)
\end{array}
\]

Recall that the truth of a given DRS is determined by the existence of two embedding functions, \(g\) and \(h\). An embedding function is a homomorphism from a set of discourse referents into a DRS model. That is, a DRS is true iff for each of its referents, there is a matching referent in the model’s set of individuals, such that the same conditions apply to both the DRS’s referent and the model’s individual. The formal implementation of DRS truth per definition 3.13 is that a DRS \(K\) is true if there is a pair of embedding functions \(g\) and \(h\) such that \(g \subseteq_D h\), and such that for all conditions \(\varphi \in C_K\), \(h \models_{W, w} \varphi\).

So, for \(K_1\) there must be an entity \(y \in \mathcal{D}_{38}\) that satisfies the conditions imposed on \(y \in C_{K_1}\). In \(K_1\), these conditions are (a) that the entity is named *Pozzo*; and (b) that the entity is loved by another entity, whose name is *Vladimir*. If there is no entity in a particular world \(w\) that meets these conditions, then the DRS is not true at \(w\). At all worlds in \(\llbracket K_1 \rrbracket\), such an individual exists. These conditions do not specify what kind of entity \(y\) must be, and so the entity could conceivably be, for instance, a human, or another kind of animal. Provided that there is an entity named *Pozzo* who is loved by another entity called Vladimir at \(w\), then \(w \in \llbracket K_1 \rrbracket\), regardless of what the entity actually is.

Now, let us consider the longer, more comprehensive common ground in (65), which is made up of three short sentences.

(65) Pozzo is Vladimir’s cat. Vladimir sees Pozzo. He loves him.

The DRS that results from the merge of the DRS of these individual sentences is the following.

\[
K_2 = \begin{array}{c}
  x, y \\
  \text{Vladimir}(x) \\
  \text{Pozzo}(y) \\
  \text{cat}(y) \\
  \text{owns}(x, y) \\
  \text{sees}(x, y) \\
  \text{loves}(x, y)
\end{array}
\]

In \(K_2\), three additional conditions are present, namely that \(y\) is a cat, that \(x\) owns \(y\), and that \(x\) sees \(y\). Since \(K_1\) and \(K_2\) share a couple of conditions, \(\llbracket K_1 \rrbracket\) and \(\llbracket K_2 \rrbracket\) both include those
worlds at which the conditions loves\((x, y)\), Vladimir\((x)\), and Pozzo\((y)\) hold. In fact, the relationship between the two extensions is stronger than this. If \(K_2\) is true, then the three conditions of \(K_1\) hold, since \(C_{K_2} \subseteq C_{K_1}\). So, \(K_2\)’s truth implies \(K_1\)’s. But this implication does not hold in the opposite direction: there are worlds in \([K_1]\) where Pozzo is not a cat, for instance, for this condition is not present in \(K_1\). Therefore, \([K_2]\) \(\subseteq [K_1]\). The chief consequence of this for the resolution of anaphora is that there are fewer entities that may resolve the anaphora in the common ground that is generated by the sentences in \((65)\) than by the common ground that is generated by the sentences in \((63)\).

The addition of information in the common ground thus restricts possible resolutions of anaphora. In general terms, the size of the set of models that satisfy a DRS is diminished by the inclusion of common ground information. For as more information of a possible world is introduced, the worlds that are consistent with this information become fewer.

I argue that this phenomenon extends beyond anaphora, and that the interpretation of utterances depends on antecedent common ground DRSs. Put differently, the common ground DRS at time \(t\) constrains the interpretations of an utterance \(m\), such that certain interpretations are more likely to be consistent with the world. This observation appears obvious in the context of what has been discussed so far. Furthermore, it is present in a number of fields of linguistics: it forms the basis of segmented DRT (see Asher & Lascarides 2003, Lascarides & Asher 2008), where particular speech acts have rhetoric relations to others present in the discourse; a similar phenomenon is important in the relevance theory literature, where it is called (semantic) narrowing (see Carston 2004, Sperber & Wilson 2004); and it is also present in theories of lexical semantics, including most strongly the Generative Lexicon, where collocates and other common ground information coerce the types of terms (for relevant examples, see Pustejovsky 1991, 1993, 1995). But, since we are dealing here with a theory of strategic selection of utterances, I claim that speakers are aware of this phenomenon, and that they approach the construction of common ground with this in mind.

In this section, I discuss the common ground of a conversation, its effects, and strategies that deal explicitly with control of the common ground. In §6.1.1, I argue that common ground ought to be considered a core part of any valid model of language use. To do so, I document the extent to which existing game theoretic models of language use rely on notions of common ground to drive models of interpretation. I go on to propose a game theoretic means of accounting for the effects of common ground on the design and interpretation of utterances which follows the general, Bayesian approach taken throughout the thesis. In §6.1.2, I describe a formal model of common ground for game theory. In §6.1.3, I advance one of the core claims of this thesis, articulated in claim B in chapter 1: that common ground arises only through an iterative process of negotiation. I make the argument that interlocutors must agree on what is included in the common ground, which implies that not all information that is expressed in utterances in a conversation enters the common ground. Indeed, it appears that interpretation itself occurs only under the same conditions of agreement. I motivate these claims in this subsection.
6.1.1 · Existing game models of common ground

I begin the task of modelling the negotiated establishment of common ground by exploring the way that common ground is treated in game theoretic treatments of language. To do so, let us consider a well-studied game theoretic description of a communicative situation which involves *scalar implicature*.

Implicature is characterised by Horn (2004) (emphasis is his own) as “a component of speaker meaning that constitutes an aspect of what is meant in a speaker’s utterance without being part of what is said”. In his seminal work on the topic, Grice (1975) introduced two forms of implicature: *conventional* and *conversational*. In an earlier work (Grice 1969), he discusses a third form, *scalar implicatures*, which are a subtype of conventional implicatures. A conventional implicature is an implicature that largely fits Horn’s definition above. Consider the following:

(67) Malcolm is a politician, but he seems to be honest.
⇝ Politicians are generally dishonest.

The implicated meaning of the second proposition is triggered by the ‘conventional meaning’ of the conjunction *but*. This is contrasted with conversational implicature, which “either refers to the act of meaning one thing by saying something else or to the content itself that is conveyed by such speech acts” (Blome-Tillmann 2013: 170), an act which is exemplified by the following:

(68) Estragon: Is Lucky a good pianist?
Vladimir: He has beautiful handwriting.
⇝ Lucky is not a good pianist.

Vladimir’s seemingly irrelevant reply implicates a negative answer to the question, without any word or any combination of words *per se* triggering this meaning. According to Grice (1975), this meaning arises because Vladimir behaves in a manner that is unexpected by Estragon. The basic idea is that, if Estragon believes that Vladimir will adhere to certain norms of conversation, then there must be a reason why Vladimir eschews these norms. As such, an utterance’s implicated meanings are *inferred* by the Receiver of the utterance. (I leave the discussion of Grice’s theory of implicates here; see Grice 1989, Horn 2004 for further information.)

Scalar implicatures are simply sentences that implicate quantity. For instance, *some* implicates *not all*; but, in the game theoretical example, this is not always clear-cut. The communicative situation that we consider is based on an example of scalar implicature given by Sauerland (2004: 367):

(69) Kai ate some peas.

The game theoretic model of this utterance is based on the signalling game model formulated by Franke (2009).
Suppose that there are two players, S and R. R wishes to know if she has to buy more peas after Kai was over for dinner the night before. Imagine that she is extremely parsimonious, so she will only buy peas if there are none remaining. She can choose between buying a packet of peas (\(a_{\text{buy}}\)) or leaving it at the supermarket (\(a_{\text{leave}}\)). Suppose that there are two states of the world: \(w_{\exists \neg \forall}\), where Kai ate some but not all of the peas, and \(w_{\forall}\), where Kai ate all of the peas. There are likewise (somewhat artificially) only two strategies in \(\mu_{S}\):

\[
\begin{align*}
\text{(70)} & \quad \text{a. Kai ate some peas. (denoted } m_{\text{some}}) \\
& \quad \text{b. Kai ate all the peas. (denoted } m_{\text{all}}) \\
\end{align*}
\]

What complicates the game model is the semantic function. Observe that \(m_{\text{some}}\) is felicitous if Kai has eaten some of the available peas, but is also felicitous if he has eaten all of the available peas; claiming to eat some peas can mean exhaustively eating all of the available peas, but it could also mean that only a portion of the total available peas are eaten. However, uttering \(m_{\text{all}}\) can only mean that all of the available peas have been eaten. The full specification of the denotation function of the utterances in 6.1 models this.\(^1\)

\[
\begin{align*}
\text{den}(m_{\text{some}}) & = \{w_{\exists \neg \forall}, w_{\forall}\} \\
\text{den}(m_{\text{all}}) & = \{w_{\forall}\}
\end{align*}
\]

For the moment, we assume cooperative agents, and so a Lewis utility function is employed. This means that S and R are both interested in coordinating R’s action with the state of the world. In this case, this means that both agents wish R to buy more peas only if there were no peas left. A full specification in normal form is given in figure 6.1.

<table>
<thead>
<tr>
<th>(w) ∈ (W)</th>
<th>(\text{Pr}_{R}(w))</th>
<th>(m_{\text{some}})</th>
<th>(m_{\text{all}})</th>
<th>(a_{\text{buy}})</th>
<th>(a_{\text{leave}})</th>
</tr>
</thead>
<tbody>
<tr>
<td>(w_{\exists \neg \forall})</td>
<td>1/2</td>
<td>∈</td>
<td>(1, 1)</td>
<td>((-1, -1))</td>
<td></td>
</tr>
<tr>
<td>(w_{\forall})</td>
<td>1/2</td>
<td>∈</td>
<td>∈</td>
<td>((-1, -1))</td>
<td>((1, 1))</td>
</tr>
</tbody>
</table>

Figure 6.1: Tabular representation of the pea signalling game.

A solution to the game presents itself when we consider two basic facts of signalling games: (1) that both players mutually expect cooperative actions; and (2) that R updates her belief of the potential states of the world given observation of the utterance. Before observation of the utterance, note that the Receiver believes that each possible world is equiprobable. Suppose that S utters \(m_{\text{all}}\). In this case, given the semantic denotation function, R knows that S is indicating that \(w_{\forall}\) is the state of the world: \(m_{\text{all}}\) is truthful iff \(w_{\forall}\). Clearly, then, \(\text{Pr}(w_{\forall} | m_{\text{all}}) = 1\), and so R will select \(a_{\text{buy}}\). The case is not so clearly

\(^{1}\)Note that a semantic denotation function is used here only because I am describing the model of this communicative situation that is formulated by Franke (2009). Recall that this is not the method of semantic denotation that I use, and that the game model of communication that I define uses a semantics function and an extension function. These integrate with a DRS language and its intensional model.
defined if $S$ emits $m_{some}$. But since $R$ knows that $S$ wishes to be completely perspicuous in her utterance selection for their utilities are aligned, and so they each wish to be understood. Since $R$ knows that there exists precisely one unambiguous way to signal $w_\forall$, she knows that $m_{some}$ will be used only if the case is $w_\exists \neg \forall$. Thus, $S$ will produce $m_{some}$ iff $w_\exists \neg \forall$, and $m_{all}$ iff $w_\forall$.

The ability of signalling games to model complex communicative problems like that shown in figure 6.1 leads Franke (2009: 18) to claim that signalling games represent an idealised and abstracted model of (certain parameters of) an utterance’s context. This is certainly the case in examples such as the implicature game that I have presented here. But does this approach also capture to the same degree the effects of context over multiple signalling games? Furthermore, does it allow for the modelling of games where the number of available worlds and available strategies is significantly larger than is available in figure 6.1—being potentially infinite—and where these types and strategies do not have such artificial constraints placed upon them? This appears not to be the case, and I argue that it is better practice to place the context (or common ground) not as a byproduct of agents’ Bayesian reasoning, but rather as an entity in itself.

6.1.2 · A new game model of common ground

Here, I devise a simple method of modelling the effects of a common ground of this kind on the strategy selections of agents in an interpretation game. To do so, I discuss the Sender’s use of her knowledge of the common ground to construct utterances, and the Receiver’s use of her knowledge of the common ground to felicitously interpret these utterances. Since the Sender has an intended interpretation, and since we presume for most situations that the Receiver wishes to select that interpretation, I call these problems of interpretation resolution. The term resolution is preferred to other possible terms for this act, such as selection or simply interpretation, because it avoids asserting the primacy of the Receiver’s decision: the felicitous resolution of an interpretation involves both the actions of the Sender and the Receiver. I specify the common ground as a constituent of the formal definition of a game. This means that the common ground is accessible to players in their deliberative processes, and that it can have effects on the strategies that are selected in an IPG. For instance, interpretation resolution can be modelled as a choice between interpretations. The conditional probability of each interpretation given the utterance being interpreted and the common ground may be used to inform the Receiver’s selection of an interpretation. I show that this approach greatly enhances game theory’s ability to account for naturally-occurring speech data.

To recapitulate, common ground is viewed in this thesis as a record of the linguistic information that has been agreed upon for the purpose of the conversation. As we have seen, this is a role that DRSs are particularly suited for. For instance, in datum 12 on page 127 in §5.4, two names are erroneously given to a single character. Datum 12 is reproduced as datum 14 overleaf. Two name attributes are ascribed to B’s character. I noted that the assignment of two names to a single character appears to violate the internal consistency of
6.1 · Context and negotiation

Datum 14 Middle of 1-A2. (Reprint of datum 12.)

1 A: cheryl i love that you’ve learnt (.) i love that you’ve changed

2 B: it’s gonna be a very confused child

3 A: that’s how you grow

4 B: right

5 A: you grow THROUGH confusion

6 B: yes

7 A: you gotta keep on evolving (.) keep on changing (0.8)

CATCHing every curveball jessica

8 P: hahhh ((uncomfortable laughter from an audience member))

the plot-world. This seems to be a result of the violation of an assumption that a character may only have a single first name. Since the two name attributes do not occur in the one turn, or even in contiguous turns, the only way that we can account for the violation of this assumption is by employing a model that can act as a register of the linguistic information that is introduced; that is, a model that can encode interlocutors’ memory and its effects on linguistic choices.

Common ground was defined in definition 3.15 on page 72 as a DRS whose extension is nonempty. Let us incorporate common ground into the definition of an interpreted signalling game.

Definition 6.1: Interpreted signalling game with common ground. An interpreted signalling game with common ground is a tuple

\[
\langle N = \{S, R\}, L, \mathcal{M}_L = \langle W, \mathcal{D}, \mathcal{S} \rangle, \rho_S, \sigma = \langle W \times \mu \times \kappa, \mathcal{G}, ||\cdot||, \llbracket \cdot \rrbracket, U \rangle \rangle
\]

where \( N \) is the set of players with two players \( S \) and \( R \); \( L \) is a DRS language; \( \mathcal{M}_L = \langle W, \mathcal{D}, \mathcal{S} \rangle \) is a model for \( L \), where \( W \) is a set of worlds, \( \mathcal{D} \) is a set of individuals, and \( \mathcal{S} \) is an interpretation function per definition 3.10; \( \rho_S \) is the Sender’s belief; \( \sigma \) is the strategy set, whose elements are triples constituted of a \( w \in W \), a message \( m \in \mu \), and a DRS \( K \in \kappa \); \( \mathcal{G} \) is a set of all possible valid common grounds; \( ||\cdot|| \) and \( \llbracket \cdot \rrbracket \) are the semantic and extension functions, respectively; and \( U \) is a utility function.

Modelling the effects of common ground on the interpretation of an utterance is simple. An interpretation \( K \in \kappa \) given a common ground \( \Gamma \) is a conditional probability of \( K \) given \( m \) and the common ground \( \Gamma \):

\[
\Pr(K|m, \Gamma) = \frac{\Pr(m \cap \Gamma|K) \Pr(K)}{\Pr(m \cap \Gamma)}. \quad (6.3)
\]

This use of conditional probability is somewhat novel when compared to those that we have encountered up to this point, because the common ground is a DRS, which is a complex entity that is constituted of a universe and conditions on the universe, while worlds and utterances are atomistic. The interpretation of the probability of a common ground,
then, should be analogous to the interpretation of assessing the probability of a series of events, where each event is a predicate.

I should point out that this is a thoroughly subjectivist view of common ground. I take as my starting point for such a conception the way that a common ground is perceived by the agent herself, at a particular turn, having observed particular data. I contrast this with an 'analyst's-eye' view that considers the interaction in its entirety before making an effort to represent the common ground.

This implementation of common ground means that messages become more likely given certain common ground. We have already seen probabilistic effects of common ground information on interpretation resolution, although it was not couched in these terms. In datum 4 on page 112 in §5.2.1.2, which is set in a Mexican restaurant, the anaphoric expression of the first line (reproduced in (71)) is resolved only by using prototypical knowledge of a restaurant’s frame.

(71)

\[
A: ((\text{mimes moving a frying pan on a stove, speaks in a stereotypically Mexican accent}})) \text{ eh compadre: (\cdot a) ah they not coming in any more }
\]

Let us recast this anaphoric resolution in probabilistic terms. The common ground information that is available to an agent who observes only this utterance includes: the existence of a frying pan, posited by A’s mime; the fact that A is Mexican, posited by A’s accent and use of the term compadre; and that some set of individuals used to enter the location of the scene but no longer do, posited by the semantics of the utterance itself. A felicitous interpretation of the utterance takes all of these facts into account. If we consider a range of possible interpretations, the most felicitous interpretation is the most probable given these three facts. Specifically in the case of the set of individuals that resolves the anaphoric expression they, the common ground renders the most likely interpretation of they a set of patrons of the restaurant. As I showed in §5.2.1.2, this is precisely the interpretation that becomes operational for the remainder of the scene.

Each of the linguistic devices used to construct plot-worlds that were described in the previous chapter can be recast as problems of probabilistic interpretation resolution. As I have suggested with the reanalysis of (71), and as I show in this chapter, a probabilistic model of interpretation is able to explain any interpretation problem that presents itself to an actor, and does so in a way that is commensurate both with game theory and with DRT. It allows for the linguistic information in the context to bear on the decision making of agents in game theory, and does so rigorously.

6.1.2.1 · Belief and the common ground

The common ground is therefore conceived in the model as a DRS that records the linguistic information that is introduced iteratively during the course of a conversation. Definition 6.1 introduces the common ground as a basic entity of interpretation games, which allows it to be used explicitly in game models. The primary implementation of the effects of common ground on the deliberation is on the interlocutors’ beliefs. There are three
beliefs that interest us: (a) the Sender’s belief; (b) the Receiver’s prior belief; and (c) the Receiver’s posterior belief. In the following paragraphs, I update the definition of each of these kinds of belief to include common ground.

The Sender’s belief. To recapitulate, the Sender’s belief describes the Sender’s expectation of how probable the Receiver is to play a certain move. In signalling games, this is a probability distribution over the Receiver’s strategy set. It is not a conditional probability in standard game theory. What the belief represents is the Sender’s expectation of the Receiver’s likely course of action before the Receiver observes the Sender’s utterance. If the Sender estimates the probability with which the Receiver will select a certain strategy, she can use it as a basis for the calculation of how these probabilities will change when the Receiver observes the utterance.

The inclusion of common ground complicates matter. As I have suggested above, the common ground makes certain utterances more salient than others. And therefore, the Sender’s belief describes her estimation of how much of an effect the common ground will have on the Receiver’s interpretation, even before the Sender produces an utterance. That is, the Sender’s belief is the Sender’s guess of the Receiver’s expectations for the direction of the conversation. Let us define this formally.

**Definition 6.2: Sender’s belief with common ground** \( \Gamma \). Let \( \kappa \) be a set of interpretations for the Receiver \( R \), and let \( G \) be a set of common ground. The Sender’s belief with common ground \( \Gamma \) is a probability distribution \( \rho_S \equiv \delta_{\kappa \mid G} \in \Delta(\kappa \mid G) \) with full support (that is, every \( \Pr_S(\kappa \mid G) \in \rho_S \) is nonzero).

It appears very odd to discuss expectations of an interpretation selection before an utterance has been produced. This seems to violate a basic fact of causality in conversation: the Sender cannot interpret an utterance that she has not yet observed. But this is not how it should be seen. This is instead the game theoretic implementation of the Sender’s anticipation of the Receiver’s expectations regarding the utterance, given the common ground that has been established. That is, if the Sender believes that certain interpretations are more likely given the common ground, then she guesses that the Receiver expects her utterance to be salient to those interpretations.

The Receiver’s prior belief. In interpretation games with common ground, the Receiver’s prior belief describes the Receiver’s judgment of the probability of each of the possible world-states given the common ground that has been established. In standard signalling games, the Receiver’s prior belief expresses her estimate of how probable a specific world is before she observes the utterance. The formal expression of Receiver prior beliefs in standard signalling games is a simple probability distribution with full support over the set of world-states. With the inclusion of common ground, this probability distribution is conditionalised on the common ground. The result in definition 6.3 is very similar to definition 6.2.
Definition 6.3: Receiver prior belief with common ground $\Gamma$. Let $W = \{w_1, \ldots, w_m\}$ be a set of world-states, $\kappa$ be a set of interpretations, and $G$ be a set of common ground. A prior belief of $R$ is a probability distribution $\rho_R(w|\Gamma) = \delta_{W\mid G} \in \Delta(W\mid G)$.

The Receiver’s posterior belief. Once the Receiver has observed the utterance, she updates her belief to incorporate the addition of information provided by the utterance. That is, the utterance provides the Receiver with some evidence of the world-state that the Sender has observed. The posterior belief quantifies how the Receiver understands this evidence.

Definition 6.4: Receiver’s posterior belief with common ground $\Gamma$. Let $W$ be a set of states of the world, $\rho_R(w|\Gamma)$ be a prior belief of $R$, and $m$ a message sent by $S$. A posterior belief of $R$ is a fully supported probability distribution $\rho_R(w|m, \Gamma) = \delta_{W\mid m \cap G} \in \Delta(W\mid m \cap G)$ such that each member of $\delta_{W\mid m \cap G}$ is calculated by

$$\Pr(w|m \cap \Gamma) \equiv \frac{\Pr(m \cap \Gamma|w) \Pr(w)}{\Pr(m \cap \Gamma)}.$$  

(6.4)

These redefinitions are perhaps the simplest method of incorporating common ground into a game theoretic model of language use. The mechanism provided by these definitions gives us a way of explicitly modelling the effects that a changing common ground can have on how an agent designs and responds to an utterance, and in addition, it captures the narrowing of probable interpretations as the common ground becomes more specified.

6.1.2.2 · Misunderstandings

A consequence of these definitions of belief is that divergences in the perception of the common ground are likely to lead to misinterpretations. If a Sender believes that some information is in the common ground, but it is not, then the Receiver is unlikely to correctly resolve an interpretation of an utterance whose interpretation depends on that information. Consider the following example to highlight this point.

Example 6.1: Verdi versus Wagner. Vladimir and Estragon arrange to meet up and see an opera. There are two operas on that night: Rigoletto by Verdi, and Parsifal by Wagner. Neither can decide which to see, so they agree to come to a decision at dinner before the show. But disaster strikes! Vladimir falls ill, and is laid up in bed on the night they were supposed to meet. A few days later, Vladimir runs into Estragon. Estragon says, “I saw the opera, it was terrific!” Vladimir does not know which opera he means, since there is nothing concrete in their conversation to suggest either is more probable.

This example describes a situation where Estragon believes (for whatever reason) that his choice of which opera to attend is in the common ground, while Vladimir does not believe the same.

Let us model it game theoretically. Estragon and Vladimir are $S$ and $R$, respectively. There are two world-states—$w_v$ for the Verdi opera and $w_w$ for the Wagner opera—that only $S$ can observe. There is only one message, $m$, which Estragon plays deterministically. This is meant to model Estragon’s assumption that he does not need to alter the content of
the utterance because the common ground constrains the possible entities that may resolve the definite noun phrase *the opera*. *R* may select from two possible interpretations: *K_v* and *K_w*, in which the Verdi opera or the Wagner opera resolves the anaphora, respectively.

So far, then, this is a typical interpreted signalling game without any common ground. Assuming a Lewis utility function, the tabular representation of this game is given in figure 6.2.

<table>
<thead>
<tr>
<th><em>w ∈ W</em></th>
<th><em>Pr_R(w)</em></th>
<th><em>m</em></th>
<th><em>(u_S, u_R)</em></th>
</tr>
</thead>
<tbody>
<tr>
<td><em>w_v</em></td>
<td>1/2</td>
<td>∈</td>
<td>(1, 1)</td>
</tr>
<tr>
<td><em>w_w</em></td>
<td>1/2</td>
<td>∈</td>
<td>(-1, -1)</td>
</tr>
</tbody>
</table>

*Figure 6.2:* A tabular representation of the Verdi versus Wagner signalling game without any common ground.

Figure 6.2 expresses the total uncertainty that Vladimir feels. He is unsure of the world-state, and so *Pr_R(w_v)* is exactly equivalent to *Pr_S(w_w)*. Furthermore, Estragon’s utterance provides no evidence of the value of *w*, because it could conceivably refer to either opera. So, both worlds are in *[[K_m]]*. This game is strategically intractable for Vladimir, and his best solution is to play either *K_v* or *K_w* each with probability 0.5.

It appears that Estragon has made a grave error of strategy selection. However, the addition of common ground alters the understanding of the game. Because Estragon believes that the identity of the correct opera is part of the common ground, he believes that Vladimir will be able to felicitously resolve the anaphora. And so Estragon’s choice can be justified.

Interlocutors may become confused regarding the precise constitution of the common ground. They may lack sufficient information to correctly identify or distinguish some part of the common ground from other potential parts of the same kind. In example 6.1, Estragon knows which opera he is referring to, while Vladimir does not. Since there are two candidates that can plausibly bind to the definite noun phrase, Vladimir becomes uncertain, as a result of the asymmetry of information between him and Estragon, of what the true common ground is. The uncertainty of common ground that an agent faces, then, is analogous to the uncertainty that arises in any signalling game from a Receiver’s lack of knowledge regarding the state of the world. In signalling games, the Sender must design her utterance so that this uncertainty is diminished as much as possible.

The same holds for signalling games with common ground, though the considerations of a Sender are necessarily more complex. The Sender must select a strategy that communicates the world-state, and to do so she must take into account the common ground that has been established until that point, as expressed by the Sender’s belief. Likewise, when the Receiver selects an interpretation, she is influenced both by the common ground—as
expressed by her prior belief—, and by her judgment of the motives of the Sender in selecting a certain message given the common ground—as expressed by her posterior belief.

Now that I have established a game theoretic means of representing common ground, let us turn to the primary means by which common ground is constructed in improvised theatre: negotiation.

6.1.3 · Negotiation

It is clear that in any given conversation (a term that excludes monologues, speeches, and so on), multiple interlocutors contribute information to the common ground of that conversation. This is a simple consequence of how conversations operate: utterances introduce linguistic information into the common ground, and all interlocutors produce utterances. I therefore refer to the process of building the common ground as a process of negotiation. Negotiation in this sense refers to an iterative process by which the inputs of multiple agents are taken into account when determining the true constitution of the common ground.

The notion that the form of the common ground might be negotiated appears intuitive. The speech acts of multiple interlocutors contribute to the common ground introducing information, or by contesting what has already been introduced. But in this subsection, I lay the groundwork to make a more radical claim in §6.2: that the interpretation of any utterance arises through a process of negotiation.

The idea of negotiation in improvised theatre is certainly not a novel idea. I have stated my agreement with Baumer & Magerko (2010) that the basic unit of analysis of improvised theatre ought to be the offer–response pair. Adapting their designation for the current purpose, I propose that the offer–response pair be seen as the basic unit of negotiation. Any negotiation is made up of a suggestion for the direction of the narrative, and a subsequent judgment of this suggestion. This is in line with a game theoretic framework, where the most basic signalling game includes both the actions of a Sender and a Receiver; and perhaps more importantly, where the understanding of each player of the meaning of an utterance has to be signalled to the other player.

But it also agrees with a paradigm that has recently experienced a surge in popularity in linguistics, and especially politeness studies (see, for instance, Arundale 2010, Haugh 2010). This ‘negotiative’ paradigm views the Sender–Receiver dyad as the basic unit of interaction, and argues that the meaning of any utterance is determined only by agreement of the interlocutors, through a process of negotiation. This does not exclude utterances from having a composed meaning per se. However, proponents of this paradigm argue that often utterance meanings are not purely the result of a compositionally semantic mechanism, and that any rigorously valid account of meaning must take into account meanings that arise from interlocutors’ joint achievement. Examples of theoretic frameworks which adhere to this broad description include: the ‘symbolic interactionalism’ of Goffman (1955, 1967); the ‘ethnography of speaking approach’ developed by Hymes (1962, 1963, 1976); Silverstein’s (1976, 1993, 2001) ‘metapragmatic’ approach; the ‘joint action’ account of Clark
(1997); the ‘sequential information paradigm’ of Cooren and Sanders (Sanders 1987, Cooren 2000, Cooren & Sanders 2002); and, most recently, the ‘conjoining co-constituting model’ devised by Arundale (Arundale 2006, 2010, 1999, 2008).

It is worth examining how this process of negotiation might operate in a linguistic setting. I base what follows mostly on the work of Arundale, for a number of reasons. First, his work informs an increasingly large array of research in linguistic pragmatics: see, for instance, recent work in studies of linguistic intention related to Griceanism (Heritage 1990, Haugh 2009), recent research in face theory (Bargiela-Chiappini & Haugh 2009, Bargiela-Chiappini 2003, Grainger, Mills & Sibanda 2010, Haugh 2010, Spencer-Oatey 2007), politeness theory (Bargiela-Chiappini & Harris 2006, Bargiela-Chiappini & Kádár 2010, Eelen 2001, Haugh 2003, Mitchell & Haugh 2015, Haugh 2004), and ethnographic studies of communication in languages other than English (Samra-Fredericks 2010, Haugh & Watanabe 2009, Haugh 2007, 2004). Second, Arundale’s descriptions are very perspicuous, and his formulations are designed to be as generalised and generalisable as possible. Finally, as we will see, his model’s conception of interaction integrates well with already existing game theoretic concepts.

Arundale (2010: 2080) provides a detailed template for how a segment of conversation is analysed under these conditions. Emphasis is the author’s.²

As [an interlocutor] Amy designs and produces her first position utterance, she anticipates [a second interlocutor] Bob’s interpreting of it. Her utterance, like any other, affords and constrains the possibilities for interpreting meaning and conversational action. Because she is cognitively autonomous from Bob, because meanings are not determinate (Arundale 2008: 244), and because (unlike an analyst) she cannot know what Bob’s second position utterance will be, Amy does not know just how Bob has interpreted her utterance until she interprets his second position utterance responsive to it […] Prior to that point, Amy’s projecting of Bob’s interpreting of her first position utterance remains provisional, because she does not know how it will be understood within the frame of their particular interaction […] In producing his second position utterance, Bob provides evidence of some aspects of his interpreting of Amy’s first utterance. Her interpreting of that evidence allows her either to confirm her provisional projecting if that interpreting appears consistent with her projecting, or to modify her provisional projecting if she finds inconsistency. At the point she has interpreted Bob’s second position utterance, Amy’s projecting of Bob’s interpreting of her first utterance, or her modification of it, becomes an operative interpreting because she now has evidence for how it has been understood within their particular conversation. If Amy is to add a third position utterance, she must in some manner take this operative interpreting into account, whether or not it is consistent with her own initial projecting. A provisional interpreting, then, is one not yet assessed in

²Two notes on Arundale’s particular vocabulary. Arundale mentions ordinal position utterances a number of times; I would hyphenate these, for instance first-position, since first position utterance makes it seem as if a position utterance is a unit of the analysis. It is not, and Arundale is here describing the utterance that occupies the first position of the segment being analysed. Second, Arundale prefers gerund forms of verbs to emphasise their status as ongoing processes, rather than fixed discrete events. This is not important to my analysis, and I do not adhere to this convention in the rest of the thesis.
view of uptake, even though one may be certain about it, while an operative interpreting is one assessed in view of uptake, even though one might change it.

What is important here is that a range of interpretations are projected by the first speaker’s utterance, and, while the first speaker might have a specific interpretation in mind, the addressee potentially selects any possible interpretation, even if the first speaker does not consider that interpretation a possibility. To demonstrate this, Arundale (2006: 196) points out that a number of interpretations are available to the utterance in (72).

(72) That’s a nice jumper.

The speaker could be simply complimenting the jumper; or she could be asking to borrow it; or she could be making a comment about the temperature; and so on. The point is that none of these interpretations is per se privileged above the others at the level of the conjoint constitution of the conversation. The speaker of (72) certainly had a particular interpretation in mind when she produced the utterance, but the interpretation of the utterance that becomes operative in the conversation is only determined by the addressee’s response.

I do not claim to include Arundale’s formalisms in the model that I propose here. Arundale’s model is largely a reaction to and a refutation of Gricean notions of intention, and I remain agnostic on such questions in this thesis. It is important to understand that these ideas underpin what follows, and that I will describe the linguistic strategies of actors only insofar as they are contributions to the conjoint constitution of the discourse of the plot. And, as I show, these ideas become an indispensable part of the fabric of the model in its entirety when the model culminates in the following chapter.

I do, however, wish to bring a basic fact about Arundale’s model to the foreground of what I propose here. Underlying the sequence in Arundale’s description of the interpretive act is the fact that the number of possible operative interpretations of the first-position utterance is diminished when the second-position utterance is produced, and likewise following the third. The plausible interpretations of an utterance change depending on the utterances that have been produced in the discourse until the point when $m$ is produced. I have already shown that this phenomenon is present in DRT: as the number of conditions in a DRS grows, the extension of that DRS shrinks. This parallel suggests that the two theoretical approaches are compatible.

To categorise interpretations, I use two terms from the work of Arundale (2010, 2008): provisional and operative. A provisional interpretation has been projected by an interlocutor, but has not been agreed upon for the purposes of the conversation. Interpretations remain provisional until they are agreed upon, whereupon they become operative. The term agreement is taken as it is in the definition of acceptance in common ground given by Stalnaker (2002): if a particular interpretation is provisional, interlocutors treat that interpretation as accepted for the purposes of the conversation. As with acceptance, this does not require interlocutors to commit to the truth of the interpretation; interlocutors
may of course interpret a statement and yet know that it is false. Instead, interlocutors commit to the notion that a particular interpretation is accepted by all parties.

This distinction is something that I would like to preserve, not only because it has a good deal of explanatory power for certain strategic selections that appear in the data, but also as it is often useful to have a means of distinguishing interpretations that agents propose (i.e., those that are provisional) and those that display a meaningful effect on the strategies that actors employ (i.e., those that are operational).

It is important to note that the characterisations of operative and provisional interpretations that I have given above are not consistent with Arundale’s conception of the categories. According to Arundale (1999), the metaphor of an interpretation Receivers ‘passing’ or ‘graduating’ from a provisional to an operative interpretation does not hold, since it implies that a particular interpretation has a privileged status. He argues that this metaphor implies that a specific interpretation has a privileged position. Instead, he argues that “a first speaker’s utterance affords a certain range of interpretings, but does not determine which one of these interpretings will be operative in the conversation” (Arundale 2006: 196).

I disagree. The formal apparatus of game theory that has been developed thus far provides for a range of interpretations of a given utterance. Certain of these interpretations may be preferred by the Sender, but they may have a zero probability. Likewise, the Receiver may select any of the available interpretations, provided that it is rational according to the information available to her to do so. Indeed, a Receiver may select an interpretation that appears impossible to the Sender. Players can therefore have preferences over interpretations, while not privileging any preferences at the level of the model.

The centrality of negotiation both to the interpretation of an utterance, and more broadly to information that is included in the common ground, is manifested in the model chiefly by definitions that rely heavily on conditional probabilities. What results is a model that foregrounds agents’ process of inferring facts based on information that is presented to them. An emphasis on inferential reasoning is appropriate for modelling situations of linguistic interaction in general, for, in conversation, each interlocutor must attempt to infer the motives of her partners based solely on the utterances that these produce. But it is especially appropriate for game theoretic models of language use, for game theory is particularly apt at descriptions of inference, especially where players lack information. This will be shown in the section that follows, in which I describe the strategies that actors use to overcome the problems introduced by this lack of information.

6.2 · Strategies for resolving interpretation
The effect of common ground on the deliberation of agents is a complex matter in isolation, even before agents’ preferences have been factored into the model. We now move towards an account of agents’ preferences. In general, the signalling games that we have examined have had agents who are interested in correctly communicating and in inferring the state of the world. With the introduction of a formal mechanism of common ground, the foci
of the players shift. Felicitously conveying the world-state requires the Sender to also consider the common ground. Likewise, felicitous resolution involves recognising how the common ground is being considered by the Sender. It is primarily the control of the information introduced into the conversation, and the emphasising of certain parts thereof, that allows agents in signalling games with common ground to maximise the possibility of felicitous resolutions. In this section, I present some data that display methods that actors tend to use to promote felicitous resolution of interpretations. Similar to the last chapter, I use these observations as a basis for proposing a number of heuristics that actors use to select their strategies in §6.3.

There are two major functions of these strategies that I wish to explore. The functions are both kinds of interpretation, but differ with respect to their targets. In §6.2.1, I explore the use of strategies to resolve the interpretation of utterances, where the resolution of an utterance relies on an operative interpretation of the term being agreed upon. In §6.2.2, I focus on strategies that actors use to signal, and resolve the interpretation of, their preferences over plot trajectories, and the narratological functions of these strategies.

6.2.1 · Negotiating operative interpretations

In this subsection, I examine data that exhibit how actors arrive at operative interpretations, and identify how prosody and repetition contribute to this end. First, let us consider datum 15.

Datum 15 Middle of 2-A2.

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>B</td>
<td>&gt;sorry&lt; something about twirlin a gun just makes me feel alive man</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>(0.2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>naw i get that (1.7) back east (0.4) wasn't the same</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>(0.7)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>B</td>
<td>no (0.4) it was frowned up:wn (0.2) you couldn't even twirl a toothpick with your tongue in your mouth</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>A</td>
<td>no (0.9) they hated that</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Some common ground is required to understand the resolution of interpretation. Throughout the scene, A continuously makes reference to the general situation of being in the American West, particularly in relation to the comforts and safety of the characters’ homes in the East. B meanwhile twirls his pistol throughout the scene, and repeatedly provides justifications for doing so, in spite of the protestations of A.

It appears, then, that the actors’ intentions for the plot trajectory differ. A’s utterances suggest that he has in mind a plot that pertains to their reasons for making the trip to the West. B’s preferred plot trajectory appears to involve conflict provoked by his twirling of his pistol: practically all of his lines, except the first, conspire to suggest this is his preference. With this conflict of preferences in mind, let us examine the lines that are included in datum 15, especially the phrase uttered by A in line 3:

(73) A: naw i get that (1.7) back east (0.4) wasn't the same
The implicit subject, *it*, has a range of possible interpretations. Perhaps the two most salient given the noted conflict in the actors’ preferences are

(i) the situation at the characters’ place of origin differs from their current situation, a discussion of which is part of A’s preferred plot trajectory, which I denote \( K_a \); and

(ii) the act of pistol twirling, which appears to form the basis of a conflict that is part of B’s preferred plot trajectory, which I denote \( K_b \).

In line 5, B provides strong evidence that he selects interpretation \( K_b \), by using the pronoun *it* to denote the act of pistol twirling, and suggesting that pistol twirling was frowned upon. \( K_b \), it appears, becomes provisional. If we ignore the redirection strategy employed by B (for I return to it below), A’s utterance in line 6 appears to confirm B’s provisional interpretation, and so \( K_b \) becomes the operational interpretation of A’s utterance in line 6. This is an instance of interpretation resolution that precisely follows Arundale’s schema: A projects a particular interpretation, B provides evidence for an interpretation, and A abandons his preferred plot trajectory by confirming that B’s interpretation is the operative interpretation for the remainder of the scene.

This process of negotiation can be represented using DRSs. The following is an approximation of (73) that concentrates on the disputed interpretation of *it*.

\[
K_{(73)} = \begin{array}{c}
    x \\
    \neg \text{the same}(x)
\end{array}
\]

Since \( \text{the same}(x) \) implies a comparison, *it* is underspecified, and the set of individuals in the universe of the common ground that it may bind to is very large. This allows for multiple interpretations of *it* to be plausibly projected. The two interpretations that are proposed by the actors are approximately (and somewhat inelegantly) represented as follows.

\[
a. \ K_a = \begin{array}{c}
    x \\
    \text{the general situation}(x) \\
    \neg \text{the same}(x)
\end{array}
\]

\[
b. \ K_b = \begin{array}{c}
    x \\
    \text{pistol-twirling}(x) \\
    \neg \text{the same}(x)
\end{array}
\]

The existence of these two provisional interpretations persists until A’s utterance in line 6, when \( K_b \) is merged with the common ground. And so, in subsequent turns of the dialogue, *it* refers to pistol-twirling, rather than the general situation.

This is a very simple example of the negotiation of an operative interpretation, since it is the words in the utterance themselves, rather than strategies like prosody or sociolin-
guistic register, that are used to reach an operative interpretation. In fact, it appears that even without the premise of the conflict between the two actors, the three utterances in lines 3, 5, and 6 would still function as an instantiation of this principle, since the meaning of \textit{it} is underspecified at line 3, and then its meaning is determined by B’s response in line 5. Using the words of the utterance only, B is able to prevent A from introducing a reinterpretation that might better suit A’s preferences. That is, suggesting that the general situation in the West can be frowned upon back in the East is not as semantically felicitous, while suggesting that pistol twirling was frowned upon is more semantically cogent. By this strategy, B makes it practically certain that $K_b$ will come to be the operative interpretation, because the probabilities of any other interpretation are greatly diminished. There are naturally effects from other strategies, such as the use of prosody in A’s second utterance in line 20: the emphasis of \textit{that} is not necessarily a product of natural stress patterns (for \textit{hated} is also stressed), but instead appears to be used to signal that A accepts the interpretation that B has projected with his utterance. But these do not appear to be particularly marked, and are not when compared to other uses of prosody, at least as they operate at the level of negotiating an operative interpretation. We will see that they play an important role in the negotiation of plot trajectories in the following subsection.

Datum 16 overleaf is a longer and more complicated instance of negotiating an operative interpretation. The factors that contribute to the datum’s complication are twofold. First, the interpretation in question is, unlike datum 15, not of a term, but is instead of an offer. That is, it is the intention that underlies an actor’s eliciting of a response from another actor that is at issue. Second, there is no felicitous resolution. Instead, one of the actors attempts to forcibly resolve the interpretation of the offer in question to suit his plot trajectory, while attempting to conform to the norms of improvised theatre. By these norms, he must not block others’ suggestions outright, but incorporate them into the plot (see Spolin 1963). Note that, while there are many participants in datum 16, we are most interested in A and C, since they perform the greatest part of the negotiation.

A general synopsis for the relevant characters of the scene follows. A proposes in turn 4 that something about the zoo is amiss (‘hey, have you ever thought, I mean ...’). This offer is taken up by C in lines 6 and 9, wherein C proposes that the animals in the zoo’s exhibits are not real, but are mere projections of animals. A disputes this assertion in line 11 (‘or …’) by asserting that the actors themselves are in a zoo in line 14 (‘what if we’re in the zoo, man’). C attempts to reconcile these two theories by asking if the human characters A and C are projections. In a thinly-veiled metanarrative move, A attempts, basically for the remainder of the scene, to repeatedly establish why C has proposed this, in lines 26, 29, 32, 40, 45, and 52. The purpose of this series of repetitions of the strategy is clear: A wishes C to conform with his desired plot trajectory.

It appears that A has, at the beginning of the scene, a preferred plot trajectory: one in which the humans are exhibited in the zoo. It seems, then, that he anticipated that his utterance in line 4 would be clear enough a suggestion that C would be able to follow A’s preferred trajectory. But this does not occur: A’s offer is vague, and does not provide any
6.2 · Strategies for resolving interpretation

**Datum 16** Most of 3-B2, from the beginning. Less salient parts of the example (lines 46–52) have been redacted.

1. A  | o look at that (0.6) pænda (1.0) .t aw look at those
deg:i i (0.5) i LOVE the zoo ma:n (. ) the zoo just really
(0.6) DOES somethin for me ma:n i mean like
(2.1)
2. B  | [ca ca:u ] ((i.e., a bird call))
3. A  | [hey have you th-] ever THOU:GHT (. ) >°i mean°<
4. (0.8)
5. C  | yes:h (1.3) i mean (. ) what if they're not REAL (. ) that's
what you're getting at [isn't it what ]=
6. P  | [ha ha ha ha ha ha]
7. B  | [ca ca:u ]
8. C  | not really there behind the bars it's just like a (0.6)
<projection> or something
9. B  | ca ca:u=
10. A  | =0:R
11. (0.5)
12. P  | XXXXX
13. A  | what'f <WE:'RE> in the zoo [ma:n ]
14. P  | [ha ha ha] xxx
15. B  | ca ca:u
16. (0.8)
17. C  | what d'ya mean like we're [(0.4) proJEctions ]
18. D  | [PPPPPPPPPPPPPPPPPPPPPPPPP] ((i.e.,
elephant trumpet noise))
19. P  | XXXXXXXX
20. B  | ca ca:u
21. P  | [###
22. C  | [(of the) ZOO?] (1.0) ca ca:u ca ca:u
23. D  | PPPPPP
24. P  | XXXxx
Chapter 6: Resolving interpretations with divergent common grounds

26 A  NO MAN NO (0.2) that's more [evidence] for my idea
27 B                                  [caiy ]
28 P  ha ha ha
29 A  i don't think we're projec[tions ]
30 B                                  [ca caiy ]
31 P  ha ha ha xxx
32 A  WHY're you so convinced that we're proj[ections]
33 E  [did you ] see that
      elephant and bird couple? (. ) that's disgusting
34 P  XXXXXX#X#X#
35 C  can you hear the penguins talking?
36 F  that was the most disgusting thing i've see:n
37 E  it's ruined my day at the [ZOO:] i'd say
38 F  [YES ]
39 C  XXXXX
40 A  CLINT
41 C  kind of bigoted penguins
42 A  CLINT, i can definititely hear that the penguins are
      talking in a biggoted way (. ) WHY DO YOU THINK WE'RE
      PRO↑J̲E̲C̲T̲IONS?
      (0.4)
43 C  sh- i ca:n't SPEAK penguin
44 P  xxxxXXx
45 A  YEAH BUT WHY DOES THAT MEAN WE'RE PROJECT-I MEAN I'M HAVIN
      A BREAKTHROUGH ABOUT OUR LIVES 'N YOU'RE DOIN A DIFFERENT
      BREAKTHROUGH ABOUT OUR LIVES AND I'M TERRIFIED

52 A  so I'M NINETY NINE PERCENT SURE THAT £WE'RE IN A ZOO RIGHT
      NOW
indication of his preferences beyond the idea that there is something about the zoo that is strange. And so C proposes something strange about the zoo, namely that the animals in the zoo are false, which is an outcome that is not aligned with the preferences of A. A’s repeated appeals to the evidentiary basis of his claim later in the scene are performed in an attempt to return the plot to his preferred trajectory. But more fundamentally, what A’s strategies appear to be trying to accomplish is the setting of a provisional interpretation that differs from the interpretation that the evidence suggests C selected. While C continues to advance the plot using the logic of the plot-world that has been established, A repeatedly employs strategies that suggest that a particular plot trajectory is preferable. The actual strategy that A employs is unsubtle: he simply repeats the term projection. The most egregious example of this is line 45, reproduced as (76).

A: YEAH BUT WHY DOES THAT MEAN WE’RE PROJECT-I MEAN I’M HAVIN A BREAKTHROUGH ABOUT OUR LIVES ‘N YOU’RE DOIN A DIFFERENT BREAKTHROUGH ABOUT OUR LIVES AND I’M TERRIFIED

The narrative premise is that the character is confused about the basis on which the character C makes his claim. The motivation appears similar, except that there is a degree of competition between the competing plot trajectories: A essentially demands that C explain why his preferred plot trajectory is preferable to A’s.

The most salient strategy that A deploys in datum 16 is the use of repetition. He insists on repeating utterances that signal his preferences a number of times, over the course of a long scene. The rationale for the use of repetition is intuitively simple: by increasing the amount of evidence that an agent observes, such that the evidence supports a particular hypothesis, the likelihood that the agent will select that hypothesis is increased. So, the more evidence that A provides to C, the more likely C is to resolve A’s original offer in a certain way. The error that A appears to commit is that he assumes that C will comply with his preferences. Meanwhile, however, C appears disinterested in arguing the operative interpretation of the first offer, and attempts to advance the plot of the scene by interacting with the other characters that are in the scene. A clear indication of this is in line 35:

C: can you hear the penguins talking?

This line responds to E’s utterance in line 33:

E: [did you ] s:ee that elephant and bird couple? (.)
that’s disgusting

The irony of the situation is that the other actors do in fact treat A’s utterance as operative. The plot that is constructed is indeed one where the human characters are exhibited animals in a zoo, and where the animals act as if they are attending the zoo, as becomes apparent in line 37, where the animal played by actor E complains about his day at the zoo being ruined by seeing an interspecies elephant–bird couple. Indeed, even C, far from dissenting, appears content to proceed with this plot trajectory. In line 35, C asks A if he can hear the penguins talking, explicitly referring to E’s utterance in line 33. This
suggests that C accepts the plot trajectory that is premised on the humans being exhibited at the zoo, with the animals acting as patrons, for E’s utterance makes sense iff this is the premise. C abandons his projection plot trajectory entirely, and does not mention it again.

In strategic terms, there are some explanations for A’s apparent failure to notice this fact. First, C’s acceptance of A’s interpretation of the initial offer may have been sub-optimal, in that it does not convey with sufficient clarity that he considers A’s preferred interpretation operative. This is plausible, since C’s response in line 18 does respond to A’s initial offer as an affirmation, but it also redirects the plot trajectory towards an integration of their ideas. Second, it could be a simple failure of awareness on A’s part in that he does not recognise that his interpretation is treated as operative by the other players.

These two instances share a basic intuition about strategies for resolving an interpretation so that it is operational: Sender strategies that provide a greater quantity and a higher quality of information to a Receiver are more likely to result in the Receiver correctly inferring what the Sender attempts to convey. Increasing the quantity of information—that is, the number of relevant utterances—increases the data that the Receiver uses to inform her decision. Likewise, selecting utterances of higher quality—those that less ambiguously mark the intended interpretation—will result in a similar increase in the conditional probability that the Receiver will infer the correct state. This intuition is codified in standard game theory, but we have seen that it is even more salient when players play repeated signalling games, as in this datum. Likewise, attempts to increase the likelihood of successful communication are found throughout the data. We can view instances in data 15 and 16 presented above, but it is also visible in the data presented in the previous chapter, and will be visible in those in the following chapter. Thus, I wish to emphasise that the process of reaching an operative interpretation underlies the discussions that follow of operations on the state of the scene, and of plot trajectories. I provide a formal mechanism for this intuition in §6.3.1.

6.2.2 · Operations on the state of the scene

A common topic of the analyses of data that I have presented is the question of how actors achieve the construction and construal of a plot. In general terms, the most important means of achieving this is by signalling their intentions over the plot trajectory. Signalling intentions over the plot trajectory involves the use of strategies that are used to control the state of the scene.

Recall from §4.1.2.3 that these strategies are divided into four categories. The inclusion of a particular response strategy is included in a category based on its polarity, and on whether it adheres to the plot trajectory that the offer projects. Table 6.1 overleaf, which summarises the categories, is a reproduction of table 4.1. According to Baumer & Magerko (2009), who devised these categories, the actions of actors suggest that these types of response, as well the more basic fact that scenes proceed through a series of states, are known to actors. I provided formal definitions that integrate these categories with DRT and the epistemological apparatus of game theory in §4.2.3. It is with this formal apparatus,
6.2 · Strategies for resolving interpretation

Almost every strategy that is part of an actor’s linguistic faculty is involved in resolving interpretations. For instance, I showed in §5.2.1 that actors use presupposition to resolve anaphora that would otherwise fail to be resolved. Although this was not couched in the vocabulary of interpretation resolution, the examples that I examined involved actors introducing and using common ground information to cause certain interpretation to be more probably selected. All of the data that have been so far discussed exhibit multiple strategies for interpretation resolution.

Since strategies for interpretation resolution are so pervasive, then, I do not aim for an exhaustive category of such strategies. Instead, I examine two devices that are relatively simple: the use of prosody, and the use of repetition. These are selected because together they represent different levels of operation on the discourse. Prosodic patterns can be employed in the context of a single utterance, while repetition requires a number of turns to deploy. Restricting our analysis to these two devices allows us to examine actors’ motives more clearly, both at the level of individual utterances, as well as across many turns. Furthermore, when compared with the setting of operative interpretations, interpreting plot trajectories is closer to the well-studied problem that agents face in classical signalling games, since it is a theatre-specific form of a Receiver inferring the Sender’s intention from the observation of an utterance.

6.2.2.1 · Prosody

First, let us consider datum 17 overleaf, which displays many instances of the use of prosody to signal an affirmative augmentation.

Some common ground is required to understand this portion of the scene. Before this moment, A and B are flirting in what appears to be their workplace, some sort of scientific laboratory, where A is transferring liquids to a beaker. At the beginning of the scene, it is established that B has been put on probation by her immediate superior, a certain Jacobs. In datum 17, the actors talk about a male who is not on the stage (see the uses of he in lines 3, 4, 5, and 7); this is Jacobs. Not only has Jacobs put B on probation, but there has been some romantic link between him and B, which A is jealous of.

The actors use stress throughout this datum to signal how they intend for their utterances to be interpreted. When, in line 1, A offers for B to take hold of the beaker, he stresses this. It appears that he means to signal his intention that, in his preferred plot trajectory,
Chapter 6: Resolving interpretations with divergent common grounds

Datum 17 Middle of t-A3.

1  A: NO by not doing take (.) take the reigns of this alright ((B takes the beaker from A)) (0.5) it's DELicate (0.2) but i trust you (0.5) you got SPA:RK
2  (0.7)
3  B: 's what he said hm hm
4  A: yeah look ( ) [he's a dick ] i get it
5  B: [although he put me on probation]
6  P: >heh heh heh<
7  A: he's got ru:les (.) and that's not what science is about it's not about rules it's about fun that was ↑fun
8  (0.8)
9  B: i love these ↑da::tes:
10 P: hah hah
11 (4.0)
12 B: ((the experiment fails to produce a result)) nothing happened=
13 A: =a:h sh:it (2.7) alright we can't see each other again i really- i thought that was gonna (0.2) i thought there was gonna be an exPLo:sion (0.2) i thought you were gonna be exCITed (.) now you HATE me i'm BO:Ring
14 B i don't HATE you

B take over the experiment from him. A complies with this request. Similarly, in line 4, A emphasises that he is a dick; and A’s feelings about him become important later in the scene. In both of these examples, it appears that the interpretation that is intended by A is resolved correctly by B, for the plot trajectory continues as A suggests.

The rate of speech is another aspect of prosody that is exploited in this datum, usually to suggest that a character is of a certain state of mind. For instance, the lengthening of dates in line 9 expresses excitement. It appears that this information is used in A’s performance of a reaction to the non-result of the experiment that is undertaken in the scene. Likewise, A’s excitement is expressed by the speed of the utterance in line 7; note the lack of even short pauses. We see rate of speech and stress combined in line 13. A long, 2.7-second pause leads to an outburst of disappointment. The words that are stressed suggest a plot trajectory set in the aftermath of a failed experiment that deals primarily with B’s reaction to the failure. This plot trajectory continues for most of the rest of the scene, which is not in the datum. But it can only do so because B replies in such a way to confirm that this particular trajectory is being taken up. B and A agree on a particular trajectory to follow, because B responds with an affirmative augmentation.

A particularly interesting example of the use of prosody to negotiate a plot trajectory, and one that constitutes an instance of negative augmentation, occurs towards the end of the scene. Consider lines 13 and 14 from datum 17. Notice that B’s utterance rejects what is conveyed by A’s utterance, namely that B hates A. B’s response to this is negative. And since it shifts the state to one that is projected by A (for a new predication is introduced), it is a negative augmentation.
Let us, then, examine the particular strategies that are used in this example. The word *hate* is emphasised in both of the actors’ utterances. As we observed above, A’s strategy is a signal that his preferred plot trajectory is one in which the actors deal with A’s disappointment at the apparent failure of the experiment. The rapidity of A’s utterance, and the exaggerated stress patterns suggest an emotional state of the character (note the use of capital letters instead of the underline notation, which indicates more extreme fluctuations in volume). B’s response, which exhibits some degree of desperation, uses the same exaggerated stress. But what is important to note is that *hate* is the word that is intentionally stressed by B, instead of any of the other words in the sentence.

Why is this the case? Suppose that B stresses either of the pronouns. It appears intuitively clear that, if a pronoun (or noun) is stressed in a declarative sentence, it triggers a presupposition that is equivalent to a cleft construction of the same sentence. That is, stress pragmatically operates in the same fashion as an alteration in word order does. This intuition predicts that, for instance, (79) triggers the same presupposition as (80) does.

(79)  *I* don’t hate you.

(80)  It is not *I* who hates you.

So if B stresses *I* or *you*, a presupposition is triggered such that someone hates A (i.e., ‘it is not I who hates you’), or that B hates someone other than A (i.e., ‘it is not you whom I hate’), respectively. In triggering unresolved presuppositions, these instances of utterances pose questions that may demand satisfaction in a plot trajectory. One can well imagine that a response to (79) might be similar to (81).

(81)  Wait, *who* hates me?

If it is the intention of B to pursue a plot trajectory that answers these questions, then this is a successful strategy. But this does not appear to be the case here; it appears instead that B is interested only in signalling that she wishes to follow the plot trajectory offered by A. There are two words left in the sentence, then. Suppose that B stresses *don’t*. I do not perceive any pragmatic differences between this and stressing *hate*. There are, I argue, differences in the dramatic function. Stressing *hate* echoes the stress pattern that A deploys, and so appears to signal with greater force that B’s interpretation of A’s utterance is at least close to the projected interpretation. It appears, then, that this prosodic pattern is optimal for this sequence of words, if the preferences of B dictate that A’s projected plot trajectory be followed. And since the response follows A’s preferred plot trajectory and yet rejects the assertion that A makes, this response is an example of a negative augmentation.

Stress may also be used to signal redirection of a plot trajectory. Lines 6, 9 and 11 of datum 11, which was examined in the previous chapter, are reproduced in (82).

---

^3^Lines 7, 8, and 10 are not included because they contain extraneous details.
A plot trajectory is proposed by C, and rejected by A. Note that rejection is not necessarily a hostile move. In this instance, A is being cooperative, and he is simply signalling his preferences over the plot trajectory. As I mentioned when this datum was analysed in the previous chapter, this rejection is a subversion of the expectation that, for these characters, to use the car’s heating is euphemistic for sexual intercourse. The words that A stresses are wind, windows, and asphyxiate. The use of stress on these words in particular shifts the frame so that the same action—using the car’s heating—signifies something other than what it is expected to signify. That is, by using the car’s heating and winding up the windows of the car, a situation when these characters would have sex becomes one in which they asphyxiate themselves. Stress is used in this instance to signal an affirmative redirection. It is affirmative because A suggests that the characters wind up the window, but suggests that they do it for a different purpose than C intended.

6.2.2.2 · Repetition

A second strategy that actors employ is the production of utterances that, taken together, re-enforce the plot trajectory that they prefer. A particularly clear instance of such a strategy is found earlier in the scene that was transcribed in datum 15, transcribed overleaf in datum 18. As I have noted in §6.2.1, this scene seems to contain an example where the actors’ preferences over potential plot trajectories diverge. In datum 18, B’s preferred plot trajectory is clearer than A’s, and it appears to involve the conflict that arises from B refusing to cease twirling his pistol in the direction of A. A’s preferred plot trajectory is not realised, but it seems that it would be related to their journey into the American West.

Let us focus on A’s use of repetition to attempt to signal his intentions. Line 1 signals his intention that the plot be set in the American West during the period of expansion in the late eighteenth and early nineteenth centuries. A does so by using an accent that is stereotypically associated with characters in Western films. This is accepted by B, and augmented by his utterance in line 3. In line 5, A appears to begin to signal a more involved plot trajectory, but is derailed by a mention of B’s action in twirling a gun. In lines 8 and 13, A attempts to return the plot trajectory to one that is closer to his original intention. There are a couple of recurring elements of A’s lines that he attempts to use to signal his intended plot trajectory. References to the journey that they have undertaken to get to their place in the West are in lines 1 (‘whirlwind of a journey’), 5 (‘glad we came aw—’ before he is interrupted by B twirling his pistol), 8 (‘we came here for good reason’), and 13 (‘coming out West was a risk in itself...’). The repetition of an aspect of the setting appears to suggest that A has some plot trajectory involving the location of the scene in mind. And further, it seems that A attempts to elicit a response that confirms this trajectory, so that it can be pursued. This trajectory becomes less likely to be taken up as the scene proceeds, for the trajectory involving pistol twirling ‘hijacks’ the trajectory that A intends. Perhaps as
6.2 · Strategies for resolving interpretation

**Datum 18** Beginning of 2-A2. The indices in this scene are consistent with 15, as this is the beginning of the scene reproduced in that datum.

1 A ((in an American accent typical of Western films)) what a whirlwind f a journey we've been on, ↑frank
   (0.5)
2 B ((in an American accent typical of Western films)) what bloody tumble weeds i've seen (0.2) jimmy
   (2.4)
3 A i'm glad we came aw- ((he notices that B is twirling his pistol)) eh don't wave that thing at me:
   (0.9)
4 B hey i'm just twirlin my gun
5 A well twirl it out therehh hhlike everyone hhelse (0.8) to the open rangees fra:nk (0.9) we came here for good reason
   (1.0)
6 B it doesn't have the feel to it (0.5) as it does when you twirl it in front of another man
7 A don't (0.4) frank
8 B now i feel like i'm twirling for a reason (. ) like i could just draw it at any [moment]
9 A [frank ] comin out west was a risk in itself (. ) we left our families (0.7) a reliable food source (0.3) this is the last kind of risk i wanna take
10 (0.6)
11 B >sorry< something about twirlin a gun just makes me feel alive man

As a result of this, A opts for a different method, and juxtaposes their present location in the wilderness with the civilised West, in line 13 (‘we left our families, a reliable food source’). Finally, after line 19, he succumbs to B’s preferred plot trajectory.

B also employs repetition to bring about his preferred plot trajectory. Lines 10 (‘it doesn’t have the feel to it as it does when you twirl it at another man’), 12 (‘now I feel like I’m twirling for a reason’), and 15 (‘something about twirling makes me feel alive, man’) all return the conversation to pistol twirling, and all endow the character with motivations for why he is interested in twirling his pistols.

The repeated reference to certain entities in the discourse appears, then, to be a strategy that actors use to emphasise their preferences over the plot trajectory. I propose that the strategy is effective because a greater number of data are provided to an agent who attempts to resolve an interpretation problem, and the conditional probability of a given interpretation is greatly increased. At the beginning of the chapter, I discussed how the constraints that are placed on the possible resolutions of an interpretation problem increase with the amount of information in the common ground. Repetition is a strategy that leverages this fact. Utterances that repeat a certain piece of information in the common ground are designed to make that information more salient to subsequent interpretation resolution problems.
6.3 · Heuristics for resolving interpretation

The two strategies identified by the analysis of the last section approach interpretation problems in different ways, yet they both have negotiation at their core. In this section, I propose four heuristics that pertain to negotiation in improvised theatre. In §6.3.1, I define a heuristic for markedness, which determines whether the interpretation of an utterance is optimally clear given a series of responses. The heuristic defined in §6.3.2 concerns the use of strategies that repeat certain facts, in order that certain interpretations be more likely to become operative. In §6.3.3, I define two heuristics: one that describes the optimal expressions of polarity for utterances according to a common ground, and another that expresses that assertions in improvised theatre are only negated if an actor explicitly rejects them.

6.3.1 · Markedness of the operative status of an interpretation

A problem that pervades the data that are presented in this chapter, and one that pervades the communication of improvised theatre generally, is the problem of arriving at an operative interpretation. The major component of this problem is one of verification. While it is trivial for a single player, be she the Sender or Receiver, to decide that a particular interpretation is the correct interpretation, it is extremely difficult for any single player to ensure with certainty that the other player has selected a particular interpretation.

Let us recapitulate the model of how operative interpretations are reached that I outlined at the beginning of the chapter, which is based on Arundale’s co-constituting model (Arundale 2008, 2010). The model that I define is a somewhat abstracted form of Arundale’s original formulation, and outlines a three-step process of arriving at an operative interpretation.

(1) The first speaker produces an utterance that projects a number of interpretations, with a particular interpretation intended to be conveyed;

(2) the second speaker produces an utterance that projects the same interpretation that the first speaker intended, called the provisional interpretation; and

(3) the first speaker produces an utterance that confirms the provisional interpretation as the operative interpretation that will be used in the remainder of the conversation.

This is clearly the most optimistic scenario for the interlocutors’ arrival at an operative interpretation. An insight of Arundale’s (2008, 2010) theory, and also of game theory (Franke 2011), is that there is no inherent reason why the intended interpretation of an utterance should be taken up by other participants in the conversation who observe the interpretation. This is true of every turn in a conversation. For instance, just as the second speaker can misconstrue the intentions of the first speaker, the first speaker can misconstrue the intentions of the second speaker, even after having observed her utterance in step 3 of the sequence above. A consequence of this high level of uncertainty is that actors tend to provide enough evidence for a particular interpretation that the interpretation becomes operative. But here we encounter another major problem, namely that there is no reliable
way for an agent to know which interpretation another agent considers most probable. This is a problem that also presents itself in game theory, though it is not usually couched in the vocabulary of co-constitution of meaning (see Benz 2012a, Parikh 2007, Asher & Williams 2006, McCready 2012). The common resolution of this problem is to stipulate that optimal signalling strategies are those that encode information about the world-state in such a way that it appears most probable to the Receiver when she revises her belief (see Franke 2011). This is a good solution, since it relies on the basic mechanics of game theory to formalise a basic heuristic of naturally occurring speech: if agents are cooperative, they will attempt to produce utterances that are able to be understood by their fellows.

Heuristic 6, then, stipulates a criterion of markedness. If an interpretation is marked, it is the most likely interpretation to be operative, following a series of utterances.

Heuristic 6: Optimal markedness. Given a particular utterance, and a series of responses such that one response is made by each of the parties to a conversation, an operative interpretation of the utterance is optimally marked if it is most probable given all of the responses.

The formal implementation is complicated, for it requires the specification of a situation in which each player is allowed to make a response.

Definition 6.5: Optimal markedness (from heuristic 6). Let $G^{n+1}$ be an iterated game constituted of $n + 1$ interpreted signalling games, such that $g^1$ has a specific Sender $i$, and such that for the following games $g^2, \ldots, g^n$, each player in $N$ (including $i$) is the Sender for exactly one game. Recall that $s^n$ denotes a strategy profile of length $t$. An interpretation $K \in \kappa$ is optimally marked if it solves

$$\arg \max_{K \in \kappa} \Pr(K|m|m \in s^{n+1}).$$

This definition states that the optimally marked interpretation is the most probable interpretation given the set of messages that are present in the history of an iterated game. This heuristic is designed not to be restricted to the cognition of a particular party to the conversation. It is instead meant to conform to the ideology of co-constitutional theories of meaning, and it is defined at the level of an emergent phenomenon. That is, it operates between players, and provides an analytic category for interpretations, rather than co-opting the players’ beliefs for this purpose.

6.3.2 · Convergence through repetition

The data presented throughout the thesis suggest that actors generally wish their interpretations—both provisional and operational—to agree with those of their fellows. In this chapter, we have seen the use of repeating strategies that attempt to provide strong evidence for a given interpretation. The most obvious example of this that I discuss in this chapter is datum 16, where A does not stop repeating utterances that signal his preferred plot trajectory, even when it appears to lead to situations that are suboptimal. In the other data that I have presented, such as datum 18, repetition is used in subtler ways, but to similar ends.
To illustrate how this operates, let us define the problem as a game. For the sake of simplicity, let us imagine the problem as a signalling game with no common ground. Suppose that a Sender sends five signals to a Receiver, without any response from the Receiver whatsoever. The game in question, then, is constituted of five signalling subgames, so that \( G^5 = g^1, \ldots, g^5 \). Suppose that there are \( k \) possible world-states, \( w_1, \ldots, w_k \), and let us assume a Lewis utility function. To keep the example simple, let us introduce the restriction that the state of the world that is selected in the first signalling subgame is the state that is selected in the subsequent subgames. With this restriction, there is only one world-state that the actors are attempting to communicate about.

The Sender attempts to convey the state of the world through her utterances, and the Receiver attempts to correctly infer the world-state that the Sender has viewed given the observation of these utterances. But unlike the general formulation of signalling games, the Receiver has multiple pieces of evidence to inform her inference; multiple utterances can act as conditions on the probability of a given state. In \( g^1 \), the Receiver’s posterior belief is given by

\[
Pr(w|m) = \frac{Pr(m|w) Pr(w)}{Pr(m)}.
\]

Recall that the DRS of each \( m \in \mu \) has an extension function \( \llbracket K_m \rrbracket \) that maps \( \kappa \) to \( W_{\mathcal{R}} \). The inclusion of a world \( w \) in \( \llbracket K_m \rrbracket \) increases the probability of \( w \) given the observation of \( m \); otherwise \( m \) would be either a lie (and thus most likely contrary to the interests of both players), or self-contradictory. The observation of \( m \) thus constrains the world-states that are possible. In \( g^2 \), the Receiver’s prior belief is identical to her posterior belief of the game before. Since posterior belief is simply a probability distribution over the game’s set of world-states, the probability distribution that is generated by the update of the prior belief by the observation of \( m \) in \( g^1 \)—that is, the Receiver’s posterior belief—, may be used as a prior belief in a second game. The observation of the signal in \( g^2 \) likewise updates the Receiver’s prior belief, which further constrains which world-states are possible, for the world-state that the Sender is trying to convey to the Receiver must be a member of \( \llbracket K_{m_1} \rrbracket \cap \llbracket K_{m_2} \rrbracket \). This continues until the fifth game \( g^5 \) when the Sender’s prior beliefs have been conditioned by the utterances of each of the five games. The formal expression of this is quite simple, in light of the discussion in §6.1 above.

\[
Pr(w | \bigcap \{K_{m_1}, K_{m_2}, K_{m_3}, K_{m_4}, K_{m_5} \}) = \frac{Pr(\bigcap \{K_{m_1}, K_{m_2}, K_{m_3}, K_{m_4}, K_{m_5} \} | w) Pr(w)}{Pr(\bigcap \{K_{m_1}, K_{m_2}, K_{m_3}, K_{m_4}, K_{m_5} \})},
\]

Repetition is effective precisely because the utterances that a Sender produces diminish the set of possible world-states by ‘shearing away’ those states of the world that are not in the intersection of the extensions of the utterances’ DRSs. But observe that this implies that, for a strategy of repetition to be effective, each utterance must be sufficiently distinct. If a Sender repeats precisely the same utterance a number of times, nothing is learned apart from what is learned by the observation of the first instance of that utterance. The Sender must ensure that there are differences between the utterances, but that the intended
interpretation is common to the extensions of all of the utterances. From these facts, we can derive heuristic 7.

**Heuristic 7: Convergence of repetition strategies.** Given an intended interpretation and a series of utterances meant to convey this interpretation, a strategy of repetition achieves optimal convergence if and only if the intersection of the extensions of each of the utterances contains only the intended interpretation.

Observe that this heuristic’s condition of optimality is designed such that it excludes any series of utterances if there are any two utterances in that series whose extensions are pairwise disjunct, for this will cause the final intersection to be empty.

The formal definition (definition 6.6) is a fairly simple application of the heuristic.

**Definition 6.6: Optimal convergence of a repetition strategy (from heuristic 7).** Let $G^k$ be a composite signalling game constituted of $k$ subgames. A series of utterances $m_1, \ldots, m_k$, where each is produced in its own subgame, is optimally convergent as a repetition strategy with respect to a particular state of the world $w \in W$ iff

1. $w \in [K_{m_1} \cup \ldots \cup K_{m_k}]$ where $K_{m_i}$ denotes that $m_i \in \|K_{m_i}\|$; and
2. there is no $w' \in W$, such that if $w' \neq w$, then $w' \in [K_{m_1} \cup \ldots \cup K_{m_k}]$.

This definition states that a series of utterances is optimally convergent if a particular plot-world is in the DRS that results from the merge of each of the utterance’s individual DRSs, and if there is no other plot-world that is plausible given this extension.

Strategies that fit this definition of optimality appear plausibly achievable if we examine abstract games with few states of the world, few utterances, and only a small number of subgames. But when we analyse naturally occurring speech, it might seem to be unrealistically strict. There are an uncountable infinity of world-states, an equally large set of possible utterances, and the extension of each utterance contains an enormous number of members. Designing and producing a series of utterances that meet the requirements of the heuristic will indeed be difficult in naturally occurring speech. Recall, however, that the purpose of the heuristics is not to designate certain strategies as achievable, but is instead to propose that there are certain ways that actors in improvised theatre are expected to act. In the case of this heuristic, actors do not necessarily expect that their fellows will execute a perfectly convergent strategy of repetition, but they do expect that, when a strategy of repetition is used, it is used with the intention that a certain interpretation be conveyed. Actors know to be watchful for a series of utterances that constitutes a strategy of repetition and, because of the expectation expressed by this heuristic, they are able to interpret the series of utterances as an attempt to convey a certain interpretation.

6.3.3 · Response strategies

We have seen in this chapter and the last that actors behave in ways that suggest an awareness, at least at an intuitive level, of different kinds of response, and of their effects on the state of the scene. Therefore, I include optimal narratological strategies as heuristics. To
do so, I do not formulate a heuristic for each of the four categories (affirmative augmentation, negative augmentation, affirmative redirection, and negative redirection). Instead, I formulate a heuristic for the characteristics that determine each: polarity (positive or negative) on the one hand, and conformity (augmentation or rejection) on the other. The reason these are treated inseparably is because they only differ in the degree to which they are successful: a response’s optimal polarity depends on the degree to which it distinctly conveys whether it is negative or positive, for instance.

Polarity, as has been mentioned, describes whether a response to an offer accepts the premise of the offer. An optimally polar response, then, is simply one that is unambiguous in its polarity.

**Heuristic 8: Optimal polarity.** A response to an offer is optimally polar if it is unambiguously the case that what is intended to be conveyed by the offer utterance is taken up and subsequently accepted or rejected.

The formal definition of this heuristic is simple.

**Definition 6.7: Optimal polarity (from heuristic 8).** Given an utterance $m$, and an interpretation $K$ such that $m \in \|K\|$, an optimally polar response strategy solves either

$$\arg\max_{m \in \mu} \Pr(K|m)$$

if it is positive, or

$$\arg\min_{m \in \mu} \Pr(K|m)$$

if it is negative.

That is, if a particular $K$ is the most probable interpretation choice given a specific $m$, then $m$ is optimally positive. Likewise, $K$ is optimally negative if it is the least probable response given $m$. To justify this claim, recall that this definition describes optimality of the polarity of a response, and therefore deals with extremes. If a response is positive, then it affirms that $K$, and so an optimally positive response is the response that most clearly accepts $K$. An optimally negative response is the response that most clearly rejects $K$. Since polarity is binary, $\Pr(\neg K|m) = 1 - \Pr(K|m)$. By decreasing $\Pr(K|m)$, then, $\Pr(\neg K|m)$ is increased.

The formal expression of this heuristic allows for a determination of how actors interpret utterance polarity in improvised theatre. There is a subtle distinction here. Clearly, the definition of the polarity of an utterance is a simple matter formally, for it is the case that any formal language has negation included in its syntax. But what this heuristic addresses is how actors recognize the polarity of their fellows’ utterances, and furthermore how they expect this aspect to manifest itself in utterances. Therefore, it also describes how actors produce their utterances so that their polarity will be successfully interpreted. Interpretation of polarity is key for an actor to signal her preferences, and therefore an account of polarity is vital to the efficacy of the model as a whole.

It was suggested in §5.2.2.3 that actors accept proposed inclusions in the common ground by default, and that it is only explicitly that an actor might reject an assertion.
This was noted in the context of a discussion of actors’ assertions regarding the attributes of characters, for it appears that in general such assertions are accepted, and furthermore that the acceptance occurs by default. I have exhibited multiple data that suggest that this is the case in terms of character attributions. In this chapter, there have also been instances where tacit acceptance also applies to information about the plot world. For instance, in datum 16, which is reproduced in (83), A asserts that the characters are in a zoo.

A: o look at that (0.6) paɪnدا (1.0) .t aw look at those
deg:i (0.5) i LOVE the zoo ma:n (.) the zoo just
really (0.6) DOES somethin for me ma:n i mean like
(2.1)
(83) B: [ca ɔː w ] ((i.e., a bird call))
A: [hey have you th-] ever THOU:GHT (.) »i mean«<
(0.8)
C: ɪə:ʃ (1.3) i mean (.) what if they're not REAL (.)
that's what you're getting at [isn't it what]=
A proposes that be_at(A, B, C, zoo) be included in the plot’s common ground. B’s response, which involves producing a bird call, explicitly references this assertion. Through proposing that there are birds at the same location, he accepts the proposal that the characters are at a zoo. Importantly, note that C raises no objection to this assertion. Indeed, he does not refer to the assertion at all, but deals with A’s next offer, which is to propose something that might be strange about the zoo. I have already discussed the latter parts of the scene, in which A tries to overturn the trajectory of the plot. This is a clear instance where rejection of assertions requires explicit action, for if A did not express his objection to that particular plot trajectory, it is probable that the other actors would have assumed that he accepted it.

And so, I introduce a heuristic that captures this assumption of actor behaviour.

Heuristic 9: Default acceptance. A response strategy is by default positive.

This heuristic is not an assertion of optimality, and so does not require a formal expression, as was discussed in chapter 1.

6.4 · Conclusion

In this chapter, I have focussed on processes of negotiation of two key components of communication in improvised theatre. The first, negotiation of operative interpretations, involves agents reaching agreement on the interpretation of a particular utterance that will be utilised for the remainder of the conversation. The second, negotiation of plot trajectories, is far closer in spirit to the normal interpretation problems that the study of signalling games address, and involves agents signalling to one another their preferences for how the plot will progress. I have argued that negotiation is fundamental to a proper understanding of communication in improvised theatre. In particular, I have identified
the effect of a common ground, the constitution of which is agreed upon by actors involved in the performance, as exerting the strongest single effect on which interpretations are proposed provisionally, and on those that are eventually rendered operative. Treating communication as a fundamentally negotiated practice has been shown to be highly appropriate both to a description of linguistic acts in improvised theatre, and to a more thorough conception of game theoretic rationality as it pertains to interlocutors engaged in interaction.

I have examined strategies that actors use to negotiate operative interpretations, and also those that are used to signal their trajectories over the plot. Since actors’ communication relies on the entire linguistic faculty—as is the case in other genres of naturally-occurring speech—, it is impossible that a single study contain a full catalogue of possible strategies.

I have established that the common ground of a conversation acts as the basis on which any future interpretation is resolved, and therefore as the basis for future selections of linguistic strategies. For two interlocutors to be mutually understood, they must ensure that the common ground is agreed upon, and that there are no disjunctions in the common ground, or misperceptions of the common ground. In the next chapter, we turn to an exposition of cases that involve the opposite: those cases in which agents exploit differences in common ground knowledge to cause differences in communication, and thereby to conceal information within their utterances.
“Ware whee are, wore ’way? And throttles happy tune me?”
Though the stream of nonsense was no less upsetting, Michael was astonished to discover that he almost understood himself. He’d asked her where they were as he’d intended to, but all the words had come out changed and they were twisted round, with different meaning tucked into all their crevices.

Alan Moore, *Jerusalem*

“It is some time since I last heard the sound of your shears. How long have you been eavesdropping?”
“Eavesdropping, sir? I don’t follow you, begging your pardon. There ain’t no eaves at Bag End, and that’s a fact.”
“Don’t be a fool!”

J. R. R. Tolkien, *The Lord of the Rings*

In the previous chapter, we were concerned with the effects of a ‘monolithic’ common ground on players’ strategies. This common ground contained the information that is introduced during a conversation, and accepted for the purposes of that conversation. We were interested in the linguistic devices that actors use to control the information that enters the common ground, and the negotiation that underpins this process.

This chapter takes what was discussed in the previous chapter and extends it in two interrelated ways. First, the singular common ground of the previous chapter is decomposed into partial grounds, such that each partial ground is unilaterally shared between two players. Whereas in the previous chapter agents had to control the discourse so that their utterances were interpreted felicitously, agents in this chapter must control multiple partial grounds to achieve resolution of the interpretations of their utterances. Second, I do away with the assumption that all players are cooperative, and instead propose a model
that is able to capture situations in which an agent wishes to conceal information from one or more of the parties to a conversation. In contrast to the games so far discussed, agents’ preferences are not aligned, and the Sender S wishes that her utterances be resolved only by a subset of the players in $N \setminus \{S\}$.

Given the discussion of the previous chapter, the means by which an agent might achieve this are clear: the partial grounds of certain players includes information that will allow them to resolve an interpretation, while the other players lack this information. Controlling the common ground under these conditions becomes far more complicated, since players must be conscious to foster and subsequently exploit differences in information. In other words, players attempt to promote divergences in partial grounds so that certain of the other players make erroneous selections of interpretations.

I aim to show that any valid model of improvised theatre must necessarily include some provision for concealment of information. But there also appears to be good justification for the inclusion of concealment in game theoretic models of naturally occurring speech generally. Consider example 1.1, which I reprint below as example 7.1.

Example 7.1: Lucky’s piano recital. Dutiful Vladimir and naïve Estragon attend a piano recital put on by their mutual acquaintance, Lucky. It is terrible. Lucky butchers the Bach, murders the Mozart, and, for good measure, wrecks the Ravel. After a marathon two-hour performance, Vladimir and Estragon, exhausted, attempt to slip away, but they bump into Lucky at the exit. After exchanging the usual polite nothings, Estragon asks Vladimir what he thought of the performance. Not wishing to offend Lucky, but also not wishing to lie, Vladimir murmurs, “it was very expressive.” Lucky is rapt, and replies, “I don’t play accurately—any one can play accurately—but I play with wonderful expression. As far as the piano is concerned, sentiment is my forte.”

Recall that, in this example, Vladimir wishes to convey two separate things: to Estragon he wishes to convey that he did not enjoy the performance, while to Lucky he wishes to convey that he did. That is, with a single utterance, Vladimir wishes to convey two complementary, mutually disjunctive interpretations.

I claim in this chapter that speech in which utterances conceal information from certain parties is best described as steganography in naturally occurring speech. Steganography is the art and science of transmitting usually private information in the presence of adversaries by concealing it within an apparently innocent ‘cover’ medium (Provos & Honeyman 2003). Steganography is closely related to the better-known cryptography, but there is an important difference between the two. Applications of the latter most often scramble the message so that (a) any adversarial observer is able only to see a seemingly random block of characters; (b) any friendly observer is able to decrypt the message with a shared secret, that is, a piece of information, that is transmitted before the secret message; and (c) it is infeasible to compute the hidden message given only the random block of characters, without the shared secret (see chapter 1 of Menezes, Van Oorschot & Vanstone 1996). Meanwhile, steganography conceals the message in such a way that the presence itself of the message is undetectable. The key difference between steganographic
and cryptographic systems, then, is the kind of output that each produces. The output of a cryptographic system appears to be a garbled sequence of random, unusable data. Text that has been encrypted, for instance, is readily identifiable to human observers, for it is a long sequence of gibberish, made up of letters and numbers. Steganographic systems, on the other hand, output data that is structured, and that appears to be a normal file. An image file that has some information steganographically concealed within it appears to be simply an image, and a user will only be able to receive the concealed information if she knows (a) that there is information hidden in the image; and (b) the method of extracting the hidden information. The security of a steganographic system is a metric which measures the perceptible difference between a message that contains no concealed information and one that does. If a steganographic system is secure, an unwarranted party should not be able to tell the difference between the two.

Steganography is more appropriate for studies of naturally-occurring speech than cryptography because of the outputs that are produced. In example 1.1, Vladimir uses a natural language utterance to communicate both that he enjoyed and that he did not enjoy the performance. The utterance appeared complimentary to Lucky, and did not appear to have any concealed message, while it did appear to have a concealed unflattering message to Estragon. The messages, then, are similar to the image files above. They must appear structured and plausibly innocent, and, importantly, they must be able to be interpreted by the unwarranted party. Some sort of linguistic encryption system would therefore not be appropriate, for it would output natural language utterances that are unable to be interpreted by the unwarranted party; an example of such a system is where two interlocutors converse in a language that a third does not speak. Since we are interested in situations where all interlocutors can interpret the utterances, but where messages are concealed and meant only to be interpreted by a subset of the interlocutors, steganographic systems are more appropriate.

I have noted a number of times throughout the previous chapters that improvised theatre has the somewhat unusual property that any speech act must be able to be interpreted both as a plausibly natural utterance in a plot-world, and as a statement of an actor’s interpretations over the plot-world. This tension is noted in the literature (see Baumer & Magerko 2009, 2010, Magerko et al. 2009), but is rarely ever dealt with other than through a cursory acknowledgment that this is a challenge that actors face. The only exception is Fuller & Magerko (2010), who manage a partially satisfying, informal account of the problem. I claim that including steganography in a game theoretic model of improvised theatre allows for a highly accurate description of this property of actors’ speech, and a rigorous means of modelling it.1

In this chapter, I endow the game theoretic model with the ability to account for interactions where two interlocutors collude to conceal information from a third party by using utterances that are able to be interpreted by this third party. In §7.1, I provide a short

---

1The production of utterances that have different intended interpretations for different audiences has also been noted by Henderson & McCready (2017) in political speech, and by Morrissey (2010) in online trolling.
introduction to information theory, and particularly to the information theoretic understanding of communication. In §7.2, I introduce steganography. Since steganography is a field whose applications greatly vary, information theory is used as the mathematical and theoretical basis for a general account of steganography. With some amendments, I apply the model of steganography to communication in natural language in §7.3. This application involves a discussion of steganography’s appropriateness for a general description of the concealment of information in natural language, before proceeding to an incorporation of steganographic formalisms in a game model. As usual, I discuss instances from the dataset and how the steganographic model might be used to account for them in §7.4. I appraise the model in §7.5, and suggest ways in which it is advantageous to the study of language, but also identify certain scenarios in which its use is inappropriate. In §7.6, I define heuristics that pertain to the concealment of information in improvised theatre. The chapter concludes in §7.7.

7.1 · Information theory

Information theory is a branch of applied mathematics that studies properties of information, and particularly how information is communicated (MacKay 2003, Appelbaum 2008). It was developed through practical approaches to engineering problems associated with transmitting digital information through a channel, such as a copper wire. It sought to answer two major problems: the maximal degree to which data can be compressed when transmitted, and the maximal transmission rate over the particular channel (Cover & Thomas 2012). In two seminal papers, Shannon (1948, 1949) devised a formal apparatus that allows for a more abstracted study of the properties and transmission of information in a system that has some element of uncertainty. If an abstract system is thought of as a machine that accepts inputs and produces outputs, the presence of uncertainty in a system implies that the outputs of that system are not able to be reliably predicted by the inputs. Shannon devised a measure of uncertainty of such a system, called that system’s entropy. Entropy provides us not only with a useful metric of how uncertain a system is, which allows us to suggest ways of decreasing or of promoting uncertainty, but it also allows us to quantify the amount that is learned about a system by feeding it an input and observing the output.

I spend this section outlining Shannon’s information theoretic theory of communication. I formally describe entropy in §7.1.1. Entropy is the basis for the quantitative measure of a steganographic system’s efficacy, and as such is extremely important to the model of natural language steganography that I formulate in this chapter. In §7.1.2, I introduce the information theoretic model of communication that provides the basis for the formal model of steganography.

7.1.1 · Information content and entropy of random variables

As I mentioned in the introduction to this section, information theory is most often concerned with modelling a basic kind of machine that accepts an input and produces an
When we describe an information theoretic system, particularly one of communication, it is useful to imagine a black box, into which a particular input is entered, and out of which a particular output is returned. The output is a transformation of the input, and this transformation results from a specific mathematical function. In information theory, the function invariably has some degree of indeterminacy: inputs and outputs often do not correlate precisely.

Suppose that we wish to learn the mathematical function that predicts the output for a given input. We might test a particular input \( x \) a number of times, to check whether an output \( y \) is consistently returned. For instance, let us suppose that inputting \( x \) returns \( y = x^2 \). The first time we observe \( y \), we learn something about the method that the black box uses to transform inputs. The second time, we learn that this transformation appears to occur consistently. In the second instance we have learned less than the first, as we have already observed that \( x^2 \) is returned. When we perform further tests, we observe more outcomes, and so become more certain of the outcome. And we thereby learn less, or are less surprised, with each observation that confirms our expectations. Before the first test, there was a wide range of potential outcomes, so our uncertainty, and therefore the amount that was learned from each observation, was extremely high, which is equivalent to stating that the entropy of the system was high. As we learn more about the system, the uncertainty diminishes until it is virtually nonexistent, and it is equivalent to say that the entropy diminishes until it is practically zero.

Shannon (1948) devises a means of quantifying the amount of information that the observation of a particular outcome produces. It is based on a very simple intuition, which was illustrated above: if an outcome is expected, then we learn less than if it was unexpected.

We have already seen examples where this applies to natural language. Recall the example from chapter 2 of Vladimir disarming a bomb. Suppose that Vladimir is completely certain about which wire to cut. He would learn nothing new from Estragon’s suggestion of which wire to cut, even if Estragon were an expert in disarming explosives. But if Vladimir considers each wire as likely as the next to detonate the device, an increase in the number of wires positively correlates with the amount of information that is learned when Vladimir tells him which wire to cut. That is, when Estragon tells Vladimir which wire to cut, Estragon effectively gives Vladimir information about all the wires: which will cause the bomb to explode, and which will not. The more wires there are, the more uncertain Vladimir is about which wire to cut, and the more Vladimir learns from Estragon’s message.

Information entropy applies to a random variable that can take a number of values, such that each value is associated with a particular discrete event, and such that each event has a probability. For instance, the flip of a fair coin has two possible outcomes. A random variable simple assigns a real number to each of these outcomes, so that, for instance, heads = 0 and tails = 1, each outcome has a probability between 0 and 1, and \( \Pr(\text{heads}) + \Pr(\text{tails}) = 1 \). As the entropy of the variable increases, the certainty of any
particular variable being observed decreases. A second, interrelated sense of the measure that entropy represents follows from this definition. Entropy also commonly refers to a quantified measure of how much is learned by the observation of a particular value of a variable. So, as the entropy of the variable increases, so too does what is learned from the observation of that variable.

The following series of definitions provide means of quantifying entropy for random variables, and for probability distributions. These definitions are chosen specifically with to provide the necessary and sufficient mathematical basis for the definition of a steganographic system in the most concise manner possible.

First, let us define a parameter that describes how much is learned from the observation of an outcome (based on MacKay 2003: 32).

**Definition 7.1: Information content of an outcome.** Let \((\Omega, \delta)\) be a probability space. The information content of an outcome \(x \in \Omega\) is found by solving

\[
h(x) \equiv \log_b \left( \frac{1}{\Pr(x)} \right),
\]

where \(b \geq 1\) is the base of the logarithm, such that \(y = \log_b(x)\) is the unique \(y\) that solves \(b^y = x\).

If we assign a real number to each possible outcome of the probability space, so that we have a discrete random variable, then we can describe the entropy of that random variable. As I indicated above, the entropy is the average information content of the variable’s outcomes (based on MacKay 2003: 32).

**Definition 7.2: Information entropy of a discrete random variable.** Let \((\Omega, \delta)\) be a probability space, and let \(X : \Omega \rightarrow \mathbb{R}\) be a discrete random variable. The information entropy of \(X\) is found by

\[
H(X) = \sum_{x \in X} \Pr(x) \log_b \left( \frac{1}{\Pr(x)} \right),
\]

where \(\Pr(x)\) is the probability that \(X\) is a given value \(x\).

**Remark 7.3:** To avoid undefined results, and in keeping with convention (see MacKay 2003: 32), we define \(0 \log_b(1/0) = 0\), since \(\lim_{x \to 0^+} x \log_b(1/x) = 0\).

The base of the logarithm determines the unit that the entropy is measured and expressed in. Three common units are (i) bits, where \(b = 2\); (ii) nats, where \(b = e\);\(^2\) and (iii) bans, where \(b = 10\). I select a unit for the base, and justify this selection below in §7.3.2.3. For the exposition of information theory that follows, bits are most often used, as this is usually the case in information theory.

\(^2\)\(e\) is Euler’s number, which is approximately equal to 2.71828, and which is the sum of the infinite series

\[
e = \sum_{n=0}^{\infty} \frac{1}{n!}.
\]
Let us apply definition 7.2 to a couple of examples: first, a coin flip, and second, a dice roll. In a coin flip, the set of outcomes is \{heads, tails\}. Let us define a discrete random variable \(C\),

\[
C = \begin{cases} 
0 & \text{if the upwards face is heads} \\
1 & \text{if the upwards face is tails}
\end{cases}
\]

and let us assume that the coin flip is fair, so that the probabilities of each of these are equivalent; \(\Pr(0) = \Pr(1) = 0.5\). The information content of either of these possibilities is \(h(0) = h(1) = \log_2 \frac{1}{0.5} = 1\) bit. The entropy of the random discrete variable, then, is simply the mean of the information contents of its values: \((1 + 1)/2 = 1\). The same result is returned if we use the equation for the entropy of the variable:

\[
H(C) = \left( 0.5 \times \log_2 \frac{1}{0.5} + 0.5 \times \log_2 \frac{1}{0.5} \right) = 1.
\]

(7.5)

Since the coin flip is fair, and since the probabilities of the outcomes are uniformly distributed, there is maximal uncertainty, and this variable has maximal entropy. But suppose that the coin flip were not fair, but was skewed towards one of the two outcomes. In this case, the degree to which the probability distribution is skewed towards one outcome or the other is negatively correlated with the entropy of the variable. This is illustrated by figure 7.1.

Now let us consider a fair dice roll. The random variable \(D = \{1, 2, 3, 4, 5, 6\}\) represents the outcomes of the roll. The probability for each value of \(D\) is 1/6. The entropy of \(D\) is given by 7.6.

\[
H(D) = \left( 6 \left( \frac{1}{6} \times \log_2 \frac{1}{1/6} \right) \right) \approx 2.58.
\]

(7.6)

Note that, when more outcomes are possible, the entropy increases, since a greater number of potential outcomes increases the uncertainty that any particular outcome will occur.
Because of game theory’s heavy use of Bayesian probabilities, it is useful to quantify the effect that the entropy of one random variable has on another, by conditionalising the entropy of the second on the entropy of the first (based on MacKay 2003: 138).

**Definition 7.4:** Conditional entropy. The conditional entropy of two random variables $X, Y$ is

$$H(X|Y) \equiv \sum_{x \in X, y \in Y} \Pr(x \cap y) \log_b \frac{1}{\Pr(x|y)}.$$  \hfill (7.7)

A useful and immediate consequence of this definition is that it gives us a means to quantify how much information one variable gives us about another. The *mutual information* between two random variables measures the average reduction in the uncertainty of the unknown value of $x$ when the value $y$ is observed, and vice versa.

**Definition 7.5:** Mutual information between random variables. The mutual information between two random variables $X$ and $Y$ is $I(X; Y) = H(X) - H(X|Y)$, such that $I(X; Y) = I(Y; X)$, and $I(X; Y) > 0$.

The final definition that I present here concerns probability distributions, rather than random variables, and gives us a way to compare the entropy of probability distributions. The *relative entropy* of two probability distributions gives us the average logarithmic difference between the probability functions of the probability distributions. It is defined as follows (based on Cover & Thomas 2012: 19).

**Definition 7.6:** Relative entropy. Let $\delta_X$ and $\delta'_X$ be two discrete probability distributions over $X$. The relative entropy of $\delta_X$ with respect to $\delta'_X$ is defined by

$$D(\delta_X \parallel \delta'_X) = \sum_{x \in X} \Pr(x) \log_b \frac{\delta_X(x)}{\delta'_X(x)}.$$  \hfill (7.8)

$\delta_X \neq \delta'_X$ implies that $D(\delta_X \parallel \delta'_X) \neq D(\delta'_X \parallel \delta_X)$.

Definitions 7.5 and 7.6 are particularly important, since together they will be used to measure the security of steganographic systems.

The definitions that I have introduced in this section are by no means an exhaustive catalogue of the definitions that pertain to information entropy. They were instead selected because of their relevance to an information theoretic model of steganography devised by Cachin (1998, 2004), which I present in §7.2.

### 7.1.2 · A model of information theoretic communication

The major contribution of Shannon (1948), which is the reason that information theory is useful to our purposes in this thesis, was to propose a general model of communication. The primary concern of Shannon’s model is the description of the effect of uncertainty on the design, transmission, and reception of a message. Shannon’s model appears very simple, and can be represented schematically as in figure 7.2 overleaf. Beneath this semblance
of simplicity, however, is a highly uncertain and complicated set of entities.

The information source is a stochastic process; that is, a set of random variables such that each is associated with a particular time. The information that the source generates cannot be directly carried across the channel. For instance, if the source is an electronic thermometer that outputs information about the ambient temperature, this number needs to be encoded in a specific fashion to be transmitted. If the channel is a copper wire, the information must be encoded as an electrical signal, so that it may be communicated to other devices to display the temperature. Any given channel is restricted in this way to a specific kind of information that it accepts for input, and that it uses for its output. The symbols that a channel accepts constitute the channel’s input alphabet, and the symbols that it uses in its output are its output alphabet. These do not need to be equivalent. The signal must be encoded using the channel’s input alphabet. In figure 7.2, the Sender performs this conversion from information to signal. The signal is sent across a channel, and is relayed to the Receiver. The Receiver performs the inverse operation to the Sender, and converts the signal to a message, which is read at the destination.

As the signal travels across the channel, random alterations of the signal are added by noise (Gray 2011). For instance, if the signal is a string of bits, noise causes some bits of the message to flip to the opposite value, with utter irregularity. The diagram in figure 7.2 is somewhat misleading, since it suggests that noise is inserted into the channel. Instead, noise is often treated as a property of the channel itself. It is in this way that it will be understood in this chapter. Cover & Thomas (2012: 183) point out that, since “the transfer of information is a physical process[, it] is subject to the uncontrollable ambient noise and imperfections of the physical signalling process itself”. Noiseless channels, then, are practically impossible. Since noise is inevitable, the Sender and Receiver (which are generally computing devices) are designed to mitigate the error that is introduced into the signal by noise.

We deal here only with discrete channels, that is, channels that are able to emit discrete signals. The signals that have been studied throughout the thesis are discrete. That is, the utterances that are produced in signalling games are elements of a set of utterances, and choices between utterances are treated as the choice between discrete elements of the set. I define discrete channels in definition 7.7 (based on Mahdavifar & Vardy 2011: 6431).
Chapter 7 · The subliminal channel in improvised theatre

**Definition 7.7:** Discrete channel. A discrete channel is a triple \((X, Y, \delta)\) such that \(X = \{x_1, \ldots, x_m\}\) and \(Y = \{y_1, \ldots, y_m\}\) are the input and output alphabets of the channel, and

\[
\delta = \begin{bmatrix}
\delta_{x_1, y_1} & \delta_{x_1, y_2} & \cdots & \delta_{x_1, y_m} \\
\delta_{x_2, y_1} & \delta_{x_2, y_2} & \cdots & \delta_{x_2, y_m} \\
\vdots & \vdots & \ddots & \vdots \\
\delta_{x_m, y_1} & \delta_{x_m, y_2} & \cdots & \delta_{x_m, y_m}
\end{bmatrix}
\]  

(7.9)

is a probability matrix such that \(\delta_{x_i, y_i}\) solves \(\Pr(y_i|x_i)\).

When I incorporate information theory and game theoretic linguistics below, I revise this definition, so that the roles of the Sender and Receiver are more clearly demarcated.

Quantification of noise is not particularly important for the model of natural language steganography that I develop here, and so I discuss it no further. But I would like to highlight a vitally important consequence of the ubiquity of noise. Under this model of communication, a message is encoded into a signal, and sent along a channel. Once it reaches its destination, it is decoded from a signal to, ideally, the same message that was sent. Provided that the channel is always noisy, then the signal that is sent along the channel is invariably altered before it reaches its destination. Therefore, the message that is received will always be different from the message that was sent. The noisiness of the channel determines to what degree.

For this reason, it is preferable to say that the output of a noisy channel will always provide evidence of the message that was sent, rather than to say that the received message is a reproduction of variable quality. What is output by a channel, then, is an estimate of the message (Cover & Thomas 2012). Using a notation that I have established for expected utilities, estimates are denoted with hats. If \(m\) is the message, the estimate of the message is \(\hat{m}\). The use of message estimates, rather than messages themselves, resonates with the game theoretic apparatus of using messages to give evidence for the state of the world. The world-state is not itself sent by the Sender of signalling games, and the signal instead is meant to convey which is the true world-state. In a similar way to signalling games, then, finding the correct message from a set of messages when the estimate has been observed is a simple matter of finding the most probable message, conditional on what has been observed. That is, \(R\) must solve

\[
\arg \max_{m \in M} (m|\hat{m}).
\]  

(7.10)

This notion is central to the information steganographic model I propose in §7.2, and it is in turn a fundamental part of the model that I propose in §7.3.

**7.2 · Steganography**

Steganography deals with the concealment of information within objects that do not ostensibly act as containers for that information. That is, the aim of steganography is to embed the secret information within a cover object so that it is imperceptible to the Adversary while it is transmitted. A steganographic system’s robustness measures how resilient
it is to detection of messages that are transmitted within it. Concerted and deliberate attempts at detection are called \textit{attacks} on the system. Implementations of steganographic systems are generally algorithmic, in that both the embedding of secret information in a cover object, as well as attacks on the system, are automated. A simple example of such a steganographic system is \textit{Least Significant Bit} (LSB) embedding, a method of altering the least significant of the bits (usually the final bit in a bitstream) that encode the colour of a given pixel, thus introducing concealed information (see Dabeer et al. 2004, Mielikainen 2006, Ker 2007b). This simplistic example is effective because alterations in the later bits in a colour specification only minutely alter the colour. “Because the eye cannot detect the very small perturbations [LSB embedding] introduces into an image”, a casual observer is not able to differentiate an image whose LSBs have been altered from one that has not been changed (Ker 2005: 441).

The academic study of steganography was commenced by Simmons (1984), who proposed a vignette to illustrate the problem of steganographic concealment. Example 7.2 is based on this vignette.

\textbf{Example 7.2: Another prisoner’s dilemma.} Vladimir and Estragon have landed themselves lengthy sentences in gaol. Without much hesitation, they begin to hatch a plan of escape by sending messages to one another through an official mailing system run by the warden of the prison. The warden reads every message that is sent in this way. Vladimir and Estragon must find a way of communicating in such a way that the warden is unaware that they are plotting anything, for if their plot is discovered, each will be disallowed from sending any further messages.

The problem that Vladimir and Estragon face involves producing messages that are (a) concealed in such a way that the warden cannot detect the presence of a message; but at the same time, (b) constructed in such a way that it is able to be read by the other prisoner. Observe that this problem does not involve a question of whether the warden is able to interpret the message once it is intercepted. While this is certainly a concern of the prisoners, the threshold of failure of steganographic systems is far lower, for it is the mere detection of messages that causes a system to fail.

Clearly such a method of communication requires the prisoners to agree upon a way of both hiding and of extracting the secret information. That is, the prisoners must have some shared information that the warden does not have; or, at least, a source of information that the warden does not know operates as such. Simmons (1984) provides that the prisoners each have access to a copy of the bible, which they can use as a source of shared information. Using the bible, the prisoners can conduct a conversation which is ostensibly an innocent discourse on certain biblical passages. In reality, they are formulating a plan of escape. In this instance, it is probable that the warden has access to the same edition of the bible as the prisoners, but it is also improbable that the warden realises that the bible is used as a source of shared information. This is precisely the goal of the system of communication.
So, the prisoners are interested in being able to plausible deny that they are conspiring with one another. Roe (1997: 55) describes plausible deniability as a part of a service, such as cryptographic or steganographic software, that allows users to prevent unwarranted access to information:

Anyone can always deny anything, even if there is an abundance of evidence which shows that the denial is false. A lie is only effective if it is believable, that is, if it is consistent with known facts. Thus, an invocation of the plausible deniability service involves the following elements:

- Event X actually happened.
- Party A (the service user) wishes to deny that event X happened.
- Party B (the adversary) wishes to demonstrate to a third party that X happened.
- Party A can produce a false account of what happened (Y) which is consistent with all the evidence that party B can present to the third party.
- Party A arranges that all evidence not consistent with Y is concealed or destroyed.

The key insight here is that a denial is plausible iff it is consistent with the observed facts. So, for instance, steganographic filesystems (see Anderson, Needham & Shamir 1998, Giefer & Letchner 2004, McDonald & Kuhn 1999, Pang, Tan & Zhou 2003, Zhou, Pang & Tan 2004) allow users to plausibly deny the presence of certain files in the filesystem, even if they are compelled to give the decryption keys for the filesystem to a third party. The steganographic filesystem proposed by Pang, Tan & Zhou (2003), for instance, relies on files being distributed to blocks which appear to be empty (Troncoso et al. 2007: 222). Plausible deniability is thus an integral part of utterance design and interpretation in a steganographic model of natural language communication.

The steganographic system that is defined here is a system of hiding information within natural language utterances. The application of steganographic concealment to natural language has a limited and very technical literature of practical techniques and systems (Bennett 2004, Taskiran et al. 2006). The chief concern of these systems is largely the embedding of concealed information or of watermarks3 in ostensibly innocent passages of natural language text (see Atallah et al. 2002, Topkara, Topkara & Atallah 2006). Such systems generally rely on certain linguistic features of text being able to be interpreted as components of a secret message. Wayner (1992, 2002), for instance, devises a system wherein constituents of syntactic structures of context-free grammars themselves may stand for sequences of bits. For instance, a noun phrase stands for a certain string of bits, a verb phrase for another, and so on. The conversation rules for each part of speech are given to the decoder and, when read left-to-right, a sequence of bits is extracted. Other

---

3Watermarks are features which identify some object as having a particular place of creation. The two most widespread uses of watermarks are, unsurprisingly, enforcing copyright restrictions (watermarking may be used, for instance, to embed information to identify an office document printer, and so where pirated copies of copyrighted materials are printed or copied; see Yu, Niu & Sun (2005) for a paper indicative of the field), and in distinguishing true bank notes from counterfeits.
models propose similar uses of syntactic structure (Atallah et al. 2001), of synonym substitution (Topkara, Topkara & Atallah 2006), of morphosyntactic alteration (Meral et al. 2009), and of manipulation of abstract semantic tree representations of sentences (Chapman, Davida & Rennhard 2001, Atallah et al. 2002) as means to achieve steganographic concealment with natural language coverforms. Attacks on such systems generally rely on statistical analyses to reveal irregularities attributable to the embedding of secret information (Bennett 2004).

The endeavour that I undertake is entirely different from these approaches to steganography: I instead look at primarily spoken language, and the means by which a human agent, rather than a machine, can embed covert communication in natural language utterances. For this reason, I do not adopt the principles of the hitherto-developed approaches to natural language steganography.

In this section, I provide a formal definition of steganography that I will use for the remainder of the chapter. In §7.2.1, I provide an information theoretic model of steganography. This model introduces many of the key concepts of steganography while providing formal definitions for the modelling and quantification of such entities. In §7.2.2, I discuss the somewhat scarce literature on game theoretic approaches to steganography.

Before I proceed, it is useful to introduce vocabulary that is used in the steganographic literature (see Cachin 2004). Most of these terms are derived from the related nomenclature of cryptography. Steganography is the study of the concealment described above. A stegosystem is a particular set of methods for concealing information; steganalysis is the study of breaking stegosystems, i.e. of devising a means to determine whether an object contains concealed information, and a means of extracting that information. There are two types of messages that we are broadly interested in. The first are messages that are innocent and legitimate, containing no concealed information, which are called coverforms.

The second are messages with some concealed content, called stegoforms.

### 7.2.1 · Information theoretic steganography

Because of the enormous variety of possible cover media, and of possible embedding algorithms, much work in steganography (see Anderson, Needham & Shamir 1998, Ker 2005, Sallee 2004) lacks a cohesive theoretic framework that can capture all possible scenarios, instead opting for practice-based descriptions of stegosystems. I have already drawn some attention to the literature on LSB steganography (Dabeer et al. 2004, Mieliikainen 2006, Ker 2007a, for instance). None of the research in this literature proposes a general theory of steganography that is applicable outside of LSB concealment, but instead only describe how information might be concealed in image coverforms.

A more general theory of steganography is proposed by Cachin (1998, 2004). It bases a formulation of stegosystems on existing information theoretic models of communication, which allows Cachin’s model to apply to steganography in diverse settings. Cachin’s infor-

---

4 In the steganographic literature, these are called covertexts. I have chosen to depart from the literature to prefer -form to -text, to highlight the primacy of the linguistic form (that is, utterance) in my model.
mation theoretic model is the preferred basis of the natural language model that I develop because it includes agents’ Bayesian reasoning about the information that is presented to them. I also draw on work by other authors, namely Sallee (2004), Moulin, O’Sullivan, et al. (2003), Hopper (2004), Cox et al. (2005). The information theoretic model is also informed by information theoretic studies of steganalysis, especially those by Zöllner et al. (1998), Westfeld & Pfitzmann (2000). A general information theoretic model of steganography is used because it integrates well with game theory, especially with the error models developed by Benz (2012a,b) and the relevance-centric models of van Rooij (2003, 2004b).

It is important to note that, while the following account of steganography is highly general and can be adapted to describe natural language steganography, it is still concerned foremost with algorithmic stegosystems. Most noticeably, it discusses bits as the basic constituent of both the concealed information and the coverform. Particularly in the latter case, I diverge from the general formulation when I introduce my own model for natural language. Furthermore, I do not present Cachin’s model in full detail, but instead I intentionally omit or alter details that are not salient to natural language.

(1) I omit everything that is not part of Cachin’s formulation of a symmetric-key model of steganography. A model of a stegosystem (and more broadly of a cryptosystem) that is symmetric operates under the assumption that the Sender and Receiver both possess a shared secret that is used in both the embedding and extraction of the secret information in and from the stegoform. An example of an application of symmetric-key cryptography is the use of a password to encrypt some data object. The password acts as a symmetric key, and is exchanged by the agents through a secure channel, which is different from the channel used to transmit the file (see Delfs & Knebl 2007). In computer scientific stegosystems, a key is usually a binary or hexadecimal number. It is a convention that, if a key is used, Shannon’s maxim (Shannon 1949: 662) be adhered to: "the enemy knows the system being used". That is, the security of the stegosystem should not rely on only the algorithm for embedding—for this will inevitably be discovered—but on the shared secret between the Sender and the Receiver(s).

(2) I omit the provision that Cachin (2004) provides for active adversaries. Instead I only describe passive adversaries. If an Adversary is passive, she merely observes the message, and does not alter it before it is transmitted to the Receiver. When discussing natural language, this motivation is clear, since utterances cannot be altered once they are produced.

The stegosystem that Cachin (1998, 2004) defines involves three agents, a Sender S, a Receiver R, and an Adversary E (for 'eavesdropper'). We assume that, regardless of whether S transmits a message with some hidden content, the following series of events occurs: S emits a signal through a public channel to R, and E intercepts a copy of the signal while in transmission. Let us make the simplifying assumptions that (a) the signal that R and E observe is identical in every case; and (b) if noise is present in the channel, the
resultant permutations of the signal occur before (and only before) $E$ observes the signal. As stated above, $E$ has read-only access to the public channel, meaning that she cannot alter the signal before it continues towards $R$. Indeed, it is only $S$ who has write access to the channel, for she is the only agent who can produce utterances that are transmitted across it.

There are two states that $S$ is possibly in. In the case that she is inactive (state 0), she sends an innocent signal that does not contain any embedded information, which is simply a coverform $c$. In the case that $S$ is active (state 1), she sends stegoform $s$. Note that, in both states, the signal that $S$ transmits may be precisely the same. However, if she is active, the signal contains hidden information, and so is more complex. While the coverform is simply a signal from a set of signals, the stegoform is generated by an embedding function $X$. The embedding function accesses both a random source $r$ which is known only to $S$, and a key $k$ which is selected randomly and distributed across a private channel so that it is known mutually to $S$ and $R$. Lastly, the embedding function also takes the coverform $c$. So, $X$ takes $c$, $r$, and $k$ as inputs, and outputs $s$.

Observe that while $X$ embeds a secret message $z$, which is a random variable drawn from a message space $\zeta$, it does not take $z$ as an argument. This notion appears rather counterintuitive at first, but it is used by Cachin to preserve the generality of the model. As I discussed above, and as is the case in general information theoretic models of communication, $s$ ought to give $R$ information about or evidence of the value of $z$, rather than delivering $z$ itself. Such a formulation is advantageous as it allows for the inclusion of hypothesis testing in the model. I discuss this point in greater depth below. The value of the random variables is probabilistic, and each has its distribution. $\delta_c$ is the probability distribution over possible cover forms, $\delta_s$ the probability distribution over potential stegoforms, and $\delta_\mu$, the probability distribution over the message space. As we adhere to Shannon’s maxim, these distributions and the embedding algorithm $X$ are public knowledge, and so are known to every agent of the system.

Figure 7.3 is a schematic depiction of the system.

---

**Figure 7.3**: Information theoretic, symmetric-key stegosystem, from Cachin (2004). The circular switch at the left end of the public channel determines whether $S$ is in state 0 or 1.
The model’s operation when \( S \) is in state 0 is trivial. Both \( R \) and \( E \) observe \( c \). When \( S \) is in state 1, the mechanism is far more complex. \( R \) and \( E \) observe \( s \). \( E \) simply intercepts \( s \) as it is transmitted. If the stegosystem is secure, then \( E \) should not be able to distinguish \( s \) from \( c \). I discuss \( E \)’s decision problem in greater depth below. \( R \) uses an extraction function \( Y \), which takes \( s \) and \( k \) as arguments, to output a value \( \hat{z} \). If the stegosystem is useful, the value of \( \hat{z} \) provides her with some evidence of the value of \( z \). This characterisation is important, for I rely on it extensively throughout the natural language steganographic model. Instead of delivering \( z \) itself, \( S \) tries to deliver something that will give \( R \) enough information to infer the value of \( s \). This move appears to be motivated by Cachin’s desire to maintain a level of generality in the model, but it resonates with descriptions of inferential interpretation that have been given thus far in the thesis. I therefore preserve this inferential basis, and its aptitude in the description of situations of natural language steganography will become clear.

This overview is distilled by Cachin (1998, 2004) into definition 7.8

**Definition 7.8: Stegosystem.** Let \( \delta_c \) be a coverform distribution and \( \zeta \) be a message space. A **stegosystem** is a pair of algorithms \( \langle X(c,k,r) = s, Y(s,k) = \hat{z} \rangle \) iff there exist random variables \( k \) and \( r \) such that for all discrete random variables \( z \) over \( \zeta \) with \( H(z) > 0 \), it holds that \( I(\hat{z}; z) > 0 \). This condition ensures that the stegosystem is useful.

Now, let us turn to the decision-making of agents using the stegosystem. In Cachin’s (1998, 2004) model, the Receiver has an oracle which informs her whether the Sender is active or inactive. Oracles are used in computer science as modelling entities which grant agents knowledge by answering questions. The knowledge is supposed to be reliable: oracles cannot lie, and they have perfect and complete knowledge of the model at every state of its operation. According to Cachin (2004), the Receiver does not need to decide if the transmitted message is a stegoform or a coverform, for she already gains this information from the oracle. Instead, the Receiver’s chief task is determining the value of \( z \) given \( \hat{z} \). The Receiver’s determination is a rather simple instance of Bayesian inference:

\[
\arg \max_{z \in \zeta} \Pr(z|\hat{z})
\]  

(7.11)

The Adversary’s decision problem, instead, involves determining if the emission is a container for hidden information. \( E \)’s perspective of both states is similar to \( R \)’s: any received emission is either innocent, or designed to have the appearance that it is so. There are two possibilities available to her. If \( R \) is in state 0, the probability of the form that she observes is a member of the probability distribution \( \delta_c \). And if \( R \) is in state 1, the form’s probability is a member of \( \delta_s \). The form that \( E \) observes is a datum that may support either of these scenarios, and therefore \( E \) is faced with the problem of determining which possibility is more likely, conditional on her observation of the emission. \( S \) is of course cognisant of \( E \)’s process of detection, and is therefore mindful to design her stegoforms and coverforms to be, ideally, indistinguishable from one another. It is by the degree to
which stegoforms and coverforms are indistinguishable that a stegosystem’s security is measured. Specifically, a stegosystem’s security depends on the relative uniformity of the distributions $\delta_c$ and $\delta_s$. If a given emission is highly probable according to one of these distributions but not the other, the stegosystem is insecure, since the likelihood that the Adversary will make an error is significantly reduced. We can thus measure a stegosystem’s security by calculating the relative entropy of $\delta_c$ to $\delta_s$, as per definition 7.9.

**Definition 7.9:** Stegosystem security. A stegosystem is $\lambda$-secure (against passive adversaries) if $D(\delta_c\|\delta_s) \geq \lambda$. A stegosystem is perfectly secure (against passive adversaries) iff $\lambda = 0$.

### 7.2.2 · Game models of steganography

The selection of a stegoform that maximises the information available for R’s estimate of the message while simultaneously appearing to be a coverform to E, would appear to lend itself quite well to a game theoretic treatment. However, the literature has very few examples of applications of game theory to the steganographic problem, as shown in this section.

Game theoretic studies of steganography have narrow foci, and describe the steganographic problem under specific conditions, rather than in its generality. Ettinger (1998) models image steganography as a zero-sum game of two players where the Steganographer attempts to maximise the amount of hidden data in a stegoform, while the active Adversary attempts to minimise the same metric by introducing distortion (noise) into the stegoform. The distortion is simplistic—flipping digits of the bitstream—and the Adversary is indiscriminate: she distorts every transmission regardless of its contents. Distortion incurs a cost which is proportional to the significance of the bit that is altered. Therefore, the Steganographer’s best strategy is to embed hidden information in the most significant bit without it being apparent that information is hidden, so as to increase the robustness of the channel against the Adversary’s suppression. Ker (2007a) employs game theory to find best strategies for agents faced respectively with the problem of batch steganography, where the Steganographer uses multiple objects to embed hidden messages, and pooling steganalysis, where the Adversary collects these objects and uses them as data to inferentially detect hidden messages. Interestingly, Ker finds that the Steganographer’s best strategies are to either spread the hidden message uniformly through a large number of stegoforms, or else to include the entire hidden message in a single stegoform. Finally, Schöttle & Böhme (2012) focus on steganographic situations where agents have incomplete information of the distributions for generation of the coverform and stegoforms. The authors argue that it is only under these conditions of epistemological poverty that realistic, rational agents are able to make choices concerning the embedding and detection procedures that are used. The agents in Schöttle and Böhme’s model used mixed strategies (that is, selected probabilistically: see §2.1.2). The coverform image is generated by Nature, and is presented to the Steganographer. The Steganographer’s best strategy depends on the heterogeneity of the image, so that she selects the most heterogeneous area of the im-
age (that is, the parts of the image where there is the least uniformity, or the parts of the image with the most entropy) and hides her message in the bits that encode that area of the image. Noise is introduced into the signal, such that the Steganographer’s alterations to the image are reversed with a probability of 0.5. Since Schöttle & Böhme (2012) model steganography as a game of incomplete information, the Adversary in unsure where in the image the Steganographer has concealed information. She must therefore use Bayesian inference. But if the Adversary has changed a highly heterogeneous area of the image, it is more difficult for the Adversary to discern meaningful alterations amidst the noise. That is, the conditional probability of any given area being the correct area to attack given the information available to the Adversary is substantially lowered.

Regrettably, none of these three models is in its entirety of particular use to us here. Ettinger’s model formulates an active Adversary, which is inappropriate for natural language since utterances cannot be altered once they have been transmitted; Ker’s model relies on different epistemological conditions than those that natural language speakers operate within; and Schöttle and Böhme’s relies too heavily on bitstreams to be of use for natural language. There are nevertheless aspects of each which may serve some use in a steganographic model of natural language. The Adversary’s use of pooling steganalysis in Ker’s model is useful in inspiring descriptions of conversations with multiple turns, and the embedding of secret information across multiple utterances. Schöttle and Böhme’s model is useful in its attention to the epistemology of the agents, which is elaborated upon in other research by Böhme (2009).

7.3 · Natural language steganography

Now let us turn to natural language. For the moment, I assume that there are, as in Cachin’s model, three participants in the conversation. I assume that the Sender wishes to convey some information to the Receiver, and not to the Adversary, using an utterance. I enforce the strong condition that the utterance must be comprehensible to all parties to the conversation. The problem that confronts agents in the model is reformulated: how does one transmit a natural language signal (that is, an utterance) across a public channel so that a given interpretation is only accessible to a certain individual, and not to anyone who might happen to observe the same signal?

In this section, I aim to provide the formal apparatus that can model situations that involve concealment of information. In §7.3.1, I introduce a means of restricting which agents have access to specific parts of a common ground, called a permissions system. To do so, I add a specification to common ground DRSs that prescribes a set of players that may view that DRS. I incorporate this version of DRS into a fully-specific formal model of natural language steganography in §7.3.2. This model only describes a natural language stegosystem, and does not apply any formalisms of game theory. Instead, this is taken up in §7.3.3. Finally, the complex array of entities introduced in this chapter is summarised in §7.3.4.
7.3.1 · DRS with permissions

Until this point, I have assumed that the common ground is a singular entity, that is understood in the same way by all participants in the conversation. This assumption is made throughout the literature (see, for instance, van Dijk 2006, Lascarides & Asher 2008, Stalnaker 2002), but it is one that I wish to examine more closely. Specifically, I now wish to attend to Claim B that I made at the beginning of the thesis: *That for every possible combination of the parties to the interaction, multiple, divergent partial grounds are generated and exist within the total discourse.* The partial grounds are called combinatoric, because each combination of interlocutors that is larger than one are a party to a partial ground.

A basic intuition lies behind this claim, which I illustrate in the following vignette. Suppose that two people, Vladimir and Estragon, are known to one another. They become acquainted with a third person, Pozzo, with whom they converse. It is clear that Vladimir and Estragon mutually know more of one another than do, say, Pozzo and Vladimir. This means that, according to what has already been discussed, Vladimir and Estragon have access to information to resolve their interpretations that is not available to Pozzo. Now suppose that Vladimir utters (84).

\[(84) \quad \text{I got the job!}\]

The use of a definite noun phrase suggests that the entity is already present in the discourse (see van der Sandt 1992). But note that this is not necessarily present in the discourse that Vladimir, Estragon, and Pozzo are parties to. Indeed, one might assume that Estragon would know which job Vladimir refers to, while Pozzo would not, unless it was mentioned in their conversation.

A given partial ground is therefore similar to the common ground of the conversation, and likewise represented by a DRS. As such, the usual rules of DRSs apply to partial grounds—that is, the same effort-based information is included in a partial ground as in a DRS—and they are represented in the same way. Unlike DRSs, however, partial grounds stand in relation to the common ground, which is the merge of all the partial grounds. Formally, these are defined as follows.

**Definition 7.10:** Partial grounds and the common ground. Let \(i, j\) be agents, and let the smallest possible set of DRSs of language \(\mathcal{L}\) be denoted \(\kappa = \{K_1, \ldots, K_k\}\). A **partial ground** \(\gamma(i, j)\) is a \(K \in \kappa\). The common ground \(\Gamma\) is redefined as the merge of all partial grounds \(\gamma\). If an agent \(i\)'s partial ground \(\gamma_i\) converges with the total common ground \(\Gamma\), \(\Gamma\) is **accessible to** \(i\).

So, the major difference between the common ground and partial grounds is that only certain agents have access to a given partial ground. To achieve this using DRSs, I introduce a system of permissions that allows certain subsets of the set of agents to view a particular DRS. The permissions system that I propose is based on the file system permissions in the filesystems of Unix-like operating systems. The idea of this system is simplistic: each specifies a protocol whereby only certain users, or certain groups of users, are allowed
to perform certain operations on a specific file. The three basic operations in Unix-like operating systems are (1) **read**, such that the user or user group is able to open and view a file; (2) **write**, such that the user or user group is able to alter the contents of a file; and (3) **execute**, such that the user or user group is able to run a file as a program (see Sandhu et al. 1996). Of these, only the **read** permission is appropriate to DRSs, since common grounds are unable to be executed in even the broadest metaphorical sense, and since common grounds in natural language are mutable by any interlocutor, and only in very specific circumstances may an interlocutor’s ability to alter a common ground be restricted.\(^5\)

The means of restricting access to partial ground DRSs is simply to specify a set of agents that are allowed to read that DRS.

**Definition 7.11**: DRS with permissions. Let \( G \) be a game with a set of players \( N \). A DRS with permissions is a triple

\[
\langle \theta, D, C \rangle,
\]

where \( \theta \subseteq N \) is a set of agents such that \( |\theta| \geq 2 \), \( D \) the DRS’s universe of entities, and \( C \) is the set of constraints on \( D \).

The box notation of DRSs is preserved, though the set of agents with read permission is included in the top compartment of the box with the universe of discourse referents. The permissions set is represented as a set of agents in square brackets. A prototypical example, where \( i, j \in N \) have access to the DRS, is produced in (85).

(85) \[
\begin{array}{c}
[i, j]; x, y \\
\vdots
\end{array}
\]

When two DRSs are merged, their permissions also merge, meaning that the permissions set of the merged DRS is equal to the union of the two DRSs that were merged.

The permissions set is developed in the context of a synthesis of game theory and DRT. The agents that definition 7.11 references are necessarily players of a game, and a DRS with a permissions set is defined as part of the game, rather than as part of a DRS language. And therefore, introduction of a permissions system does not alter the model theoretic semantics of the DRS language that is used in the thesis.

The introduction of partial grounds does, however, affect the definition of game theoretic beliefs. Beliefs with partial grounds are almost identical to beliefs with common ground as defined in §6.1.2.1, save for one key fact: the probability of an utterance is conditional on the partial ground that an agent shares with the recipients of the utterance, if the recipient is the Sender, or with the producer of the utterance, if the recipient is the Receiver.

\(^5\)Improvised theatre is in fact one of the few situations here such restrictions are uniformly operational: the audience’s contributions, such as laughter or heckling, are generally disregarded as valid inclusions for the plot’s common ground in improvised performances. This is an interesting consideration for future work on improvised theatre and common ground, but it is orthogonal to the project at hand, since audience interaction is not of primary concern.
Definition 7.12: Sender’s belief with partial ground \(\gamma(R, S)\). Let \(\kappa\) be a set of possible interpretation selections for the Receiver. A Sender’s belief is a conditional probability distribution with full support \(\rho_S \equiv \delta_{W|\gamma(R, S)}\), such that each element of the probability distribution is a conditional probability \(\Pr(K|\gamma(R, S))\).

Each Receiver’s beliefs regarding \(W\) are conditionalised by the partial ground that that Receiver shares with the Sender. This is the case regardless of the number of Receivers.

Definition 7.13: Receiver’s prior belief with partial ground. Let \(W = \{w_1, \ldots, w_m\}\) be a set of world-states. A prior belief of \(R\) is a probability distribution \(\rho_R(w|\gamma(S, R)) = \delta_{W|\gamma(S, R)} \in \Delta(W)\).

Likewise, the Receiver’s posterior belief is conditionalised by the common ground that that Receiver shares with the Sender, and by the observation of the utterance.

Definition 7.14: Receiver’s posterior belief with partial ground. Let \(W\) be a set of states of the world, \(\rho_R(w|\gamma(S, R))\) be a prior belief of \(R\), and \(m\) a message sent by \(S\). A posterior belief of \(S\) is a fully supported probability distribution \(\rho_S(w|m \cap \gamma(S, R)) = \delta_{W|m, \gamma} \in \Delta(W)\) such that each member of \(\delta_W\) is the probability of a \(w\) given \(m\) and given \(\gamma(S, R)\):

\[
\Pr(w|m, \gamma(S, R)) \equiv \frac{\Pr(m \cap \gamma(S, R)|w)}{\Pr(m \cap \gamma(S, R))}.
\]
duces. Unlike Cachin’s model, there is no oracle that informs $R$ of which state $S$ is in. In Cachin’s model, performing an extraction function on a data file confirms which state $S$ is in: if the output is random noise, there is no concealed information. In natural language, instead, interlocutors must infer what their fellows mean, and there is no trivial way of checking whether one’s inference is correct that does not require a clarification request. I therefore omit the oracle. In my model, both $R$ and $E$ use the same datum to infer the value of the stegoform utterance.

The Adversary. The presence of an Adversary in the model for natural language is somewhat artificial, in that most situations that can be modelled are not adversarial. It is useful however to build the model for extreme cases where the Sender actively wants to hide a given message from the Adversary. The Adversary is not different in any meaningful sense from the Receiver: both are boundedly rational agents who select an interpretation upon the observation of an utterance produced by the Sender. However, as I discuss in §7.3.3, the difference lies in the effects of the Adversary’s presence on the calculation of utilities.

7.3.2.2 · Cover- and stegoform utterances

The order of events in the model is straightforward, and follows that of Cachin’s (2004) model. The Sender transmits an utterance. The nature of the message depends on her state. In state 0, $S$ sends a coverform utterance $c$. In state 1, $S$ sends a stegoform utterance $s$.

Definition 7.15: Coverform and stegoform utterances. Let $S$, $R$, and $E$ be a Speaker, Receiver, and Adversary, respectively. An utterance $m$ with projected interpretations $\bar{K}_m = \langle K_R, K_E \rangle$ is a coverform if $K_R = K_E$, and a stegoform if $K_R \neq K_E$. Denote coverform and stegoform utterances $c$ and $s$, respectively.

To be clear, any given utterance may be both a coverform and a stegoform. The difference between the two is not the constitution of the utterance; instead, the difference lies in the Sender’s intended interpretations. Both a naïve and an active Sender may use precisely the same utterance for different purposes.

Since they are both kinds of utterances, coverforms and stegoforms are drawn from the same message set. This is in keeping with Cachin’s model, which implicitly holds that coverforms and stegoforms are selected from the same set of objects. Were this not the case, then statistical patterns might occur which would allow for successful attacks to compromise the security of the stegosystem. So, let us define two probability distributions over the message set $\mu$, for the appearance of a coverform and a stegoform respectively. Unlike Cachin’s model, the distributions are conditional: since utterances are meant in game theory to reveal something about the state of the world, agents must always select utterances to act as coverforms or stegoforms with this fact in mind. Therefore, the probability of a particular utterance is always conditioned by the world-state that an agent believes is the true world-state. This yields definition 7.16.
Definition 7.16: Coverform and stegoform distributions. Let \( \mu \) be a set of viable natural language sentences, and set a world \( w \in \mathcal{W}_{\text{wr}} \). \( \delta^c_{\mu|W} \) and \( \delta^s_{\mu|W} \) are conditional distributions over \( \mu \), such that each member of each distribution is a conditional probability \( \Pr(m|w) \). The probability mass function of \( \delta^c_{\mu|W} \) describes the probability that a particular \( m \) will be a coverform in world \( w \), and likewise, mutatis mutandis, for \( \delta^s_{\mu|W} \).

When an utterance is produced, \( R \) and \( E \) observe the utterance, and each selects an interpretation. We should note that the particular order in which they make their interpretations does not matter, since neither can know which utterance the other has selected. To remain consistent with steganographic models, the Adversary \( E \) observes the utterance before the Receiver \( R \); though in actual fact, they observe it and make their selections simultaneously. The interpretation problem that faces them is identical: each attempts to guess which \( \mathcal{K} \) is the true one. Thus, interpretation in natural language steganography involves determining which interpretations were meant for other agents, as well as oneself. This means that \( E \) and \( R \) both identify multiple potential interpretations, some of which might be mutually contradictory.

Stack notation. To make the representation of stegoform utterances and the interpretations that are meant for different agents clearer, I present a notation which uses the relative position of DRSs to represent their permissions. To do so, I first define a permission implication function. It gives us a way to compare the permission sets of different DRSs, particularly with respect to whether the permission set of a particular DRS is a subset or a superset of another.

Definition 7.17: Permission implication function. Let \( \kappa \) be a set of valid DRSs. The permission implication function \( \Rightarrow \) is a relation \( \Rightarrow : \kappa \rightarrow \kappa \) such that \( K \Rightarrow K' \) iff read access to \( K \) implies read access to \( K' \); \( K \Rightarrow K' \Rightarrow \theta_K \subseteq \theta_{K'} \). It follows that if \( K' \Rightarrow K \) and \( K'' \Rightarrow K \), then \( \theta_K = \theta_{K'} \cup \theta_{K''} \).

The permission implication function is the basis for a DR theoretic representation of stegoform utterances. I use a notation for stegoform utterances called stack notation. It is predicated on a simple idea: if a DRS \( K \) is above a second DRS \( K' \), then \( K' \Rightarrow K \). I call such structures DRS stacks, and they represent the projected interpretations of an utterance, which is, per definition 7.15, a vector of DRSs. Observe that this vector is not itself a DRS, for it does not strictly represent a partial model of a possible world—as is the case with DRSs that represent other sorts of utterances, or those that represent the common ground. Instead, it should be seen to represent the communicative intention of an interlocutor. That is, a DRS stack is the formal description of a strategy. To be clear, both the ostensible and the concealed messages are treated as normal DRSs, and so the definitions laid out in chapter 3 apply as usual to these parts of the utterance. But the stack itself merely represents the communicative intent.

Let us examine an example of this notation in use. A stegoform utterance is always constituted of at least two DRSs, which represent the utterance’s ostensible interpretation, meant for the Adversary, and the concealed interpretation, meant for the Receiver. For
instance, recall the example given a couple of times in the thesis: S wishes to communicate to E that she enjoyed E’s piano recital, and wishes to convey the precise opposite to R. An utterance that achieves this is represented by the DRS stack in (86).

\[
K_1 = \begin{array}{c}
[S, R, E]; S, x \\
\text{performance}(x) \\
\text{enjoys}(S, x)
\end{array}
\]

\[
K_2 = \begin{array}{c}
[S, R]; \\
\neg K_1
\end{array}
\]

The ostensible meaning, which is available to all parties to the conversation, is that S enjoyed the performance, while the DRS that is available only to S and R negates the ostensible meaning entirely. This situation is captured well by this notation, since the notation overcomes the chief problem with the semantic representation of stegoform utterances: that they may contain self-contradictory information. If the semantics of a stack notation were defined in terms of the conjunction of the ostensible and concealed messages, cases of contradictory meaning, as in (86), will result in counterintuitive truth conditions; that is, (86) will be false, which does not appear to be appropriate for this utterance. Since stegoform utterances behave differently to more conventional utterances, the representation of stegoform utterances should also behave differently to the representations of conventional utterances. Stack notation, then, is appropriate for this class of utterance.

7.3.2.3 · Security

The major concern of any stegosystem is the security of that system against steganalysis. Security is measured in Cachin’s (2004) model by the degree to which an emission sent by S in state 0 is indistinguishable from an emission sent by S in state 1. The inability to distinguish between the different kinds of utterances derives from the Adversary’s lack of grounds on which she may confidently claim that the message was generated with the stegoform distribution instead of the coverform distribution, and vice versa. I will retain this notion, though the nature of concealment in natural language is quite different from that in traditional steganography, and so the model does not use a function.

Stegoform utterances are designed to exploit differences that the Sender perceives to exist between the partial ground that she shares with the Receiver, and the partial ground that she shares with the Adversary. As I discussed in the previous chapter, divergences in perceptions of the common ground lead to misinterpretation. In natural language stegosystems, the Sender wishes to exploit this fact, and attempts to design her utterances in such a way as to maximise the divergence in the partial grounds. Secure stegoform utterances, then, are most easily constructed when little common ground exists between R and E. The reason for this is clear. If the hidden interpretation $\tilde{K}$ relies on information that is available to S, but not available to E, E will not select that interpretation, since the
probability of that interpretation conditional on the common ground is extremely low, if not zero. And since the Adversary also speaks the language of the Sender and Receiver, she will be able to interpret any utterance that is transmitted between the two. So there is the additional security requirement that a plausible interpretation be available to the Adversary. That is, the partial grounds that inform the selection of an utterance cannot be too divergent, since she wishes to conceal the fact that she is communicating covertly at all. If she produces an utterance that cannot be interpreted at all according to the partial ground that she shares with the Adversary, then the Adversary may become wary that covert messages are being exchanged. The simplest case of design is where there is some part of the partial ground that is mutually shared by all agents, such that \( \gamma(S, R) \cap \gamma(S, E) \neq \emptyset \). To be optimally secure, the Sender should design her utterance to target this intersection, since it allows for plausible deniability: by making the utterance available to the Adversary, the Sender ensures that she has an alibi that the communication is in fact innocent.

The information that \( S \) shares with \( R \), and not with \( E \) is called the natural language key in the model.

**Definition 7.18:** Natural language key. *Fix a partial ground \( \gamma \), and let there be three agents, \( S, R, \) and \( E \). A condition \( \phi \) is an **natural language key** iff
\[
\phi \in C_{\gamma(S, R)} \setminus C_{\gamma(S, E)}.
\] (7.14)*

There may therefore be multiple keys. I discuss key management further in §7.3.2.4.

Recall that, in Cachin’s (2004) model, the message itself is not an input to the embedding function: in definition 7.8, the embedding function only takes \( c, k, r \). This is ostensibly to preserve the generality of the model, but it is also felicitous to have a mechanism of indirect effect of the selection of a coverform, since by definition a channel cannot contain a message but may only include a signal that represents the message. Instead, the message affects the selection of a coverform similarly to how the message affects the generation of a signal in information theoretic communication. This formulation is also advantageous for the purposes of this thesis, for it translates well to a game theoretic model of communication. It is impossible for a Sender to precisely and exhaustively convey a state of the world with a natural language utterance. Even in situations where the two states of the world are distinguished by only one fact, such as in the Bomb Dilemma example in §2.1.3, the message still only gives evidence of what the state of the world is, rather than expressing the state of the world in any meaningful sense. Instead, the utterance gives the Receiver evidence that she uses to infer the true state of the world. This is essentially equivalent to Cachin’s model: the selection of an utterance is based on the state of the world, rather than taking the state of the world as a proper input.

As stated, both \( R \) and \( E \) are unsure of \( S \)’s state. This fact causes the deliberation of both to be similar, differentiated perhaps only in the quality of information that is available to each. Otherwise, it is only through observation of the utterance and perception of the partial ground that they infer \( \tilde{K}_m \). As in the model of traditional steganography, each has
two competing hypotheses. First, that the observed utterance is generated by $\delta_{\mu|W}^c$; and second, that it is generated by $\delta_{\mu|W}^s$.

The security of a natural language stegosystem is a measure of the uniformity of the distributions $\delta_{\mu|W}^c$ and $\delta_{\mu|W}^s$. If an utterance is obviously a coverform or obviously a stegoform, then the stegosystem is insecure. Note that the measure of security is a measure of entropy, and so is measured in units. I use base 10, because this preserves consistency with the use of decimal numbers in utilities that has been made so far in the thesis.

**Definition 7.19:** Security of a natural language stegosystem. Fix a message set $\mu$, and three distributions over it: a stegoform distribution $\delta_{\mu|W}^s$, a coverform distribution $\delta_{\mu|W}^c$, and a message set distribution $\delta_{\mu}$. Let $W$ be a set of world-states. A natural language stegosystem is $\lambda$-secure against adversaries if $D(\delta_{\mu|W}^c \parallel \delta_{\mu|W}^s) \geq \lambda$.

This definition captures what has been discussed so far in this section. If a stegosystem results in utterances that are poorly designed, and that reveal information that is supposed to be hidden, then the security is trivially low. I also mentioned a less intuitive vulnerability of a natural language stegosystem: if reaching any interpretation of the message that the Sender sends relies on information that the Adversary does not possess, then it will be obvious to the Adversary that the utterance was selected from $\delta_{\mu|W}^s$, rather than from $\delta_{\mu|W}^c$. Since definition 7.19 is indifferent to which distribution causes the security of a natural language stegosystem to fail, and requires merely that the two differences are indistinguishable, too great or too small a difference between the two distributions may both contribute equally to an insecure system.

### 7.3.2.4 · Key management

As with the problem of interpretation, key management is distinguished in natural language by its symmetry. Since the keys are valid DRS conditions, and since partial grounds are shared by every permutation of the parties to a conversation, key management in cases of the emission of secret information involves a preservation and promotion of differences between the partial grounds that the Receiver and the Adversary respectively share with the Sender. Management of the common ground of a conversation was in essence, but not necessarily in terminology, the subject of the previous two chapters. In chapter 5, agents built plot-worlds, which are in essence special instances of common grounds. Likewise, in chapter 6, agents leverage the common ground to resolve interpretations in preferred ways. We have, in a certain sense, zoomed in from the construction of common ground, to the use of common ground, and now to fine-grained control of access to the common ground.

To construct a formalism of key management in natural language, we can turn to Krifka’s (2007) notion of common ground management. Strategies of common ground management are utterances whose primary purpose is to signal “information about the manifest communicative interests and goals of the participants” of a conversation (Krifka 2007: 17). Krifka points to, for instance, the function of questions, which do not generally add information to the common ground, but express a requirement of one of the parties.
to a conversation for information about information that may or may not be therein. Or, as Repp (2012) argues, acts such as illocutionary negation—that is, negation of what is intended by an utterance rather than of the utterance’s semantics per se—, and illocutionary acceptance are strategies whose target is the status of particular propositions in the common ground. What is relevant about recent work on common ground (of which there is regrettably little; see Krifka 2007, Stalnaker 2002, Repp 2012) to the present project is that it incorporates agents’ deliberation about the status, and by consequence the effects, of the information that is in the common ground; or, as it is understood in this project, the partial grounds that they share with the other parties to the conversation.

The major question of key management in natural language is how information about the common ground is exchanged between the Sender and the Receiver so that the Adversary is unaware of it. Put differently, how might a Sender and a Receiver build a common ground without the participation and presence of the Adversary? Any model of symmetric-key encryption, including Cachin’s model of steganography, requires that agents must have secrets that are exchanged independently of the emission of the encrypted signal. To include the secret with the signal would either render the cryptosystem useless, for either the Receiver would have no way of using the key to decrypt the signal if the key itself is encrypted, or render decryption of the signal trivial if the key is sent unencrypted with the encrypted signal. It is clear then that there must be some amount of shared information between a Sender and a Receiver that is forbidden to the Adversary. We can imagine such a state of affairs without necessarily imagining any particularly purposeful collusion between the Sender and the Receiver. As I alluded to in §7.3.1, differences are always present if the Sender and Receiver have had conversations even briefly in the past that do not involve the Adversary. And so the Sender and Receiver will, in almost every case, share information that is not available to the Adversary. This is also true, we should note, of information shared by the Sender and the Adversary, and by the Adversary and the Receiver. That is, it is not a unique state of affairs for the Sender and the Receiver.

With these facts in mind, I stipulate that secrets per se may only be established prior to game in which they are used as keys for natural language steganography, and cannot be introduced during a play of that game. The reasons for this are analogous to the reasons for not including a symmetric key in an encrypted emission. If a secret is generated and used as a key during a signalling game, it will not be able to be understood by the Receiver, or it will be easily detectable by the Adversary. I must emphasise that this is not equivalent to stating that secrets may not arise in the course of a conversation, which is constituted by multiple signalling games. Indeed, it is the entire purpose of the steganographic model to allow for a sort of subliminal discourse to be entered into by the Sender and Receiver, such that information is exchanged that is, by virtue of its concealment from the Adversary, secret. A secret shared by the Sender and the Receiver opens a space in which the Sender and Receiver are able to trade such information.
7.3.2.5 · A formal definition

Now that the constituents of a natural language stegosystem have been described, we can proceed to a formal definition of the system as a whole. To begin, I present a schematic diagram of the system in figure 7.4.

![Diagram of a natural language stegosystem](image)

**Figure 7.4:** Natural language stegosystem. The switch in the circle decides whether the Sender’s utterance contains secrets (state 1) or not (state 0).

The order of events in the diagram is as follows.

1. A world-state \( w \in \mathcal{W} \) is selected by Nature, and the Sender observes it. While knowledge of the world-state ‘travels’ from the State selector to the Sender, there is a switch on the path, which determines whether the Sender is in state 0 or state 1. The Sender’s state decides how she treats the partial grounds that she shares respectively with the Receiver and with the Adversary. In state 0, where she produces an innocent coverform, she uses the intersection of these partial grounds to design an utterance that will be understood in the same way by both the Receiver and the Adversary. To be clear, this is not to claim that the difference is used to construct a stegoform utterance, but is instead meant to express that the Sender finds the difference of the Receiver’s and the Adversary’s partial grounds and uses this to inform her selection. In state 1, she finds the difference between the partial grounds, which allows her to design a stegoform utterance that will be understood differently by the Receiver and the Adversary. The result of this process is transmitted across the public channel.

2. The Adversary and Receiver both observe the utterance and select an interpretation. The interpretations of both are informed by the partial ground that they each share with the Sender. This effect of the partial ground is represented by the dashed lines, which illustrate that the Sender and Receiver either use the intersection or the symmetric difference of their respective partial grounds with the Sender. Since the partial ground affects both the design of the utterance that is sent and its interpretation, then the partial grounds effectively function as keys in the standard steganographic model. Both of the interpretation processes output evidence of the world-state, denoted \( \hat{w} \); the outputs are utilised by agents in the game theoretic model to infer the true state of the world.
There are two major points of departure from Cachin’s model.

First, in Cachin’s model, the selection or generation of a coverform, and the subsequent design of the stegoform, were performed prior to the switch that determined the Sender’s state. This is possible because coverforms may be created before it is even decided to embed a message in the coverform. One can, for instance, use the text of Sterne’s *The Life and Opinions of Tristram Shandy, Gentleman* to conceal information, even though the techniques for electronic embedding of information were developed centuries after the creation of the novel. Natural language is not able to operate in this way. One cannot produce a natural language utterance and then, post facto, embed information therein. It is necessary the Sender have the information that she intends to embed in mind before the utterance is produced. It is also necessary that this information be used in the design of the utterance. Therefore, the switch of the Sender’s state must be before the Sender comes to design the utterance.

Second, instead of there being a single key that is unilaterally shared by the Sender and the Receiver across a secure channel, the Sender shares a partial ground with each of the Receiver and the Adversary. These partial grounds exist prior to any run through of the model, and are not necessarily distributed in any sense comparable to Cachin’s model. It is chiefly the operation that is performed on both of the partial grounds that differentiates the Sender’s states, rather than the kind of utterance that is sent.

We can now give a formal definition of a natural language stegosystem.

**Definition 7.20:** Natural language stegosystem. Fix a coverform and a stegoform distribution \( \delta^c_{\mu|W} \), \( \delta^s_{\mu|W} \), and a world-state \( w \in W \). A natural language stegosystem \( \Xi \) is a tuple

\[
\langle L, \mathcal{M}_L = \langle W, \mathcal{D}, \mathcal{I} \rangle, \mu, ||\cdot||, \delta^c_{\mu|W}, \delta^s_{\mu|W} \rangle,
\]

iff \( I(\hat{w}; w) > 0 \).

The stegosystem that I have proposed here goes beyond simply stipulating that natural language signals may be used as coverforms. Even the most basic utterance requires a great deal of linguistic information to be properly interpreted. In the following subsection, I turn my attention to the use of this complex apparatus in designing and interpreting utterances.

### 7.3.3 · Stegogames

Stegogames as a general class of games are simply signalling games that include a natural language stegosystem. As such, the elements of a stegosystem are incorporated into the game theoretic model of natural language with common ground.

**Definition 7.21:** Stegogame. A natural language steogame is a tuple

\[
\langle N, S, \Xi, U \rangle,
\]

(7.16)
where: \( N = \{ S, R, E \} \), such that \( R \) and \( E \) are sets of Receivers and Adversaries, respectively;

\[
S = W \times \mu \times \kappa \times \cdots \times \kappa
\]

(7.17)

is the strategy space; \( \mathcal{Z} \) is a natural language stegosystem as per definition 7.20; and \( U : S \rightarrow \mathbb{R} \) is a utility function.

A few remarks should be made about this definition. (1) I have chosen to avoid stipulating that a stegogame may only have three participants, although in this section I discuss mainly games with this number of players. Allowing for any number of Receivers and any number of Adversaries, such that these terms are treated as classes of players rather than individual players, produces a very general model that can accommodate the modelling of a variety of communicative situations. For instance, an analyst might be interested in the effect of multiple Adversaries on the transmission of information to a single Receiver: a very thorough and complex model of improvised theatre has a large number of Adversaries in the audience, and a small number of Receivers who are part of the performance.

(2) Attention should be paid to the definition of \( S \). Since each Receiver and each Adversary selects an interpretation, the definition allows for flexibility in the length of a strategy profile. The number of types in strategy profiles that occur following the message depends on the number of Receivers and Adversaries, which is precisely what this definition stipulates. For instance, if there is one Receiver and two Adversaries, then \(|R|+|E|= 3\), and so a strategy profile will include three interpretation selection moves. (3) There is no semantic or extension functions included in the game tuple, since these are present in the natural language stegosystem. This applies equally to the message set and the world set.

Epistemological considerations. Schöttle & Böhme (2012) note that incomplete information for both Steganographer and Adversary creates the sufficient and necessary condition for a game theoretic treatment of steganography. They argue that neither player ought to have access to the coverform distribution \( \delta^c_\mu|_W \) and that \( \delta^s_\mu|_W \) should be available only to the Steganographer. Their reasoning proceeds as follows. If the Steganographer has complete information about both distributions, she would be able to perform perfect steganography. Likewise, the availability of both distributions for the Adversary, “which would be the case for a strict interpretation of Kerckhoff’s principle”, would grant her the ability of perfect detection (Schöttle & Böhme 2012: 128; see also Wang & Moulin 2008). I concur with Schöttle and Böhme, though it is not information of the coverform and stegoform distributions that players of natural language stegogames require to perform perfect natural language steganography and steganalysis. Instead, information of the respective partial grounds is far more important.

As such, the epistemic conditions of a signalling game, I believe, are closer to those that are required in a natural language stegogame than those of game theoretic treatments of steganography discussed in the information theoretic literature (Ker 2007a, Schöttle & Böhme 2012, Ettinger 1998). Under the epistemic conditions of a natural language sig-
nalling game, the true world-state is the Sender’s private knowledge, and this privacy of the access is known by all players. The Receiver has a prior belief about \( w \) based on which world she considers to be the correct and true possible world. This aspect of the epistemology of signalling games is most simply applied to stegogames, since \( S \) has access to the true value of \( w \), while \( R \) and \( E \) have beliefs thereof. \( S \)’s utterance conveys some information about the value of \( w \), rather than \( w \) itself, a condition which is reminiscent of Cachin’s steganological model, where the output \( \hat{z} \) of the extraction function provides information which allows \( R \) to infer the value of \( z \). The compatibility of the epistemology of information theoretic stegogames and natural language stegogames ends here, a result in large part of the increased complexity of information of interpretation selection. Since the strategy sets of players only include natural language utterances, the epistemology of natural language stegogames must adhere more closely to the model of naïve interpretation presented above.

However, the players’ epistemic conditions are far less clear than it is in the information theoretic literature, for any decision is informed by players’ subjective judgments, rather than by any other, objective criterion. Since natural language steganography relies on the partial ground between \( S \) and \( R \), the distribution over stegoform utterances is Bayesian, with the probability of a given utterance conditional on the dataset formed by the partial grounds. The same is true of \( S \) and \( E \), though \( S \)’s desire to keep secrets from \( E \) means that she exploits the difference across her partial grounds with the two other players in very different ways. Insofar as it will enable her to optimally select an utterance to conceal information from the Adversary and convey it to the Receiver, it is only iff a Sender has complete information of every permutation of the common ground that she will be able to perform perfect natural language steganography. On the other hand, optimal steganalysis relies on \( E \)’s beliefs of the partial ground shared by \( S \) and \( R \), which resembles \( E \)’s lack of knowledge of the key in the information theoretic model. \( E \) may perform perfect detection only if she knows the partial grounds that are shared by (a) \( S \) and \( R \); (b) \( S \) and herself; and (c) \( R \) and herself.

So, the epistemological conditions of players in a natural language stegogame are as follows: (1) \( S \) knows the distribution for a stegoform \( \delta_{\mu|\overline{W}}^s \), where any single probability is conditional on \( S \)’s partial ground with \( R \) and with \( E \), and her beliefs about \( R \) and \( E \)’s partial ground; (2) \( R \) and \( E \) do not have access to \( \delta_{\mu|\overline{W}}^s \); (3) no players have access to \( \delta_{\mu|\overline{W}}^c \), but do have beliefs about it, which follows from signalling games.

The Lewis utility function. The Lewis utility function, which has been used throughout this thesis, may be extended for use in stegogames. What I name Lewis stegogames are stegogames which use a form of the Lewis utility function redefined for stegogames: \( S \) must design her signal so that \( R \)’s move responds to \( w \) and not \( w' \), and where \( E \)’s move responds to some \( w' \). Formally, the function is as follows (reproduced from definition redef:lewisUtility in §2.1.3.1).
**Definition 7.22:** Lewis utility. Let \( s = (w, m, K) \) be a strategy profile of a signalling game. A utility function \( U \) is a Lewis utility function if it holds that for each world-state \( w \),

\[
u(s) = \begin{cases} 
(1, 1) & \text{if } s = (w, m[w], K[m(w)]) \\
(0, 0) & \text{otherwise.}
\end{cases}
\]

(7.18)

That is, if \( R \)’s strategy responds to \( w \), the game is successful.

This utility function has proved to be successful in games that model the use of natural language. The reason for beginning a description of natural language stegogames with the Lewis utility function is intuitively clear. In situations of steganography, \( S \) wishes to convey some information to \( R \), but with the additional condition that the same information is not available to \( E \). Let us suppose that the information that \( S \) wishes to communicate is regarding \( w \). With this condition, the utility function becomes equation 7.19.

\[
u(s) = \begin{cases} 
(1, 1, -1) & \text{if } s = (w, m[w], K_R[m(w)], K_E[m(w)]) \\
(-1, -1, 1) & \text{if } s = (w, m[w], K_R[m(w)], K_E[m(w)]) \\
(0, 0, 1) & \text{if } s = (w, m[w], K_R[m(w)], K_E[m(w)])
\end{cases}
\]

(7.19)

That is, if \( E \) responds to the correct world-state, she always benefits, and she is penalised otherwise. \( S \) and \( R \) have a slightly more complicated chain of preferences. Their most beneficial outcome is one in which \( R \) responds to the correct world-state and \( E \) responds to the incorrect world-state. It follows that their most detrimental outcome is one in which \( R \) fails to respond to the correct world-state, but \( E \) succeeds. In between these two extremes is a third possibility: if both \( R \) and \( E \) responds to the correct world-state, then \( S \) and \( R \) have at least some benefit, for the concealed information has successfully been transmitted.

**7.3.4 · An interim summary**

In this section, I have introduced a means by which formal apparatuses of steganography might be extended so that communicative situations that involve some element of concealment or secrecy are able to be described. To do so, I have posited that:

1. each subset of players unilaterally shares a partial ground, which may be represented by a DRS, such that each DRS is assigned a set of agents that have read access to that DRS;

2. a natural language stegosystem relies on the exploitation of the differences in these partial grounds, such that a single utterance might be interpreted in manifold ways given diverse partial grounds;

3. such a system has two probability distributions over the natural language message set, such that one describes the probability that a particular utterance is selected innocently, and the other describes the probability that a particular utterance is selected as a medium for concealment;
7.4 · Strategies to conceal information

(4) a natural language stegosystem’s security relies on the utterance’s interpretability to all agents, and its suitability for its position in the conversation;

(5) stegogames are a class of game that includes a natural language stegosystem; and

(6) the boundedly rational agents of stegogames select strategies according to the principles of steganographic concealment, which may be reflected in the utility function.

The abstract apparatus of natural language steganography and stegogames having now been described, let us turn to actual instances of covert communication.

7.4 · Strategies to conceal information

Natural language steganography is a phenomenon that pervades the speech of improvised theatre to such a degree that practically every utterance has some embedded information that is to be hidden from the audience. This has been noted in the literature (Baumer & Magerko 2009, 2010, Fuller & Magerko 2010, Magerko et al. 2009), and it is proposed in the same research that the hidden communication in improvised theatre is chiefly concerned with planning of the plot. Indeed, this research appears to suggest that improvised performances are possible if and only if there is some degree of covert communication between the actors. Otherwise, the performance would lack direction, or it would fall apart. In the vocabulary of this thesis, planning communication involves signalling preferences over potential plot trajectories. This is achieved by an iterated process of proposal of information, and subsequent acceptance or rejection. I show in this section that an actor signals her preferences with every utterance, and that actors furthermore expect that every utterance that is produced in improvised theatre signals an actor’s preferences.

The previous data analysis chapters have attempted to provide examples of how specific narrative functions are achieved, and which linguistic phenomena they are most often associated with. In this data analysis section, the data that are presented are instead meant to demonstrate the existence and the extent of the phenomenon in improvised theatre, and to give the reader a concrete idea of how the model that has been introduced operates in concrete situations.

The first question that we might ask is naturally what kind of information are actors transmitting covertly? In most of the data that have been discussed in the previous two chapters, the actors appear to be most concerned with signalling their plans for and preferences over the plot trajectory. In chapter 5, I addressed the number of ways that actors build worlds through the signalling of their preferences; and in chapter 6, I addressed how actors resolve the interpretation problems that arise when they do so. While there was variation in each situation that was described, there was a goal that was basic and pervaded all of these communicative situations: the proposal of information about the plot or its existents, and the subsequent acceptance or rejection of this information. To do so, actors use the offer–response architecture of improvised theatre that was identified by Baumer & Magerko (2010). But it is important to observe that the existence and execution
of this pattern of offers and responses, and of the states of the scene that they entail, are ideally hidden from the audience. The vast majority of speech acts in improvised theatre, then, covertly negotiate some metanarrative fact, in the Arundalean sense (Arundale 2006, 2008, 2010): an actor $i$ produces an utterance that projects a particular interpretation, and a second actor $j$ responds in a way that provides evidence for a particular interpretation. And so the basic covert communication that actors engage in involves an offer—which proposes that a certain piece of information enter the common ground—and a response, which gives evidence for one of the scene state operations.

Consider, for instance, the opening lines of datum 3 in section §5.2.1, reproduced in (87).

(87)

1 A well they fired me
2 (1.1)
3 P xxx
4 B: why’d you show up to ↑work then dad

A’s utterance posits a fact about the state of the world. In the plot-world, this is taken as a statement about the actual state of affairs. At the metanarrative level, however, we might embed this clause within another to express the covert communication.

(88) I propose that they fired me.

The predicate fired $(x, i)$ is proposed. Observe that this occurs in concert with the use of presupposition that was documented in §5.2.1.

The hidden content of B’s utterance is more complex, since it is both an affirmative augmentative response, but it is also a subsequent offer that proposes two facts about the plot-world.

(89) I accept your proposal AND I propose that you (and by implication, we) are at work AND I propose that you are my character’s dad.

The method of determining the hidden messages is as follows. By heuristic 9\(^6\), the response is positive precisely because B does not contradict any of the conditions that are asserted by A. The two facts that are asserted by B’s response are proposals of new information. Neither of these facts is established prior to B’s utterance. And so, B’s is necessarily a proposal that these facts be included in the common ground of the scene.

I show in this section that all covert communication in improvised performance is reducible to the proposal, and subsequent acceptance or rejection, of information. The data suggest that this is a fact that all actors are aware of. That is, actors appear to expect that any utterance in improvised theatre is used to propose information about the plot-world, and/or to accept or reject information that has been proposed by another actor. This expectation exhibits a powerful effect on actors’ interpretations of their fellows’ utterances.

\(^6\)Heuristic 9 states that “A response strategy is by default positive.” See §6.3.3.
This, in turn, affects the design of utterances. I therefore use these communicative functions as the basis of this chapter’s description of covert communication.

Fortunately, this kind of communication is able to be represented easily using stack notation. Using stack notation, we may summarise the different types of offer–response pairs that are possible, and their prototypical representations. To do so, let us use a standard situation: suppose that A proposes that B’s name is Bob, which is represented by $K_1$, and suppose that B responds by proposing that they are in a bank, represented in $K_3$. Note that I am not concerned with the precise utterances that are used, and I only use these examples to introduce the notational conventions. These proposals are intended by the respective actor to be included in the plot common ground, but the act of proposal itself must be hidden from the audience P. This is represented in (90).

$$K_1 = \begin{array}{|c|} \hline \text{[A, B, P]; B} \\ \text{has_attr(B, NAME, Bob)} \\ \hline \end{array}$$

$$K_2 = \begin{array}{|c|} \hline \text{[A, B]; A} \\ \text{proposes(A, K_1)} \\ \hline \end{array}$$

Note that the predicate proposes has two arguments: an agent who utters the proposal, and a DRS that expresses the proposed clause.

A response that accepts A’s proposal, while making a second proposal that the characters are at a bank, is simple to represent in the same way.

$$K_3 = \begin{array}{|c|} \hline \text{[A, B, P]; x, A, B} \\ \text{bank(x)} \\ \text{be_at(A, B, x)} \\ \hline \end{array}$$

$$K_4 = \begin{array}{|c|} \hline \text{[A, B]; A} \\ \text{accepts(B, K_1)} \\ \text{proposes(B, K_3)} \\ \hline \end{array}$$

This response is positive. Suppose that B rejects the attribution of a name to her character, that is, she responds negatively, and yet continues to assert that the characters are at the bank. What results is the following stack.
It is in this way that stack notation can be used to model both the ostensible meaning of utterances, which covers what DRSs have been used for up to this point of the thesis, as well as the concealed information that explicitly regards planning. Observe that the concealed messages represented in (90), (91), and (92) are limited to only three conditions: propose, accept, and reject. I propose below in §7.6 that these are the only three possible conditions in a concealed message in improvised theatre. I suggest that this arises from the actors’ expectations of the narrative function of any given utterance.

An important question that we might ask is the following: if most (if not all) utterances that are produced as part of improvised performance contain some concealed content, then is the security of the stegosystem always compromised? It is certainly the case that the security of the stegosystem that actors employ is extremely weak. Audience members who are familiar with improvised theatre generally understand that there is a degree of communication that passes between the actors, and that is not meant for the audience. But it appears that part of the enjoyment of improvised theatre is derived precisely from watching this sort of linguistic and cognitive acrobatics: actors set themselves complicated situations which they need to navigate, and problems they need to solve, all extemporaneously before the audience. So, I propose that it is not the case that the audience is entirely ignorant of covert communication, but it is also not the case that the audience is entirely suspicious and is interested, as the warden was in the steganographic prisoners’ dilemma, in discovering what is communicated.

The security of a linguistic stegosystem that operates in improvised theatre is not required to be as robust as machine equivalents. It suffices in improvised theatre that the actors do not violate the verisimilitude of the plot-world. So, the security requirement for a natural language stegosystem for improvised theatre is precisely that which preserves the integrity of the plot-world. Obvious metanarrative communication between actors, for instance, would violate this security requirement, since it interferes with the believability of the world. And yet, the weakness of the security requirement of improvised theatre does not, I believe, make the covert communication of improvised theatre an uninteresting object of study. Indeed, it may be the case that a low threshold of security makes it a more interesting communicative situation to study, since actors must communicate while telling a story. In Simmons’s prisoners’ dilemma, the warden was never addressed directly, and so the prisoners were able to send seemingly innocent messages without the warden nec-
essarily needing to understand anything they discuss, provided that the obscurity of their discourse does not betray their secrets. This is not possible in improvised theatre. The purpose of improvised theatre is to construct credible plots, and so the plots need to mean something to the audience. It is finding a balance between clear, covert communication between actors regarding the plot and believable dialogue between characters that makes improvised theatre worthy of serious study as a genre of naturally occurring speech.

To reflect the weakness of the natural language steganography in improvised theatre, I turn away from a Lewis utility function in favour of one that is more appropriate for the linguistic genre. Since it states that the Sender and Receiver suffer losses if the Adversary detects the concealed message, and since the Adversary benefits unconditionally if she successfully detects the concealed message, it implies that the Sender and Receiver are in direct conflict with the Adversary. In improvised theatre, this is not the case, for the audience is willing to suspend its disbelief. In a natural language stegosystem, this equates to a willingness to interpret the utterances on face value. And so the utility function that is used should stipulate that, as long as the audience is provided with the opportunity for a plausible interpretation, the concealment of metanarrative information is successful. A utility function of this kind may be couched in the same terms as the Lewis utility function given in equation 7.19 in §7.3.3 above.

\[
 u(s) = \begin{cases} 
 (1, 1, 1) & \text{if } s = (w, m[w], K_R[m(w)], K_P[m(w')]) \\
 (0, 0, -1) & \text{if } s = (w, m[w], K_R[m(w)], K_P[m(w)]) \\
 (-1, -1, 1) & \text{if } s = (w, m[w], K_R[m(w')], K_P[m(w')]) \\
 (-1, -1, -1) & \text{if } s = (w, m[w], K_R[m(w')], K_P[m(w)]) 
\end{cases}
\] (7.20)

This formulation of the utility function provides a strong disincentive for the audience to detect the hidden communication of the actors; the audience is penalised for detection, regardless of whether the Sender and Receiver communicate successfully. Like the Lewis utility for stegogames, the Sender and Receiver prefer to communicate successfully and conceal information from the audience. They are penalised if the audience detects their concealed communication, but not as heavily as the audience: it is more important to the actors to be understood than not. The least-preferred option for the Sender and Receiver is to fail to communicate successfully, regardless of whether they conceal the metanarrative information. The descending chain of preferences for the actors that arises is the following: the actors communicate successfully and conceal metanarrative information; the actors communicate successfully but fail to conceal metanarrative information; and the actors are indifferent whether they succeed at concealing information if they fail to communicate. For the audience, on the other hand, it is always better to not detect the concealed information.

This utility function represents a situation in which the audience is extremely forgiving. The audience’s best strategy is to select an interpretation that does not contain any metanarrative information whatsoever, and they will attempt to do so at each period. And
therefore the actors only need to supply sufficient information for such an interpretation to exist and the audience will select it.

In the following subsections, I discuss the linguistic strategies that actors use to make their utterances appear as naturalistic as possible, while ensuring that what they wish to convey is clear enough for the other actor or actors. In §7.4.1, I analyse the use of simple declarative utterances as containers for concealed information. In §7.4.2, I focus on the more complicated use of questions for the same purpose. Finally, I turn to even more complicated strategies still in §7.4.3, in which I describe the use of implicature to conceal information.

7.4.1 · Declarative utterances

Perhaps the simplest place we can begin to document the inclusion of concealed information in utterances in improvised theatre is with a description of declarative utterances. Let us consider datum 19.

**Datum 19** Beginning of 4-C1.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A i'm a wizard thompson (1.1) why do i have to go to the post office to pay my bills?</td>
</tr>
<tr>
<td>2</td>
<td>B well, master, well, well, master, they---they need to see i.d., you see? they need to see i.d..</td>
</tr>
<tr>
<td>3</td>
<td>A aye thompson (0.3) i make people disappear (0.5) i can fly</td>
</tr>
</tbody>
</table>

This is an extremely simple example, where A predicates something of his own character, namely that he is a wizard, and makes another attribution of B, that she is named Thompson. B’s utterance appears to accept both of these facts: it augments the first, and it remains neutral on the second, and appears to accept it tacitly since no objection is raised. Note that of course A’s utterance also conveys a fact about the plot-world through his use of a question, which is subsequently handled by B’s response. I return to this question-answer pair below in §7.4.2, and for now I focus only on the self-attribution of A’s being a wizard, and on the attribution of a name to B.

The two attributions that A makes are simple, and are represented in (93). For the sake of simplicity, I ignore the other predications made in the sentence.

\[
K_1 = [A, B, P]; A, B \\
\text{wizard}(A) \\
\text{has_attr}(B, \text{name}, \text{thompson})
\]

This is a clear and rather simple case of proposal. A’s offer proposes that these facts be accepted into the common ground, and that they become facts about the plot-world in general. As in the majority of linguistic moves in improvised theatre, the knowledge of the proposal is restricted to the actors alone. So, as with the exemplar above, line 1 of datum 19 as a stegoform utterance is represented in (94).
As I noted above, B appears to accept both of these proposed predications. Acceptance is represented in a very similar method to proposal. Recall that at present I deal only with the offer of and response to A’s utterance, and not with the contents of B’s subsequent utterance. (95) illustrates this succession.

The offer–response pair, and so the process of negotiating the interpretation of the first line of the datum, concludes in line 3. Importantly, A’s utterance affirmatively augments both of the proposed facts of A’s utterance in line 1, while not responding to the facts introduced by B. When the interpretation becomes operative, it is merged into the common ground. Since this is the beginning of the scene, and so since the common ground is empty, the facts are merged free of any conflict.

An interesting observation of the first line of datum 19, which applies to declarative sentences generally, is that while they appear to be statements of fact, they are still offers in the improvised theatrical sense, and so require some response to be accepted as true in the plot’s common ground.

We see a similar use of declarative sentences in line 16 datum 20 overleaf. The other lines of the datum are presented to give context to the utterance. The utterance in line 16 is a declarative statement that is used by B to assert some fact about the state of the
world, namely that she is prevented from showering because of the threat posed by birds. It is similar to (94), in that a predication is made of the character, and the predication is self-assigned. But instead of being a simple attribute, the utterance in line 16 references a continuing state of affairs, and suggests that this state of affairs precedes the scene. This statement requires a response which validates its inclusion in the plot common ground, which is the adjacent utterance in line 18. A’s and B’s utterances in lines 16 and 18, respectively, constitute a typical example of an offer–response pair with an offer that shifts the state, and a response that affirmatively augments the offer. The hidden information of the utterance of line 16 is a proposal that A’s character has not been able to shower for a certain amount of time. While the proposed fact is accessible from the ostensible meaning of the utterance, the fact that it is proposed by the actor for inclusion as a fact of the plot-world is not. B’s utterance in line 18 accepts the predication, and so it is free to enter the common ground without conflict or uncertainty. B accepts the offer, but the acceptance is only meant for the other actors who are a party to the performance, and not for the audience. Instead, the audience’s interpretation is guided towards a justification for what may seem like an arbitrary rule imposed by B on A.
How is this achieved? I stated above that actors expect any utterance to conceal information regarding the plot trajectory. This expectation is not necessarily shared by the audience. To account for this formally, I propose in §7.6 that the actors share a partial ground which dictates that any utterance include a hidden metanarrative message. This \textit{a priori} partial ground acts like any other partial ground in guiding the interpretation, and functions in the natural language stegosystem as the key that is shared by the actors. It operates in the examples I give for the rest of this section, and I describe it more fully in §7.6.

7.4.2 · Questions

A common means of signalling plot trajectories is to frame the proposal as a question. Questions can be effective since they propose the introduction of elements of the plot across the subliminal channel, while appearing ostensibly to be a request of information by a character in the scene. Questions may function, then, in two seemingly contradictory ways: a covert transmission of information, and an ostensible signal of a lack of information.

To illustrate this, let us consider datum 21. B’s questions in line 2 fulfil both of the functions of questions that I have identified.

Datum 21 Middle of 2-A1. The indices are consistent with datum 4, discussed in §5.2.1. Note that there are some grammatical errors in the actors’ speech, to emphasise that the characters are Mexican. These have been preserved in the transcript.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A ((speaks in a typically Mexican accent)) O! I don't make it ... I don't make it so good now because everybody want som-- salt! Too much salt in this days.</td>
</tr>
<tr>
<td>2</td>
<td>B ((speaks in a typically Mexican accent)) Oh, darling! Are you feeling sad again about your sauce? Oh, darling. Do you cry?</td>
</tr>
<tr>
<td>3</td>
<td>A ((begins to cry))</td>
</tr>
<tr>
<td>4</td>
<td>B Do you cry you baby?</td>
</tr>
<tr>
<td>5</td>
<td>A Jorge, don't do that to me, don't make me little like you’re bigger than me...</td>
</tr>
</tbody>
</table>

Ostensibly, the questions appear to be requests for information, though perhaps a somewhat strange one that is probably only likely in an intimate context. But at the subliminal level, the questions posit something about A: that A is crying.

What is surprising about this proposal is that prior to B’s utterance in line 2, A does not cry, and it is only after B’s assertion that A cries that he begins to do so. Ostensibly, B’s utterance requests information about why A is upset, but communicated between the actors is a signal that informs A that B would prefer a plot trajectory where A cries. A’s response is to cry, which accedes to B’s offer. Not only does this example demonstrate well how questions may be used by actors to covertly convey instructions, it also illustrates the pervasiveness of the practice in improvised theatre.
Questions are used in improvised theatre in ways that are apparently similar to naturally-occurring speech. Krifka (2007) observes that questions are often used by an agent to signal that agent’s requirement for information. In naturalistic genres of naturally occurring speech, the information that is needed is commonly that which the agents require in order to felicitously resolve the interpretation of an utterance. In the genre of improvised theatre, while being more complex, the function of questions is in essence the same. As a question asserts the existence of the entities that are the subject of the question, it also signals a need for direct feedback regarding the status of the interpretations that are available for the question. A clear instance of this is in datum 19, in which A uses a question to propose a plot trajectory. (97) reproduces A’s first line.

(97) A: i’m a wizard thompson (1.1) why do i have to go to the post office to pay my bills?

Observe that, like B’s line 2 in datum 21, this utterance conforms to the property of questions that I have noted. It appears to be a request for information; in this case, the wizard’s information for why the absurd situation has arisen that he has to do menial tasks such as pay the bills in person. But it is also a statement of an intention to bring about a plot trajectory framed around such a conflict. B’s response in line 2 of the datum, reproduced as (98), appears to validate the proposal. The offer of an answer to the question implies both that the question has been taken up, and also that the premise of A’s offer has been accepted.

(98) B well, master, well, well, master, they---they need to see i.d., you see? they need to see i.d..

The uses of questions that we have examined so far are salient for performers of improvised theatre because they are particularly apt at introducing information about the plot-world. For instance, asking a question about an entity or a fact presupposes that that entity or fact exists. In datum 21, a question is asked whose covert purpose is to make an assertion that causes another actor to perform a certain action. In datum 19, asking a question about why the wizard needs to go to the post office to pay bills despite his being a wizard presupposes that there is a post office to go to, that there are bills to pay, and so on. In both of these examples, questions are used not only as a strategy in an offer–response pair, but also have implications for the information globally of the plot-word.

7.4.3 · Implicature

The final general class of utterances that I wish to discuss is the most problematic, since it is the class whose utterances have the most overlap with the hitherto described steganographic model. Implicated meanings are here taken in the broadest sense, as “a component of speaker meaning that constitutes an aspect of what is meant in a Sender’s utterance without being part of what is said” (Horn 2004: 4). One immediately notices the problem here in drawing a boundary between what is an implicated meaning and what is a steganographically concealed meaning. Like implicatures, steganographically concealed
meanings are meant by a speaker without literally being part of what is stated. And so any steganographic meaning might be considered at first consideration a type of implicature.

But a more tenable position, and one that I prosecute here, is that a boundary does not need to be drawn, and in fact it appears that implication is used as a means of increasing the possible range of meanings that may be concealed. If we consider the concealed meanings of the utterances so far, they have been fairly simplistic, and revolve most often around three functions on propositions: the actor may propose, accept, or reject a proposition that is linked to the plot. As I propose in §7.5 below, a satisfactory explanation of this fact is that since both the ostensible and hidden meanings of an utterance must make sense to the respective agents, the range of possible propositions that may be embedded within an utterance is fairly narrow. Senders must design natural language utterances that are at once (a) interpretable by both the Receiver and the Adversary; (b) appropriate at a certain point in a conversation (that is, not overly tangential); and (c) designed to cause the Receiver and the Adversary to interpret utterances differently. The complexity of this task means that the information that is able to be embedded is closely bound to the ostensible meaning of the utterance. But interlocutors in natural language conversations need not only use the ostensible meaning when they speak, but may use meanings that arise through the use of implicature.

Consider, (99), which is lines 20–22 in datum 7, and which was discussed at length in §5.3.1.

(99)
A: THAT’S A REAL KNIFE
B: YEAH i got it when i was cutting cheese earlier

To recapitulate the premise of the scene, A and B are watching a horror film. It is implied that the scene is set at the home of one of the characters. B becomes the murderer in the horror film, and produces a knife. It is important to note that, as I discuss in §5.5.3, the plausibility of the scene’s setting in a naturalistic setting (that is, friends sitting at home watching a film) is challenged by B’s production of a knife: it seems odd that someone would be sitting on a couch holding a knife. I suggest, then, that A’s utterance is a request for justification for the inclusion of this fact in the plot’s common ground framed as a declarative utterance, that effectively acts as a question which expresses the same. A appears to signal that the introduction of a knife into the scene is not plausible according to the common ground of the scene that is at that point established, and appears to demand that B be accountable for the introduction of this fact. The request of this information is an implicated meaning, since it arises without being explicitly stated, but also because it may be cancelled. A might say something like (100) to explicitly ‘override’ the meaning that arises from her utterance in (99).

(100) No, I meant that’s a good quality knife, not the crap you usually buy!
This example, amongst others in the dataset, suggests that implicature is used by actors to extend the possible meanings that a stegoform utterance may take. But even more fundamentally than this, implicated meanings may also act as the concealed information in stegoform utterances. Stegoform utterances that contain an embedded implicature behave precisely like those that embed a normal utterance. The implicature may be available as part of the ostensible meaning of the utterance, it might be available as the concealed meaning, or both of these meanings could be implicated.

7.5 · The advantages and the limits of steganography

The data that I have presented suggest that the concealment of information in utterances is a widespread practice in improvised theatre. I have demonstrated that a steganographic model is capable of describing the practice formally. The state of the art in game theoretic linguistics does not allow, as far as I can see, a method of accounting for the function of a question in improvised theatre, where to one observer, the statement is a request, and to another, the statement is an instruction, such that the Sender has designed the utterance to achieve both purposes. It is certainly possible that two agents may have divergent beliefs that cause them to understand an utterance in radically different ways under current game theoretic models. But it is impossible, I believe, that this can be explained as an intentional, and especially desirable, state of affairs.

But we need to be careful not to use steganography for every conceivable communicative function, and must be circumspect to defer when appropriate to the formalisms of game theory. Consider, for instance, (101), which includes lines 18, 22, and 26 of datum 16 (the intervening lines are irrelevant) in §6.2.1.

```
C: what d'ya mean like we're [(0.4) proJECtions ]
D: [PPPPPPPPPPPPP]> ((i.e.,
   elephant trumpet noise))
P  XXXXXXXX
B: ca ca:wa
(101)  P: [####  ]
C: [(of the) ZOO?]
B: ca ca:wa (1.0) ca ca:wa ca ca:wa
D: PFFFFF
P: XXXxx
A: NO MAN NO (0.2) that's more [evidence] for my idea
```

Recall that earlier in the scene A posits that something is odd about the zoo, and offers C a turn to provide a suggestion of what is odd. C proposes that the zoo animals are all merely projections of animals. A appears to dislike this idea so much that he attempts throughout most of the remainder of the scene to overturn this fact, even when C has begun to try to move the plot of the scene forward. The lines in (101) are part of this portion of the performance. A’s line, and those that are similar in datum 16, signal with a great degree of
clarity and force, that he prefers a plot trajectory which includes the fact that the actors are in the zoo, rather than one where the zoo animals are projections.

While A’s communication attempts to bring about a change in plot trajectory, I do not believe that this should necessarily be included in the concealed information that the stegoform utterance contains. There are two reasons for this.

First, it is implausible that the hidden message is able to greatly deviate from the ostensible message. We have seen so far that all of the examples in improvised theatre appear to be embedded within proposal, acceptance, and rejection clauses. Likewise, in hypothetical examples, we have seen that the negation of DRSs may be concealed within the positive statement of those DRSs, such as in the example where a speaker tries not to offend the pianist (see example 1.1 in the introduction to this chapter). Yet observe that, even if this utterance seems to demonstrate a hidden and ostensible meaning that are highly distinct, the hidden meaning only differs at the level of a single logical operator from the ostensible meaning (that is, negation), instead of, for instance, containing predications of some other entity in the discourse’s universe entirely. It is inappropriate, then, to suggest that information about an actor’s preferences over the plot be included in the hidden contents of the message. When modelling utterances, only the proposal, acceptance, and rejection of information should be included, and not descriptions of the motives for selecting that utterance. If we are modelling (101), the hidden message of A’s line should be a rejection of C’s proposal that the characters are projections, instead of a message about the plot trajectory.

The second reason for restricting the information that a stegoform might contain is that, even if the hidden information of an utterance is restricted to a narrow range of possible meanings, the preferences of an actor are still able to be inferred by her fellows because the utterance that is produced will contain information about the actor’s manifest preferences. For instance, A in (101) appears to prefer the plot trajectory where the humans are exhibitions of the zoo, rather than the plot trajectory where the humans are projections. A produces a line that illustrates why his plot trajectory is more plausible according to the known facts of the plot-world. This signals his preference for a specific plot trajectory, but this signal is not part of the utterance’s semantics. Instead, the signalling occurs in the same way that signals give the Receiver evidence of the Sender’s intention without encoding the intention itself. That is, actors select their stegoforms precisely in such a way as to signal their preferences, regardless of whether the utterance per se includes such information.

While I have discussed the use of implicature as a strategy for concealment in the previous section, it is important to realise that its use also has limits. In some situations, it is inappropriate to use a steganographic mechanism where the standard Gricean explanation of implicature (see Davis 2014) is sufficient. In the opening sentence of his PhD thesis, Franke (2009: 2) notes “it is a near-platitude that under normal circumstances we reliably learn more from observing the honest utterance of a declarative sentence than we would learn from the direct observation of infallible evidence that the proposition ex-
pressed by that sentence was true” (emphasis is the author’s). It is the performance or
production of an utterance that endows it with meaning beyond simply the semantics of
the utterance itself. Franke gives the example of the sentence

(102) It’s raining.

uttered by John, who stands by a window. One can look outside and see that it is raining,
but John’s act of uttering this sentence might inform the observers of John’s performance
that

(103) a. John advises we should take an umbrella, or that
     b. John (hereby) declares the picnic cancelled, or that
     c. John is sick of living in Amsterdam.

The hidden information content of an utterance constituted of a declarative sentence is re-
lated to certain information that may be obtained by witnessing such an utterance. These
instances are not necessarily steganographic situations, since, as Franke (2009) demon-
strates well using his Iterated Best Response model, these meanings may be inferred by
agents who assess the Sender’s reason for producing an observation that is known and
obviously observable to anyone within view of the window that John stands by.

As we discovered in § 7.4.3, implicature may be embedded within stegoform utterances.
But it is important to emphasise that implicated meanings have the same narrow set of pos-
sibilities as explicit meanings do. I have identified that actors seem to be limited to three
types of hidden message: proposal, acceptance, and rejection. These each take proposi-
tions, which are most probably the proposition expressed by the ostensible meaning given
the complexity of embedding in natural language steganography. So while implicature
may be incorporated into the model of natural language steganography, the latter is not
best explained by the former.

Steganography has limits, which are dictated simply by what it is possible to conceal
within an utterance. And furthermore, modellers must be careful not to introduce elements
that reduplicate mechanisms that are already well established. I provide a more formal
definition of the limits of steganography in §7.6.

7.6 · Heuristics for concealed communication

In the following three subsections, I propose three heuristics that describe optimal strate-
gies in improvised performance games that include a natural language stegosystem. In
§7.6.1, I posit the existence of a partial ground that is always prior to an improvised per-
formance. It is not suggested that the complete constitution of the common ground be
defined, but instead that it stand for the actors’ knowledge of their own status as an actor,
and of the mutual recognition of their fellows as the same. The introduction of this partial
ground gives us a reasonable basis to suggest that actors share enough common ground to
make covert communication possible, even in instances where no information about the
plot has been established. In §7.6.2, I propose a heuristic that specifies the three metanarrative moves that can be concealed in natural language utterances in improvised theatre: proposal of a plot trajectory, and acceptance and rejection of a proposal. No other metanarrative moves were found in the dataset. In §7.6.3, I return to the notion of verisimilitude that was discussed in §5.5.1. In that section, I recognised that the model did not have the capability to effectively define verisimilitude. Now that the model includes steganography, a key part of which is that the utterances produced be verisimilar, we may return to this notion and formally define it.

7.6.1 · An a priori common ground for actors

In this chapter, and in others, I have suggested that any subset of the set of actors involved in an improvised performance shares a partial ground that is nonempty, and whose existence precedes any given improvised performance. This partial ground is necessary for the correct functioning of the steganographic mechanisms that have been identified in the speech of improvising actors, for if it did not exist, then there would be no difference in information for actors to exploit in order to conceal information. The partial ground that any two interlocutors share necessarily depends on their shared history, familiarity, and manifold other factors. It is not tenable to include such information in the model, and nor will it increase its efficacy; any game theoretic model is predicated on abstracting away information that would unnecessarily increase the amount of noise in the model’s findings.

So, to counteract the unnecessary introduction of noise, I introduce a heuristic that stipulates a partial ground that is unilaterally shared by any and every pair of actors that takes part in an improvised performance. The partial ground is minimal, in the sense that it is the smallest context that allows natural language steganography to operate in improvised theatre.

Heuristic 10: An a priori actor partial ground. For every subset of the set of actors involved in an improvised performance, there is always a partial ground that they share, such that the partial ground contains the information that

1. the other agent is an actor;
2. the activity they are performing involves constructing a plot;
3. the structure of communication in the genre is that of iterated offer–response pairs; and
4. there is an audience present.

There is no formal representation for the prior actor partial ground, for it is not a definition of optimality. To summarise this information, however, we can use a predication that is shorthand for this information, which applies to the actors in the performance: actor(A). Since this is the most basic fact, and since it implies the other information in the discourse mentioned above, it can stand for all of this information.
7.6.2 · Proposal, acceptance, rejection

The data that we have examined in this chapter suggest that there are three basic kinds of concealed messages, each of which takes a DRS as its complement. The three types are proposal, acceptance, and rejection. These three options are the only available metanarrative functions that I define in the model. This decision has been made because these three function allow the model to sufficiently account for any given concealed message in the discourses that are found in the data. Note that this does not imply that other metanarrative functions do not exist. It may be the case that actors utilise other functions that have not been identified, or (perhaps ironically, given the discussion at hand) cannot be detected from recordings alone.

For a concealed meaning to be successfully identified as any one of these, it must be the most probable of the three options. This, as with most of the formal definition of the heuristics, is expressed as a Bayesian probability.

**Heuristic 11: Metanarrative function.** An embedded meaning is optimally functional if the probability of one of the predicates proposes, accepts, or rejects being the embedded meanings, conditional on the actors’ shared partial ground, is greater than the respective probabilities of the other two.

The formal definition is simple.

**Definition 7.23: Optimal functionality, from heuristic 11.** A stegoform $s$ is optimally functional if there is an $o \in \{\text{propose, accept, reject}\}$ such that, for each of the other functions $o'$, $\Pr(o|s) \geq \Pr(o'|s)$.

This formal definition simply states that an utterance is optimally functional if a particular function is greater than either of the other two possible functions.

This heuristic allows us to account for the data that we have observed in this chapter, but it also permits us to model the occurrence of certain errors that have been observed. It can be used, for instance, to describe the mistake committed in the datum 16 in §6.2.1, which takes place in a zoo. The relevant lines are as reproduced in (104).

(104) A: o look at that (0.6) pa\-nda (1.0) .t aw look at those
dee\-r i (0.5) i LOVE the zoo ma:n (.) the zoo just really
(0.6) DOES somethin for me ma:n i mean like
(2.1)

(104) B: [ca ca\-w ] ((i.e., a bird call))

(104) A: [hey have you th-

(0.8)

(104) C: ye\-h (1.3) i mean (.). what if they're not REAL (.). that's
what you're getting at [isn't it what ]=

---

7I point out in chapter 7 that future game theoretic research on improvised theatre might include interviews with the actors following a performance, so that their decisions can be interrogated.
There are two important factors of (104) that should be reflected upon: the fact that this exchange occurs at the beginning of the scene, and the presence of the 0.8-second pause in line 5. The first is important because the first linguistic move of a scene is necessarily a proposal, since there is no move to accept or reject that precedes it. This is the case even if the scene resumes a previously-established common ground, for the resumption of a common ground requires at least one signal to come into effect. That is, without at least one signal, it is unclear which of the previous partial grounds is being utilised. And therefore, even if the line responds to an offer in a previous scene, it necessarily establishes the scene and introduces new information, and it is therefore an offer. Naturally, the probability that A’s move is a proposal is at least extremely high, if not certain. The second factor is important because the pause may have been misconstrued as indicating the end of the utterance. A plausible scenario is that A’s utterance was meant to continue, and that what he manages to produce is just a preamble that sets the scene, and justifies his further conspiracy theories about the zoo. But because B correctly infers that the function of A’s utterance is a proposal, and since he misinterprets the lengthy pause in A’s line, he reasons that the utterance is an invitation to provide some justification for A’s character’s apparent disquiet. In this way, this heuristic may also account for these kinds of errors.

7.6.3 · A return to verisimilitude

In §5.5.1, I introduced heuristic 1, which describes the properties of a strategy that is optimal in its verisimilitude with respect to the plot-world. The heuristic is reproduced as heuristic 12.

Heuristic 12: Verisimilitude. A strategy is called optimally verisimilar if it appears to believably belong to the plot-world; that is, if it does not appear to be produced by an actor pretending to be a character.

When I introduced this heuristic in §5.5.1, I noted that, although this is a fundamental feature of the genre of communication, we lacked a mathematical vocabulary that would allow us to describe this phenomenon in formal terms. Now, finally, with the introduction and inclusion of steganographic concealment in the game theoretic model of communication, we are able to describe actors’ strategies that seek to preserve the verisimilitude of the plot-world while subliminally planning the plot. The problem of verisimilitude is almost identical to the problem in other forms of steganography of emitting secure messages. In the same way that emissions generated by a stegosystem should conceal information in such a way that any eavesdroppers are unable to detect that information, the audience should ideally be unable to detect any communication that makes reference to extra-narrative elements of the performance. The ability of a stegosystem to felicitously achieve this goal is the security of that system. And it is security that I propose is best able to model strategies that are optimal according to this heuristic.

The security requirements of improvised theatre differ somewhat from general stegosystems. As I have mentioned, the genre of speech in improvised theatre imposes unique
constraints on what is available to actors both as stegoforms and as coverforms. In other forms of steganography, the coverforms need only make sense only insofar as any randomness might give an Adversary a clue that there is some concealed information. The security of a natural language stegosystem relies on a similar principle. An utterance that is selected according to a secure natural language stegosystem is one that is suitable to the plot-world at the time that it is produced. Utterances which appear at best tangential, and at worst inconsistent with the plot-world, poorly conceal the actors’ planning communication. According to Cachin (1998, 2004), the security of a stegosystem is a measure of the entropy discrimination of its stegoform and coverform distributions. The intuition that underlies this idea is that the degree to which the Adversary is unsure of which distribution an emission was generated by, or was selected from, provides a metric of how likely successful concealment will be.

This applies equally to improvised theatre. There is, however, a major difference: messages that are generated according to the coverform distribution are utterances that are suitable to the plot-world, while those that are generated according to the stegoform distribution are not, and are instead suitable to conversations that engage metanarrative elements of the performance. This conforms, I believe, to the facts thus far observed about improvised theatre. Actors produce utterances that either build or augment the internal logic of the plot-world, and they conceal signals about their preferences within these utterances. The security of the natural language stegosystem that actors employ is measured, then, by the degree to which their metanarrative communication resembles communication that is suitable to the plot. Observe that the notion of plausible deniability is central to this steganographic understanding of actors’ communication. If she has designed and produced her utterances correctly, an actor should be able to deny that any covert communication has occurred at all.

A strategy is optimally verisimilar, then, if it is the product of a natural language stegosystem with perfect security. In terms of improvised theatre, this represents the notion that a strategy’s verisimilitude derives from the degree to which an observer, whether the Receiver or the Adversary, is uncertain of which distribution an utterance is generated by. The formal definition follows.

**Definition 7.24:** Optimal verisimilitude (from heuristic 1). Let $\Xi$ be a natural language stegosystem, and fix coverform and stegoform distributions $\delta^c_{\mu|W}$ and $\delta^s_{\mu|W}$ over $\mu$, respectively. A Sender strategy $m \in \mu$ is optimally versimilar iff $D(\delta^c_{\mu|W}||\delta^s_{\mu|W}) = 0$.

There is a nuanced effect of actors’ preferences that underpins the operation of this heuristic. As we have seen repeatedly, actors prefer strategies that do not reveal the mechanics of the performance. And so in effect the heuristic of verisimilitude binds the security of a natural language stegosystem in improvised theatre to how suitable an utterance is relative to the plot-world. While it may appear appropriate to define this formally, there is no need to do so, since the definition of relative entropy satisfies this requirement. Recall the final statement of definition 7.6, that the relative entropy of a distribution with rela-
tion to another is asymmetric, so that if $\delta_X \neq \delta'_X$, $D(\delta_X \parallel \delta'_X) \neq D(\delta'_X \parallel \delta_X)$. By specifying that verisimilitude relies on the relative entropy of $\delta^\pi_{\mu|\mathcal{W}}$ with respect to $\delta^c_{\mu|\mathcal{W}}$, the actors’ proclivity towards preserving the verisimilitude of the performance is captured.

7.7 · Conclusion

In this lengthy chapter, I have proposed that the covert contents of actors’ communication, which is problematic for standard game theoretic models of language, may be modelled by drawing on information theoretic formalisms from the study of steganography. This descriptive of natural language steganography that I have devised has attempted to remain as general as possible. To do so, a number of choices have been made. First, the model is based on information theory, and particularly on that theory’s model of communication. The information theoretic model of communication was not formulated for natural language interactions, but is a general theory of how information is transferred between two entities. The model of steganography was based on this model of communication, which allowed for the definition of natural language utterances as the messages that are sent in the system. The compatibility of the theories allowed for the incorporation of a game theoretic model of communication and an information theoretic model of steganography. The resulting model was applied to some communicative situations in improvised theatre, and the concealment of information was shown to be present in declarative utterances, questions, and utterances that contain some implicated meaning. Finally, I defined heuristics that pertain to natural language steganography as it appears in improvised theatre, as well as returning to a previous heuristic regarding verisimilitude that was not formally defined.

With the introduction of the formal model of natural language steganography, the formal model of improvised theatre has reached its culmination. The model, which is the sum total of the definitions and heuristics that have been presented throughout the chapters, can provide a formal, game theoretic account of the use of language—including the transmission of concealed information—in any longform improvised performance. It thereby allows for standard accounts of the linguistic aspect of improvised performance, and permits rigorous explorations of actors’ cognition, particularly under differing conditions. In the next chapter, I demonstrate how a complete description of a scene might be conducted.
Chapter 8

An analysis of a scene

In this chapter, I provide an example analysis of a scene from the data. In §8.1, I offer a brief but complete description of the model, which links its parts together. In §8.2, I present the scene that I analyse in full. The state transitions of the scene are listed in §8.3, and a description of the first four offer–response (o–r) pairs as periods (that is, rounds, or iterations) of an iterated game follows in §8.4. I conclude the chapter in §8.5.

8.1 · To recapitulate: an outline of the model

We begin with a set of plot-worlds. Each plot-world is a possible world. A plot occurs within a plot-world, and is a vector of events such that the probability of each element of the vector is nonzero. Events are visible actions. This includes most prominently utterances and gestures, and does not include the selection of interpretations. The plot is partitioned into scenes, which are generally restricted to a single location, and to a finite, established ensemble of characters. Each scene is partitioned further into states of the scene (or, simply, states). States are identified by their extensions; a given state is identified with the plot-worlds at which it is true. States may shift, but only by the deliberate actions of the actors. The specific mechanism for this is the iterative performance of offers and responses. Offers may signal that the performer of the offer wishes to shift the state (SSM), or to maintain the state (SMM). Responses are categorised according to two axes—polarity relative to the assertion made by the utterance (affirmative or negative), and conformance to the intended state operation (augmentation or redirection)—so that the possible combinations are affirmative augmentation (AA), affirmative redirection (AR), negative augmentation (NA), and negative redirection (NR). AR responses are impossible after SSM offers, and NA responses cannot follow SMM offers.

A character is identified by a set of traits, and is therefore formally represented as an attribute–value matrix. A character’s traits are ascribed by the deliberate actions of the actors in the scene. Intrinsic traits of the actors (accent, physical appearance, gender, and so on) are not automatically inherited by their characters. A character is consistent if there is an existent in the plot-world that felicitously matches that character for the entirety of a scene. Inconsistencies most often occur when a trait has been assigned, and then another
trait is assigned that is inconsistent with the first. A simple but common example involves applying two different names to a single character.

An improvised performance is modelled as an iterated improvised performance game. Improvised performance games (IPGs) are essentially interpretation games, but with two key differences. (1) In interpretation games, there are two players: Sender and Receiver. Meanwhile, in iterated improvised performance games, these players are recast as roles. (2) Players may choose to remain silent by selecting the empty utterance $\varepsilon$. Each period of an iterated IPG has the following order.

1. Nature probabilistically selects a state of the scene. Only the Sender observes this selection.
2. The Sender produces an utterance, which the Receiver observes.
3. The Receiver selects an interpretation.

Iterated IPGs also have a period at the very beginning of the game, period 0, in which Nature selects a plot-world, rather than a state of the scene. Only the player who is the Sender in period 1 views the selection of a plot-world.

The first actor to utter a line in a scene sets the scene by establishing it in a certain location—which may be heavily underspecified—, and by introducing certain characters. This actor effectively determines which plot-world the plot will be set in. This determination is made at the discretion of this actor. Nature’s selection in period 0 is the formal implementation of this, for it is only revealed to the first Sender of the iterated game. Nature’s selection of a scene-state in each subsequent period represents a similar process. During a performance, an actor’s preferences over plots change. As the selection of a plot-world models the first Sender’s desired plot-world, the selection of a scene-state models the desired state of the Sender of that period. Once more, Nature’s selection is only revealed to the Sender, and so it can be used to model the Sender’s mental state.

An actor’s strategy at a particular period of the iterated game depends on the strategies that have been played prior to that period. Modelling an improvised performance as an iterated game allows for a formal account of the long-term effects of strategies throughout the scene.

The utility of the performance is revealed to the players only at the conclusion of the iterated game rather than at individual periods. The actors are able to estimate the utility they receive at the end of each period, however, because they know their private histories. Each player’s private history at period $t$ includes the utterance that is produced at each period preceding $t$. But there are differences in the histories of the players in a given period, which depend on their role as Sender or Receiver. The Sender’s history of a single given period includes the state of the scene that was selected, but since she is uncertain about which interpretation the Receiver has selected, the history includes the interpretation that she considers most probable. Likewise, the Receiver’s history of a single given period includes the interpretation that she has selected, and an estimation of the state of the scene.
that is selected by Nature. On the basis of a player’s history at period \( t \), that player can estimate the utility that she will receive given the expected utility of each period until \( t \). The players are aware that they lack certain knowledge regarding their fellows’ interpretations, and so they are aware that their estimate may be defective.

To describe the natural language used by actors in improvised theatre, I define (a) a DRS language, and (b) an intensional model for that language, whose set of worlds is the set of plot-worlds. The DRS language formally expresses the meaning of natural language sentences. Each strategy that is performed by the actors is captured by the DRS, including those that are not traditionally included in standard DRSs, such as gestures and mime, which are used to introduce entities into the plot-world. The common ground of the performance is defined as a DRS that acts as a record of the information that has been introduced and accepted by the actors. As such, the common ground is a partial model of the plot-world. As the scene proceeds, and more information is accepted into the common ground, the partial model becomes more specific, and the number of worlds within which the partial model may be embedded diminishes. In other words, the common ground represents what the audience knows about the plot-world.

Actors share information that is not available to the audience. This primarily includes the information that any communicative act in improvised theatre should be understood as in part proposing that such information be included in the common ground, or accepting or rejecting that information. Actors almost always attempt to conceal this kind of information from the audience, for it interferes with the verisimilitude of the performance which appears to be dispreferred in the vast majority of cases.

Concealment is modelled as a natural language stegosystem, which exploits differences in the information that is shared by the actors and the audience. Utterances designed according to the principles of the natural language stegosystem seek to be secure, which means that the actors attempt to design utterances so that their concealed contents (a) are unlikely to be detected by the audience, while (b) are clear enough that the other actors can detect the hidden message. Utilities for the actors are calculated according to two criteria: (1) whether the actors successfully communicate amongst themselves; and (2) whether the audience is provided with sufficient information to reach an interpretation that is plausible in the plot-world. The utility for the entire performance is the sum of each period’s utility, and so a successful performance is one in which the mechanism of the performance is hidden from the audience.

8.2 · The scene

The scene that I analyse is 1-A1, which is presented in its entirety in datum 22 overleaf. Its opening lines have been referred to a number of times throughout the thesis, most importantly in §5.2.1.1.

The analysis that I offer provides an example of the kind of analysis that may be achieved with the model that has been formulated so far. It demonstrates how the use of multiple formal frameworks—in this case, game theory, Discourse Representation Theory,
8.2 · The scene

Datum 22 Scene 1-A1.

1 A well they **fired** me
2 (1.1)
3 P xxx
4 B: why'd you show up to *work* then dad
5 P [x ]
6 A [i've] got (.) i've got nowhere else to go do i↓ ((mimes unpacking an unidentified object)) (4.2) been here for (1.5)
7 forty years (1.8) this is all i know
8 (1.4)
9 B well you've pulled out the old typewriter again
10 A HHhh
11 P xxxxxxxxx
12 A ((begins to work on the typewriter)) a job doing (1.0) a job
13 B ←s turning on my p.c.
14 P xxxxxxxxx
15 A mine's already warmed "up"
16 P xxxxxxxxx
17 B dad i (2.7) dad i'm really worried about you (1.3) you know↑
18 (2.0) i'm gonna- look i see you as a bit'v a fath're figure
19 (0.7)
20 P xxxxxxxxx
21 B you know (.) in- in kind of an abstract sort of way (3.3) *and*
22 (1.2) and the fact that (1.2) that- i mean this is desp'ratly sad (1.9) you'd (1.9) maybe you should (2.9) you've got nowhere else to go:↑?
23 (1.6)
24 A i've got nowhere else i wanna be
25 (2.3)
26 B than next to me?
27 (1.5)
28 A than at *work*
29 P xxxxxxxxx
30 (0.8)
31 B "(the) computer's started up" =
32 A =just because i'm old it doesn't mean i can't contribute (1.7) i've been typing out these letters for the last forty years (1.0) not one of them's been lost not one of them's had a spelling error
33 (1.8)
34 B i just think you that- i mean if you're gonna use the ticketing software (1.4) you'll end up in a better position i (1.0) this could be- (0.9) is it why? they fired- why? are you?
A why do you wanna replace things? what IS it with you guys? what IS it with your generation why are you tryin 'a replace things?

(0.8)

B sorry i just got a: message

A d'you think every time you came home and disappointed me in some way i thought fuck it let's have another kid?

P XXXXXXXXXX[xxx ]

A [it's not] how it works, james. you try and fix what you've got to make it the best that it can be you don't replace it with a model that takes ten minutes to warm up that's got no [soul to it]

B [to- to ] boot (. ten

A [i don't know]

B warm up's what you do with engines or what you used to have to do with engines they're pretty good now

P [XX ]

B [you use those] little key clickers as you approach the car you don't even the- to- it's contactless start↑

(0.3)

A i('d) get up a half hour earlier to turn on the car

P XXXXXX

A i make myself a coffee i go back down and i check that the car hasn't overheated

P xx

A and then i turn it off for a bit

P XXX

A cause the old girl can only run for a certain amount of time before she'll conk out

P XX

((another actor terminates the scene))
and information theoretic steganography—can be combined to form a rigorous descriptive model of interlocutor behaviour. The analysis shows how these interoperate, and how the model describes actor behaviour. It is split into two sections. First, I describe each of the scene-state transitions, in §8.3. And second, I focus on a segment of the scene, in §8.4. The analysis is split because the model operates simultaneously on a number of levels. The list of scene-state transitions describes the course of the plot throughout the scene at a general level, while an analysis of a particular segment of the scene models the specific utterances that are used.

8.3 · Describing the plot through state transitions

In this subsection, I list the transitions of the state of the scene that occur throughout datum 22. In the following list, three pieces of information are given in the item’s title: the transition that occurs; the turn that the offer is made in, and the type of offer; and the turn that the response is made in, and the type of response.

$q_1 \rightarrow q_2$. Offer: 1, SSM. Response: 4, AA.
At turn 1, the common ground $\Gamma^1$ is empty, and so $\llbracket \Gamma^1 \rrbracket = \emptyset$. The offer, then, is an SSM, because A’s utterance requires that there be respective entities that satisfy the pronouns me and they. The state is shifted because the extension shifts from $\emptyset$ to a set of worlds for which A’s utterance holds. I claim that B’s utterance is an affirmative augmentation.

$q_2 \rightarrow q_3$. Offer: 4, SSM. Response: 6, AA.
B’s response to A’s offer in line 1 operates simultaneously as an offer. The reason it is an offer is that, should the DRS of the utterance be incorporated into the common ground, the extension of the common ground will be altered so that only those worlds in which two facts hold will be included in the set of plot-worlds that satisfy the model. These two facts are that (a) the scene is set in the workplace of both of the characters, and (b) B is A’s son. A’s response in line 6 is an affirmative augmentation, since it provides a reason for the unexpected fact of his character being at work, given that he has been fired.

$q_3 \rightarrow q_4$. Offer: 6, SSM. Response: 8, AA.
The response to the previous offer also acts as an offer in its own right. By miming that there is an object on a surface in front of him, A asserts the existence of that object. A also asserts that he has been at the workplace for forty years. As such, the offer shifts the state. Note that the object introduced by A’s offer remains unidentified until the response is made. Since the object is identified by B, his response is categorised as an AA.

$q_4 \rightarrow q_5$. Offer: 8, SSM. Response: 11, AA.
The offer posits that A uses a typewriter. This assertion is subsequently accepted by A’s utterance, but also by his miming the use of a typewriter. Since this fact is incorporated into the plot-world, the response is an affirmative augmentation.

$q_5 \rightarrow q_6$. Offer: 12, SSM. Response: 14, AA.
The offer posits that B sits in front of a computer, and that the computer is booting. The response accepts that the computer exists, and further provides a point of difference between the two devices that becomes important later in the scene. This is not counted as an offer,
however, because it does not introduce new information, and because the primary purpose of the utterance is to accept the information in B’s utterance.

$q_6 \rightarrow q_7$. Offer: 17 & 20, SSM. Response: 22, AA.  
B’s utterance introduces some facts about his character’s mental state, and particularly about his feelings towards his father. Note that this line does not overturn the assertion that A is B’s father. The mention of “father figure … in an abstract way” are meant ironically; this is indicated by B’s intonation, which is represented in datum 22 by the frequent pauses and interruptions in B’s delivery of the line, and by B’s primary stress on the word abstract. The most salient part of the offer is the final sentence of line 20. This provides A with something that he can directly respond to. The response is an affirmative augmentation. It is affirmative simply because it answers this question directly in the affirmative, and it is augmentative because it shifts the state.

$q_7 \rightarrow q_8$. Offer: 24, SSM. Response: 26, NR.  
The offer made by B responds to the information that was introduced by A in his response in the previous pair. Here, we arrive at our first negative response, and our first redirection. A’s response rejects the assertion of B’s utterance, namely that the only place where A wants to be is by his son’s side. It also redirects the state to one that is not intended by B, simply because the change to the extension that was posited would be such that only the worlds consistent with A’s wanting to be with his son would be included in the extension of the common ground. Instead, a common ground whose extension contains only those worlds that are consistent with A’s desire to be at work is made operative by A, and so this response is a redirection.

This is a fringe case where the offer does not have a very salient response. The offer posits that B’s computer has completed booting. While A’s utterance in line 30 might be construed as a response, it is not entirely clear whether A meant to respond to B’s utterance, for A interrupts B’s line, which is indicated by the equals signs that join the two lines 29 and 30. Another factor that contributes to this characterisation of the situation is that B’s line in 29 is uttered more quietly than the normal volume of the conversation. This o→r pair, then, is an instance of an interrupted pair, where no response is provided to an offer. By heuristic 9, any information is by default accepted into the common ground, and may only be rejected if it is explicitly challenged by one of the actors. Therefore, because no objection is raised by A, that the computer has completed starting up enters the common ground, and so the state is shifted.

$q_9 \rightarrow q_{10}$. Offer: 30, SSM. Response: 32, AA.  
The offer that initiates this change of state is unlike the other offers that have been examined thus far in the scene. While the other offers comment on events that are contemporary with the scene, this offer constrains the worlds that are satisfied by the common ground by providing a history of events in the plot-world that precedes the scene. The effect on the extension of the common ground is precisely the same as the effects of other utterances that we have observed so far in this chapter: a plot-world’s history is constituted of facts, and only possible worlds that are consistent with these facts remain viable candidates for the

---

1Heuristic 9 states that “A response strategy is by default positive.” See §6.3.3.
A segment of the plot as an iterated game

In this section, I focus on a specific segment of the plot, and describe it as an iterated game. My intention is that this will demonstrate the efficacy of the model that I have sketched in this thesis. It is important to note that the analysis that I perform in this section is chiefly concerned with the strategic use of utterances. That is, I am primarily interested in utterances as strategies in an iterated game. As such, I do not necessarily provide a thoroughly rigorous linguistic analysis of these utterances besides their strategic
function. Since the linguistic devices used by actors in improvised theatre were the focus of chapters 5, 6, and 7, I turn now to a more abstract, game theoretic description of utterances.

Many of the phenomena that are described in this section will be restatements of phenomena that have previously been observed in chapters 5, 6, and 7. All o–r pairs tend to have very similar descriptions, and differ only in the particular devices that are employed, and in the heuristics that we might employ to describe them. For this reason, this section is quite brief, and is meant to give an idea of the mechanism of formalising o–r pairs, and thus of how the model can be used in the description of actual, naturally-occurring data.

Since it has been used throughout the thesis as an exemplar, I use the beginning of 1-A1 as presented in datum 23, as an example of how a portion of a plot might be examined as an iterated game.


<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>(1.1)</td>
</tr>
<tr>
<td>3</td>
<td>P</td>
</tr>
<tr>
<td>4</td>
<td>B</td>
</tr>
<tr>
<td>5</td>
<td>P</td>
</tr>
<tr>
<td>6</td>
<td>A</td>
</tr>
<tr>
<td>7</td>
<td>(1.7)</td>
</tr>
<tr>
<td>8</td>
<td>B</td>
</tr>
<tr>
<td>9</td>
<td>A</td>
</tr>
<tr>
<td>10</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>A</td>
</tr>
</tbody>
</table>

8.4.1 · The first offer–response pair

The first o–r pair is constituted of lines 1 and 4. Note that an o–r pair is made up of two periods of an iterated improvised performance game, because A produces an utterance—the offer—that is interpreted by B and the audience P, and then B produces an utterance—the response—and this is subsequently interpreted by A and P.

8.4.1.1 · \( t = 0 \)

The scene begins with Nature selecting a plot-world \( w \in \{W\} \).

8.4.1.2 · \( t = 1 \)

The first period of the game involves the utterance of the first line, and its subsequent interpretation by B and P. A observes \( w \), and produces an utterance \( m \in \mu \), which is given in \((105)\).

\((105)\)  A: well they fired me

The intended meanings of this appear to be fairly simple. \((105)\) ostensibly states that an entity fired A. The referent that refers to the entity is unbound, and yet this is felicitous in
the genre of improvised theatre, as was discussed in §5.2.1. The concealed meaning of (105) is a simple proposal that the ostensible meaning holds in \( w \). The stack notation follows.

\[
K_1 = \begin{align*}
[A, B, P] ; x, A, B \\
\text{fired}(x, A)
\end{align*}
\]

\[
K_2 = \begin{align*}
[A, B] ; A \\
\text{proposes}(A, K_1)
\end{align*}
\]

Note that, since \( B \) is on stage, and since \( A \) addresses \( B \) directly, \( B \) is included in the utterance’s DRS, for the utterance asserts that \( B \) exists, and is part of the scene.

At the end of the period, the utilities are awarded. In §7.4, I proposed a utility function (reprinted as equation 8.1) that takes into account the audience’s willingness to suspend disbelief, and its aversion to detecting concealed metanarrative information.

\[
u(s) = \begin{cases} 
(1, 1, 1) & \text{if } s = (w, m[w], K_R[m(w)], K_P[m(w')]) \\
(0, 0, -1) & \text{if } s = (w, m[w], K_R[m(w)], K_P[m(w)]) \\
(-1, -1, 1) & \text{if } s = (w, m[w], K_R[m(w')], K_P[m(w')]) \\
(-1, -1, -1) & \text{if } s = (w, m[w], K_R[m(w')], K_P[m(w)])
\end{cases}
\]

Suppose that there are two salient interpretations: \( K_1 \), which is the ostensible meaning only, and \( \bar{K}_1 \), which is the full stegoform. Let us also make the simplifying assumption that there are two salient states that are mutually exclusive: \( q \), which is the true state of the plot-world, such that \( \models_{\text{plot}} q \), and \( q' \), some alternative such that \( \models_{\text{plot}} \neg K_1 \). The strategy that \( A \) employs in this instance appears to be successful, but a true determination of its success is not made until the following period. The utterance communicates his proposal for \( K_1 \) well to \( B \), and it is entirely plausible that it is an utterance that is consistent with the plot-world, and as such is verisimilar in the plot-world. Let \( w \) be a plot-world such that \( \models_{\text{plot}} w \). In this case, the strategy profile that represents the first line of the performance is \((q, m(q), \bar{K}_1(m[q]), K_1(m[q']))\). That is, a state \( q \) is selected by Nature; \( A \)’s utterance \( m(q) \) responds to \( q \); \( B \) appears to select \( \bar{K}_1 \), which is the concealed information; and \( P \) appears to select \( K_1 \), which is the ostensible information. By equation 8.1, the utility profile for this period is \((1, 1, 1)\).

8.4.1.3 \( t = 2 \)

\( B \)’s response is the second period of the game, which is similar to the first period, but for two chief differences. First, the state differs from period 1 so that it represents \( B \)’s preferences over the plot. Second, the players’ choice of strategy depends on the history of \( B \)’s discourse, which is \( \eta^B_2 = \langle \max \delta_B(q), m^1, K_B \rangle \) at the beginning of the second period. Recall that \( A \) is the only player who knows the state of the world. This is true in improvised theatre, for the first actor to speak in a scene selects and sets the initial conditions of a plot,
as was discussed in §5.2. As such, the state is only part of his history. The other players’ histories contain an estimate of the state given their beliefs. Furthermore, A is unaware of which interpretation the other actor has selected, and so it is not part of his history. Each of B and P know which interpretations they have selected, but are unaware of the state. The only common element to all histories is the utterance, which is the primary condition for B’s and P’s inferences of the state of the world. This captures the co-constructive nature of improvised theatre: utterances provide evidence of the mental states and preferences of their speakers to their hearers.

Since B has selected the entire stegoform $\tilde{K}_1$, he knows that the previous move was an offer, and he therefore knows that he must respond in such a way that either accepts or rejects $K_1$, or accepts some part thereof. Moreover, $[[K_1]] \neq [[\Gamma^0]]$, and so B infers that $m$ is an SSM. It is also important, according to heuristic 2, that the DRS of B’s utterance be consistent with the partial model of the plot-world, that is $K_1$. His utterance, $m^2$, is selected with these facts, and with heuristic 11, in mind. In this regard, $m^2$-as-(107) is an extremely effective strategy.

(107) B: why’d you show up to ↑work then dad

The DRS of $m^2$ is represented in (108). Recall that the type of specification given in the DRS of (108) is not comprehensive, but instead includes only those aspects of the utterance that are most salient to its strategic function.

\[
\begin{align*}
K_3 &= [A, B, P]; y, A, B \\
\text{workplace}(y) \\
\text{be_at}(A, B, y) \\
\text{father_of}(A, B) \\
K_4 &= [A, B]; B \\
\text{accepts}(B, K_1) \\
\text{proposes}(B, K_3)
\end{align*}
\]

This utterance appears highly functional in the sense articulated in heuristic 11. While B’s utterance does not directly address the assertion made by A, recall that responses by default accept what is proposed in the preceding offer. As such, the utterance is not clearly polar, but its metanarrative intent is very clear. By heuristic 12, the strategy is also verisimilar, since it is entirely consistent with the plot-world, and as such appears to have been drawn

---

2Heuristic 2 states that “A strategy is optimally compatible if the facts expressed by that strategy do not contradict what is established about the plot-world prior to that strategy’s deployment.” See §5.5.1.

3Heuristic 11 states that “An embedded meaning is optimally functional if the probability of one of the predicates proposes, accepts, or rejects being the embedded meaning, conditional on the actors’ partial ground, is greater than the respective probabilities of the other two.” See §7.6.2.

4Heuristic 12 states that “An action is called optimally verisimilar if it appears to believably belong to the plot-world; that is, if it does not appear to be produced by an actor pretending to be a character.” See §7.6.3.
from the coverform distribution, rather than the stegoform distribution. That is, it appears to be an utterance that might plausibly occur in the plot-world, rather than an utterance that might be produced by a pair of actors who are planning an improvised performance.

Since $K_4$ suggests that $B$ accepts $K_1$, the information in $K_1$ is included in the total common ground of the scene; that is, the common ground that is available to all actors and the audience. Recall that the common ground at a particular period $t$ is a register of the information from periods $1, \ldots, t-1$. Therefore, $\Gamma^3$ includes information from $K_1$, since it is accepted into the common ground in the second period.

$$\Gamma^3 = [A, B, P]; x, A, B$$

The common ground includes information that $A$, $B$, and $x$ exist, and that $x$ has fired $A$. So, the common ground behaves as a register of the information that has been established about the plot-world.

The partial ground that $A$ and $B$ share, $y(A, B)$, includes the proposal and acceptance of the DRSs that are available to the audience. At $t = 3$, then, $y^3(A, B)$ is the merge of (106) and (108), plus information that derives from the a priori common ground. Recall that, by heuristic 10, actors always possess information that states that they and their fellows are actors, and therefore that concealed information is always included. I summarised this using the predicate actor($x$). I include this in $y^3(A, B)$.

$\Gamma^3$ suggests that $B$ accepts $K_1$, the information in $K_1$ is included in the total common ground of the scene; that is, the common ground that is available to all actors and the audience. Recall that the common ground at a particular period $t$ is a register of the information from periods $1, \ldots, t-1$. Therefore, $\Gamma^3$ includes information from $K_1$, since it is accepted into the common ground in the second period.

$$\Gamma^3 = [A, B, P]; x, A, B$$

The common ground includes information that $A$, $B$, and $x$ exist, and that $x$ has fired $A$. So, the common ground behaves as a register of the information that has been established about the plot-world.

The partial ground that $A$ and $B$ share, $y(A, B)$, includes the proposal and acceptance of the DRSs that are available to the audience. At $t = 3$, then, $y^3(A, B)$ is the merge of (106) and (108), plus information that derives from the a priori common ground. Recall that, by heuristic 10, actors always possess information that states that they and their fellows are actors, and therefore that concealed information is always included. I summarised this using the predicate actor($x$). I include this in $y^3(A, B)$.

$$\Gamma^3 = [A, B, P]; x, A, B$$

The common ground includes information that $A$, $B$, and $x$ exist, and that $x$ has fired $A$. So, the common ground behaves as a register of the information that has been established about the plot-world.

The partial ground that $A$ and $B$ share, $y(A, B)$, includes the proposal and acceptance of the DRSs that are available to the audience. At $t = 3$, then, $y^3(A, B)$ is the merge of (106) and (108), plus information that derives from the a priori common ground. Recall that, by heuristic 10, actors always possess information that states that they and their fellows are actors, and therefore that concealed information is always included. I summarised this using the predicate actor($x$). I include this in $y^3(A, B)$.

$$\Gamma^3 = [A, B, P]; x, A, B$$

The common ground includes information that $A$, $B$, and $x$ exist, and that $x$ has fired $A$. So, the common ground behaves as a register of the information that has been established about the plot-world.

The partial ground that $A$ and $B$ share, $y(A, B)$, includes the proposal and acceptance of the DRSs that are available to the audience. At $t = 3$, then, $y^3(A, B)$ is the merge of (106) and (108), plus information that derives from the a priori common ground. Recall that, by heuristic 10, actors always possess information that states that they and their fellows are actors, and therefore that concealed information is always included. I summarised this using the predicate actor($x$). I include this in $y^3(A, B)$.
Note that the partial ground behaves differently from the common ground. The common ground requires the explicit acceptance of information for that information to be included. Meanwhile, the partial ground only requires that the information be presented. This, I argue, is a consequence of heuristic 10. Actors expect that their fellows will conceal information that is related to the ostensible meaning of the utterance, and they furthermore anticipate that this information is necessary for a felicitous interpretation of the utterance. By quantifying the partial ground that the actors share, and the common ground that is available to the audience, the extent of the information that the actors conceal from the audience is visible.

At the period’s conclusion, the provisional expected utilities are revealed to the players for this period. As with the previous period, it appears that the metanarrative information concealed in the utterances is hidden sufficiently well. The common ground at the conclusion of the second period, \( \Gamma^3 \), contains enough information for the audience to select an interpretation of B’s response. Furthermore, B’s utterance signals his acceptance of A’s concealed offer. The requirements for successful natural language steganography have been met, and so the provisional expected utility profile for each player is \( (1, 1, 1) \).

8.4.2 · The second offer–response pair

Lines 4 and 6 of the datum make up the second o–r pair.

8.4.2.1 · \( t = 2 \) (again)

The response of the first o–r pair, line 4 of the datum, acts simultaneously as the offer for the second. The DRS of \( m^2 \) contains acceptance of \( K_1 \), and makes a proposal of \( K_3 \). As an offer, it also appears to be a successful strategy, since what is being proposed is clear to the actors, but the utterance is sufficiently verisimilar. I discussed in §7.4.2 that questions are often used as offers for the target to provide further information about the plot-world, and this is an instance of its effective use.

8.4.2.2 · \( t = 3 \)

Line 6 of the datum, reproduced as (111), is the response of the second o–r pair.

\[
A \quad [i’ve \, got \, .] \, i’ve \, got \, nowhere \, else \, to \, go \, do \, i↓ \, ((mimes \, unpacking \, an \, unidentified \, object)) \quad (4.2) \, been \, here \, for \quad (1.5) \, forty \, years \quad (1.8) \, this \, is \, all \, i \, know
\]

This response is selected under similar epistemic conditions to those of the response in the previous o–r pair. However, note that A’s decision is complicated by a private history that contains more information than it did in the previous o–r pair. A’s private history at the beginning of this period is \( \eta^A_2 = \langle q, m^1, \max \rho_A \rangle \). The final element of the history is the interpretation of his own utterance at line 6 that A infers is most likely given his Sender belief.

Upon observing \( m^2 \), A has evidence that B accepts the initial state. The inclusion of both \( m^1 \) and \( m^2 \) in \( \eta^A_2 \) allows for the observation of both utterances to affect the strategy that A selects, as it does in this instance. The felicitous selection of the concealed DRS
8.4 · A segment of the plot as an iterated game

\( K_4 \), such that \( K_4 \rightarrow K_3 \), allows \( A \) to select which of the facts should be included in the common ground.

The stack notation of \( A \)'s response to (111) is in (112). Given that I am interested here in the strategic function of this utterance, I identify the most important information in this line as (a) to propose a reason for \( A \) being at work; and (b) to assert the existence of an object, which is presumably used by \( A \) at work. I do not therefore give a full specification of the DRS of the utterance. Note, too, that the DRS language that was defined in chapter 4 did not include any modal operators, in particular possibility and necessity. Therefore, what I provide is an approximation of the utterance meaning.

\[
(112) \quad \tilde{K}_3 = \begin{array}{|c|}
\hline
\text{object}(x), \text{workplace}(y), \text{unpacks}(A, x), \text{be_at}(A, B, y), \text{employed_for}(A, \text{forty years}) \\
\hline
\text{can be_at}(\neg y) \\
\hline
\end{array}
\]

\[
K_5 = \begin{array}{|c|}
\hline
[A, B, P]: x, y, A, B \\
\hline
\text{object}(x), \text{workplace}(y), \text{unpacks}(A, x), \text{be_at}(A, B, y), \text{employed_for}(A, \text{forty years}) \\
\hline
\text{can be_at}(\neg y) \\
\hline
\end{array}
\]

\[
K_6 = \begin{array}{|c|}
\hline
[A, B]: A \\
\hline
\text{accepts}(A, K_3) \text{ proposes}(A, K_3) \\
\hline
\end{array}
\]

While the response is a simple affirmative augmentation, it also acts as an offer, though a subtle one. It behaves as an offer because it introduces the existence of an object in front of \( A \). I discuss this further below. But it also adds further information about the character’s motivations, which are a part of his traits. It does so by attributing a desire to the character; namely a desire to be at his workplace despite his being fired. This becomes the major conflict at the centre of the scene’s development, and it is a premise that is referenced throughout the rest of the scene.

With \( A \)'s acceptance of \( K_3 \), \( K_3 \) merges with \( \Gamma^3 \) to become \( \Gamma^4 \).

\[
(113) \quad \Gamma^4 = \Gamma^3 \uplus K_3 = \begin{array}{|c|}
\hline
x, A, B \\
\hline
\text{fired}(x, A), \text{workplace}(y), \text{be_at}(A, B, y), \text{father_of}(A, B) \\
\hline
\end{array} = \begin{array}{|c|}
\hline
[A, B, P]: x, y, A, B \\
\hline
\text{fired}(x, A), \text{workplace}(y), \text{be_at}(A, B, y), \text{father_of}(A, B) \\
\hline
\end{array}
\]

(112) and \( K_3 \) are merged with \( \gamma^3(A, B) \) to make \( \gamma^4(A, B) \). I do not present a full specification of this, for it would simply repeat what has already been stated, and the representations would become too large.
8.4.3 · The third offer–response pair

The third o–r pair, constituted of lines 6 and 8, is similar to the second. Its offer is in fact the response of the previous o–r pair. It differs however in the kind of information that makes the utterance an offer.

8.4.3.1 · \( t = 3 \) (again)

Once again, the response to the previous offer turns out to be an offer in a new o–r pair, but in a radically different way. He appears to be unpacking an object that rests on a surface in front of him, an action that predicates the object’s existence. Here we see the benefit of including gestures and mime in the common ground of the performance. Since the o–r pair introduces a new entity, this offer is an SSM, even though the entity is unidentified. Merely asserting that there is an object in the plot-world is introducing a new entity.

8.4.3.2 · \( t = 4 \)

The response of the third o–r pair, denoted \( m^4 \), is line 8 of the datum, reproduced as (114).

(114)  B: well you’ve pulled out the old typewriter again

The epistemic conditions under which the utterance is made follows the pattern that has been established. What is interesting about this response is that it does not posit the existence of the typewriter, since the existence of an object that sits in front of A has already been posited, but it instead assigns an identity to that object. As such, the stegoform representation is as follows.\(^6\)

\[
K_7 = \begin{bmatrix}
[A, B, P]; A, x \\
typewriter(x) \\
unpacks(A, x)
\end{bmatrix}
\]

\[
K_8 = \begin{bmatrix}
[A, B]; B \\
accepts(B, K_5) \\
proposes(B, K_7)
\end{bmatrix}
\]

The concealed acceptance in (115) updates the total common ground once again, so that it becomes \( \Gamma^5 \), represented in (116). Recall that, when DRSs are merged, the variables that denote entities may change. This occurs in \( \Gamma^5 \) with the predicate object(\( x \)), which becomes object(\( z \)) to avoid clashing with fired(\( x, A \)).

---

\(^6\) I gloss to pull out with unpack, which is a better rendition of B’s utterance. To pull out is a colloquial expression which means roughly to unpack for use.
### 8.4 · A segment of the plot as an iterated game

8.4.1 · \( t = 4 \) (again)

As I argued above, (114) is an SSM, because the typewriter is not assigned prior to this utterance’s production. The stegoform DRS of this utterance has been given above, in (115). This is a typical instance of the introduction of an entity into the plot-world. Its primary purpose is to introduce the typewriter as an existent of the plot-world, but it also reinforces the premise that \( A \) is at his workplace. Instead of \( A \) merely sitting at his old workplace, he is sitting at his old workplace doing the job that one presumes he was fired from. This is a fact that endows the plot-world with a plausibility that it might otherwise lack; or, put differently, it removes a strangeness that might diminish the scene’s plausibility.

8.4.2 · \( t = 5 \)

The response, line 11 of the datum, is an affirmative augmentation. It is reproduced as (117).

\[
(117) \quad \text{A: } \text{((begins to mime working on the typewriter)) a job doing (1.0) a job worth doing's worth doing right son}
\]

Since this is the last utterance of this brief analysis, let us examine it only as a response. This means ignoring the offer, which involves ascribing a curmudgeonly demeanour to A’s character.

The most salient part of the utterance with respect to its function as a response is A’s use of the typewriter. Before B asserts that it is a typewriter in line 8, A’s miming of the object is ambiguous, and the object is not identified as anything in particular. Only after B proposes that the object is a typewriter does A mime the use of a typewriter—inserting paper, winding the platen, and so on. The \textit{mime itself} signals that A accepts B’s proposal. And so, the stegoform of A’s utterance, with his subsequent offer excluded, is given in (118).

\[
(116) \quad \Gamma^5 = \begin{array}{|c|c|}
\hline
\text{fired}(x, A) & \text{workplace}(y) \\
\text{be_at}(A, B, y) & \text{father_of}(A, B) \\
\text{object}(z) & \text{unpacks}(A, z) \\
\text{employed_for}(A, \text{forty years}) & \neg \text{can be_at}(\neg y) \\
\hline
\end{array}
\]

Once again, \( \gamma^5 \) is derived by merging \( \mathcal{K}_4 \) with \( \gamma^4 \). And once again, the strategies selected in this period appear to be successful applications of natural language steganography, so the provisional expected utility at the conclusion of this period is (1, 1, 1).
That the object is a typewriter is thus incorporated into the common ground at $\Gamma^6$.

But there is more going on here: A no longer unpacks the object, but instead uses the typewriter. The entity that was previously an unidentified object is now a typewriter, and where A was previously unpacking it, he now uses it. These predications are updated. As we saw above, B proposes that the object is a typewriter, but he does not propose that A is using it. Instead, A’s mime predicates that he is using the typewriter. What results from these updates is the common ground DRS in (119).

At the end of this short segment, all periods appear to have successful steganographic concealment. At each period, then, the provisional expected utility profile is $(1, 1, 1)$, and so at the end of the fifth period, all players expect to receive 5 at the termination of the game. Recall what this engenders for the performance. It means that at each period, the actors have concealed metanarrative information from the audience, and provided the audience with enough information to select an interpretation that is plausible within the plot-world. According to the model of natural language steganography, the strategies of the actors in this segment of the scene are successful.

8.5 · Conclusion

This short application is designed to demonstrate how the model that has been developed in this thesis might be used to describe improvised performances. It provides a short example of an analysis, which deals with both the level of the plot, and the level of the individual utterance. From even this short analysis, certain general observations of communication in improvised theatre can be made.
1. The information that is available to the actors is far richer than the information that is available to the audience. That is, $\gamma^t(A, B)$ includes a far greater amount of information than the common ground $\Gamma^t$, at a given period $t$. While $\Gamma^t$ contains only that information that is accepted for inclusion in the common ground, $\gamma^t(A, B)$ contains all of the utterances that have been produced and the series of proposals and responses. The primary consequence of a substantial difference of information is that actors have many more contextual cues to guide their interpretations than the audience. Since the audience only has access to the common ground, and since information may only enter the common ground as a result of explicit acceptance by the actors, the actors are able to guide the interpretation of the audience. Meanwhile, the actors have the entire history of their covert communication to guide their interpretations, so they are able to interpret utterances with respect to their strategic function as a signal of other actors’ preferences over the plot trajectory.

In this way, the formal apparatus that I have defined is able to capture a peculiarity that is fundamental to communication in improvised theatre. It is furthermore able to express the differences in information that lead to differing interpretations of utterances, and to represent the interpretations themselves.

2. The analysis of this short segment and the description of the state transitions suggest that most utterances in improvised theatre function both as offers and as responses. There was no utterance in this excerpt that did not introduce new information about the plot-world. This suggests that actors prefer to constantly introduce information, especially early in a scene, perhaps so that the plot of the scene has a sense of progression, or of purpose. This may be explored by further analyses.

3. Conditions that are applied to existents are determined through multiple actors’ utterances. The most vivid illustration of this is the introduction of the typewriter in datum 23. First, the typewriter was introduced as an unidentified object by A, who mimed unpacking the object. B proposed that the object was a typewriter, which was subsequently accepted by A, by A’s miming the use of a typewriter. The identity of the object—that is, the condition typewriter($x$) imposed on the object—was determined through the actions of two actors, and in two stages.

This supports the suggestion that improvised theatre ought to be analysed as a communicative situation in which linguistic action can only be understood as co-constituted and negotiated. That is, interpreting any utterance involves the concerted action of multiple agents, and that any information may be accepted or rejected by any actor. Given this and other cases that have been examined throughout this thesis, it seems clear that only a co-constitutinal account of interpretation can thoroughly describe communication in improvised theatre.
In the preceding chapters, I have devised a model that allows for the formal description of the strategies that actors employ in longform improvisation. The model’s primary purpose is a general descriptive account of the interpretation of utterances based on common and partial ground, particularly over long interactions. It also provides an account of strategic devices that are more specific to improvised theatre. These include the means that actors use to generate characters and their traits (see chapter 5); the linguistic devices that actors use to create plot-worlds, and to maintain their consistency (see chapter 5); the ways that actors control the information that is shared amongst themselves, and with the audience (see chapter 6); and the methods by which actors conceal information regarding their preferences over the plot from the audience (see chapter 7). The descriptive capacity of the model was shown in chapter 8 through its application to a scene of improvised performance. This application showed that the model can adequately describe existing data, but also suggested that the inclusion of heuristics in the model improves the model’s capacity.

To build a model that has these capabilities, I added various components to a basic game theoretic model. To handle the semantics of the game model, I incorporated Discourse Representation Theory. This allows not only for a representation of utterance semantics with a greater degree of precision than the approaches generally used in game theoretic linguistics, but also for an account of the cumulative effect of a common ground on the interpretation of utterances, and a plausible representation of concealed information. To allow the model to describe concealed information, I adapted a general, information theoretic model of steganography. An information theoretic model was chosen because of its similarity to existing models of game theoretic communication, which permit the use of natural language signals to be emitted across the communication channel. The inclusion of other frameworks in game theoretic models is certainly not novel per se. For instance, Franke (2009) uses stepwise, levelled cognition that originates in research in psychology,
and Benz (2012a,b) uses an information theoretic model of communication as the basis of his error model. However, as far as I am aware, a general semantic theory such as DRT has not hitherto been used in game theoretic analysis, nor has an apparatus that utilises differences in players’ shared information been used to model interlocutors’ rational behaviour.

In chapter 1, I made three claims:

A That common ground in improvised theatre is built through iterative processes of negotiation.

B That for every possible combination of the parties to an improvised performance, multiple, divergent partial grounds are generated.

C That actors consciously exploit differences between the partial grounds to transmit information which is inaccessible to the audience.

I have substantiated each of these claims.

Claim A has been substantiated in chapters 6, 7, and 8. In chapter 6, I showed that actors form narrative common ground through an iterative process of proposal, acceptance, and rejection of information. This was made especially clear when covert communication was discussed in chapter 7, for the communication that is concealed by actors explicitly concerns the proposal and acceptance or rejection of information about the plot-world. In chapter 8, I reinforced this demonstration by analysing a segment of a scene. The resulting picture was one of conscious, concerted effort to establish a narrative common ground through a process of negotiation.

The formal implementation of claim B was defined in chapter 6, when partial grounds were introduced. Their existence was shown to be necessary for a description of covert communication, which appears to be a basic feature of language use in improvised theatre. Without partial grounds, there is no way of accounting for covert communication, for there is no means of expressing the differences between the information that is available to different agents. In chapter 8, I showed how the partial ground shared by the actors in a scene substantially differs in quantity from the common ground that is shared by the actors and the audience. I suggested that actors’ interpretations can differ from the audience’s interpretations precisely because the actors have access to much more information than the audience.

I have demonstrated that the exploitation of differences described in claim C pervades communication in improvised theatre. The formal implementation of a natural language steganographic system was defined in chapter 7, and I showed that a natural language stegosystem that exploits such differences in information can explain an actor’s assessments of her fellows’ preferences over the plot trajectory. It appears, too, that, since actors exploit these differences in contextual information, they are interested in ensuring that such differences exist. This intuition was supported by the analysis performed in chapter 8.
Chapter 9 · Conclusion and future work

9.1 · Future work

To conclude the thesis, I wish to highlight four directions for future research that flow from the discussions therein.

1. The model may provide a valuable basis for the advancement of the machine agent-based, decision theoretic research into the cognition of actors, which includes Magerko et al. (2009), Bruce et al. (2000), Baumer & Magerko (2009, 2010), Fuller & Magerko (2010), amongst others. The major contributions of this model to the formal study of improvised theatre are twofold. (1) By focussing on the linguistic aspect of improvised theatre, which is not addressed in the aforementioned literature, the effects of the progressive introduction of information about the plot-world can be better understood. Of particular importance is the ability of the model to deal with scenes in their entirety, and the information that constitutes an entire scene, rather than dealing only with devices used by actors at the level of the o–r pair. (2) The co-constitutive approach that the model takes to the construal of common ground and utterance meaning means that scenes can be described as truly emergent behaviour. The aforementioned literature identifies the practice’s emergent properties. But since it views improvised theatre as a series of decision theoretic problems, it does not adequately model emergence. Decision theory models the perspective of a single agent whose choice is made in the presence of environmental constraints. Therefore, under a decision theoretic approach, only a single actor’s decisions are modelled, while the decisions of the other actors are treated as constraints. On the other hand, game theoretic models are emergent by definition. For instance, calculation of utility in game theory requires the input of a strategy profile, which includes every player’s action. As such, the emergent properties of improvised theatre, which occur only as a result of concerted action by a number of actors, can be captured using the approach advocated here.

2. The two major advances that this model makes to the game theoretic study of natural language are: the inclusion of iterated games; and the incorporation of an intensional semantics into the game model. The first of these permits game theoretic linguistics to proceed from the present situation in which only conversations that are the length of a standard signalling game are studied. By using iterated signalling games to model conversations, this thesis provides a means of analysing longer interactions. While the iterated games that were defined in this thesis were designed primarily for use in describing data of improvised theatre, I believe that they may be successfully adapted for use in describing natural language generally.

The second major advance, which sees the inclusion of DRT in the game model, also provides future analysts with the means of extending the purview of game theoretic models of language use. It does so by providing a more thorough account of natural language semantics than is possible by mapping utterances onto propositions. The advantages of having a more complete utterance semantics as part of the game model have been shown.
9.2 · Limitations

The strongest of these is the inclusion of contextual information in the representation of the players’ deliberations, which allows for Parikh-style modelling of situations and their effects on the probabilities of certain meanings, but does so in a clearer, and more transparent, style.

3. Casting a wider net for more varied datasets will improve the toolkit available to game theoretic linguists. To my knowledge, there are no existing, extensive uses of data to inform game models of language use. I have argued throughout the thesis that, despite the constraints on communication in improvised theatre, it is a valid source of naturally-occurring speech data, for it is spontaneous, and the strategies employed by the actors span the actors’ full linguistic faculties. Other datasets of more naturalistic genres of naturally-occurring speech will further refine the findings that are offered in this thesis.

4. This thesis lays the foundation for a more thorough development of a model of natural language that includes steganography. Admittedly, the data that were used in this thesis to argue for the existence of natural language steganography, and to propose a means of modelling it, were from a single genre of naturally-occurring speech. This genre has the specific quality that most utterances produced in the speech of the genre have concealed information to some degree. This means that a model that describes the language of improvised theatre must include a provision for concealed information, and therefore devising a model of natural language steganography was necessitated for the thesis’s project.

Yet, other genres of naturally-occurring speech may also benefit from descriptions of steganography. Recall the example concerning Vladimir’s response to Estragon’s question about the pianist Lucky’s playing. Vladimir wishes to communicate that he did not enjoy the performance to Estragon, while signalling precisely the opposite to Lucky. While this example is an invented vignette, it is plausible to suggest that such situations, and such instances of language use, can be found in many naturally-occurring conversations. The study of natural language steganography will benefit from its application to data of more naturalistic genres of naturally-occurring speech.

9.2 · Limitations

My approach also has limitations, the rectification of which may render the model more effective and precise.

1. The analysis presented throughout the thesis was performed without any input directly from actors. This was a deliberate choice, and was made to conform to the usual, introspective approach taken to game theoretic studies of language. However, more may be revealed of actors’ processes of deliberation if interviews are conducted with the actors immediately following performances. For instance, strategies that are agreed upon prior to the performance may only be discovered by the analyst by questioning the actors directly.
2. Certain aspects of live performance were ignored for the model. Most saliently, neither the live improvised music that accompanied the improvisations in the data nor audience interaction were included in the definition of the model. Both of these aspects were omitted due to the complexity involved in treating them adequately. However, a comprehensive model of improvised theatre would benefit greatly from their inclusion. This can be addressed in future work.

3. Character could be better defined. In particular, I have ignored the fact that a character’s motivations may have a notable and serious effect on an actor’s decisions. As I pointed out in §2.2.2.1, situations may arise in which an actor’s preferences differ from those of the character she portrays. The assignment of utilities for characters, in addition to those for actors, was eschewed for the present study, because of the project’s primary focus on language use rather than on dramaturgy. A more faithful account of an actor’s deliberation would include some provision for character utility and actor utility, and how these interact.

4. With respect to the model of steganography, qualification of the degree to which the Adversary suspects that the other parties are communicating covertly has been ignored. It appears intuitively reasonable that there are various kinds of Adversaries in natural language. There are those who are completely naïve, and do not suspect that the interpretation intended for them is not exactly that which is intended for the Receiver. But of course the Adversary may be suspicious about the communication of the other agents. If the Adversary is suspicious, then the Sender must be more cautious about ensuring that the information is concealed, and that the message appears to be an innocent utterance of a normal conversation. While these considerations were not relevant to a model of improvised theatre, because the audience makes concessions in suspending disbelief when watching an improvised performance, there appear to be appropriate concerns for discussions of steganography in other genres of naturally-occurring speech.

5. Any restrictions on what is able to be transmitted across the subliminal channel in naturally-occurring speech remain unexplored, for it was only the three major functions of proposal, acceptance, and rejection that were identified in this study. Other kinds of message may be able to be concealed. Further study of other improvised performances, and studies involving data from other genres of interaction, may be able to identify such messages.

6. In certain cases of natural language steganography, the analyst may have a similar epistemology to the Adversary, and therefore may be unable to detect hidden messages in data of naturally-occurring speech. This problem is overcome in improvised theatre by having restrictions on what is possible in the genre. That is, since actors must plan while maintaining the plot-world’s verisimilitude, the domain of hidden messages is limited to
the three narrative functions that were identified in chapter 7: proposal, acceptance, and rejection. In other genres where these domain restrictions are not present, hidden messages may be far more complex, and the information required to detect them may be far more comprehensive. In these situations, the analyst would need to design studies in such a way that the participants of the study are able to self-report. Self-reporting gives the participants a means of informing the analyst what the concealed messages are. Using this information, the analyst would be able to examine the linguistic devices that are used by interlocutors to conceal messages. Applied to other genres of naturally-occurring speech, such research may give rise to further studies of natural language stegosystems.

The multidisciplinary nature of this thesis means that the potential future directions for research are manifold. I hope that this thesis has primarily demonstrated the utility of approaches that combine theoretical frameworks in order to better understand the deliberative processes of interlocutors engaged in improvised performance, and to understand the rational, strategic use of language that is fundamental to this genre of naturally-occurring speech.
References


References


References


Schöttle, Pascal & Rainer Böhme. 2012. A game-theoretic approach to content-adaptive steganography. In Matthias Kirchner & Dipak Ghosal (eds.), *Proceedings of the Four-


### List of Symbols and Notation

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathbb{R}$</td>
<td>the set of real numbers</td>
</tr>
<tr>
<td>$\mathbb{N}$</td>
<td>the set of natural numbers</td>
</tr>
<tr>
<td>$\wp(X)$</td>
<td>powerset of set $X$</td>
</tr>
<tr>
<td>$\emptyset$</td>
<td>the empty set</td>
</tr>
<tr>
<td>$Pr(x)$</td>
<td>probability of event $x$</td>
</tr>
<tr>
<td>$Pr(x</td>
<td>y)$</td>
</tr>
<tr>
<td>$\delta_X$</td>
<td>probability distribution over $X$</td>
</tr>
<tr>
<td>$\Delta(X)$</td>
<td>set of probability distributions $\delta_X$</td>
</tr>
<tr>
<td>$a, s$</td>
<td>action, strategy</td>
</tr>
<tr>
<td>$S_i$</td>
<td>strategy set of player $i$</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>strategy space</td>
</tr>
<tr>
<td>$U : \sigma \to \mathbb{N}$</td>
<td>utility function</td>
</tr>
<tr>
<td>$\hat{u}_i(s)$</td>
<td>expected utility of player $i$ for $s$</td>
</tr>
<tr>
<td>$W$</td>
<td>set of possible worlds</td>
</tr>
<tr>
<td>$S$</td>
<td>Sender</td>
</tr>
<tr>
<td>$R$</td>
<td>Receiver</td>
</tr>
<tr>
<td>$\rho_{S,R}$</td>
<td>sender, receiver belief</td>
</tr>
<tr>
<td>$\mu$</td>
<td>set of natural language messages</td>
</tr>
<tr>
<td>$\varepsilon$</td>
<td>the empty utterance</td>
</tr>
<tr>
<td>$\eta^t$</td>
<td>history at period $t$</td>
</tr>
<tr>
<td>$K$</td>
<td>DRS</td>
</tr>
<tr>
<td>$\kappa$</td>
<td>DRS set</td>
</tr>
<tr>
<td>$\mathcal{L}$</td>
<td>DRS language</td>
</tr>
<tr>
<td>$\mathcal{M}_\mathcal{L}$</td>
<td>model for $\mathcal{L}$</td>
</tr>
<tr>
<td>$\mathcal{D}$</td>
<td>set of individuals</td>
</tr>
<tr>
<td>$\mathcal{I}$</td>
<td>interpretation function</td>
</tr>
<tr>
<td>$</td>
<td></td>
</tr>
<tr>
<td>$\triangleright$</td>
<td>DRS merge</td>
</tr>
<tr>
<td>$\vDash$</td>
<td>permission implication function</td>
</tr>
<tr>
<td>$A, B, C, \ldots$</td>
<td>actors/characters</td>
</tr>
<tr>
<td>$P$</td>
<td>audience</td>
</tr>
<tr>
<td>$q$</td>
<td>state of the scene</td>
</tr>
<tr>
<td>$Q$</td>
<td>set of states of the scene</td>
</tr>
<tr>
<td>$\psi$</td>
<td>set of scenes</td>
</tr>
<tr>
<td>$Q$</td>
<td>scene</td>
</tr>
<tr>
<td>$E$</td>
<td>plot</td>
</tr>
<tr>
<td>$G$</td>
<td>set of contexts</td>
</tr>
<tr>
<td>$\Gamma$</td>
<td>common ground</td>
</tr>
<tr>
<td>$\gamma(i,j)$</td>
<td>partial ground shared by $i$ and $j$</td>
</tr>
<tr>
<td>$h(x)$</td>
<td>information content of outcome $x$</td>
</tr>
<tr>
<td>$H(X)$</td>
<td>information entropy of a discrete random variable $X$</td>
</tr>
</tbody>
</table>

262
List of Symbols and Notation

\[ H(X|Y) \]: conditional entropy of \( X \) given \( Y \)  
\[ D(\delta_p \| \delta_q) \]: relative entropy of \( \delta_p \) with respect to \( \delta_q \)  
\( \mathcal{E} \): natural language stegosystem  
\( E \): Adversary  
\( \bar{K} \): stegoform DRS  
\( c, s \): coverform, stegoform  
\( \delta_{\mu|W}^c, \delta_{\mu|W}^s \): distributions over the coverform, stegoform
Appendix A
Heuristics, their definitions, and assumptions

Heuristic 1: Verisimilitude. An action is called optimally verisimilar if it appears to believably belong to the plot-world; that is, if it does not appear to be produced by an actor pretending to be a character.

Definition A.1: Optimal verisimilitude (from heuristic 1). Let \( \Sigma \) be a natural language stegosystem, and fix coverform and stegoform distributions \( \delta^c_{\mu|W} \) and \( \delta^s_{\mu|W} \) over \( \mu \), respectively. A Sender strategy \( m \in \mu \) is optimally versimilar iff 
\[
D(\delta^s_{\mu|W} \parallel \delta^c_{\mu|W}) = 0.
\]

Heuristic 2: Compatibility. A strategy is optimally compatible if the facts expressed by that strategy do not contradict what is established about the plot-world prior to that strategy’s deployment.

Definition A.2: Optimal compatibility (from heuristic 2). Let \( \Gamma^t \) be a context at time \( t \). The set of optimally compatible strategies \( S^* \) is the smallest set such that, for each \( s^* \in S^* \), \( s^* \in \|K\| \), such that \( \|K\| \subseteq \|\Gamma^t\| \).

Heuristic 3: Salience. If a Sender intends that a scene be resumed, a strategy is optimally salient with respect to that scene (called the target scene) if there is no other scene that is more probably the target scene given the strategy.

Definition A.3: Optimal salience (from heuristic 3). Let \( Q_j \) be a scene, and \( Q_1, \ldots, Q_{i-1} \) be the preceding scenes. Let each scene \( Q_j \) have a context \( \Gamma_{Q_j} \), and let \( \Gamma_{Q_1} = \emptyset \). A scene-initial strategy for Speaker \( S \), \( s^*_S \in S_S \), is optimally salient with respect to a target scene \( Q_j \in \{Q_1, \ldots, Q_{i-1}\} \) iff

1. there is no \( s' \in S_S \) such that \( \Pr(\Gamma_{Q_j}|s') > \Pr(\Gamma_{Q_j}|s^*_S) \); and
2. there is no scene \( Q'_j \in \{Q_1, \ldots, Q_{i-1}\} \) such that \( \Pr(\Gamma_{Q_j}|s^*_S) > \Pr(\Gamma_{Q_j}|s^*_S) \).

Heuristic 4: Distinction. If a Sender intends that a scene be resumed, a strategy is optimally distinct if only the intended differences between the target scene’s context and the present scene’s context are communicated.
Definition A.4: Optimal distinction (from heuristic 4). Let $Q_0$ and $Q_1$ be two scenes, such that $Q_0 < Q_1$, with contexts $\Gamma_{Q_0}$ and $\Gamma_{Q_1}$, such that initially $\Gamma_{Q_1} = \emptyset$. A strategy $m^*$ is optimally distinct iff the following conditions are met:

(i) after observation of $m^*$, $\Gamma_{Q_1}$ becomes $\emptyset \cup \Gamma_{Q_0} \cup K$ such that $m^* \in \|K\|; and

(ii) $[\Gamma_{Q_1}] \subset [\Gamma_{Q_0}]$.

Assumption A.1: If a scene is resumed, all facts are assumed to be identical to those in the target scene, unless this assumption is explicitly overturned.

Heuristic 5: Plausibility. A strategy is optimally plausible in the plot-world if, according to the information that is already known about the plot-world, the plot-world has a high probability of being an element of the extension of the strategy’s DRS.

Definition A.5: Optimal distinction (from heuristic 5). Let $m$ be an utterance, let $\kappa$ be a set of DRSs such that for each $K \in \kappa$, it holds that $m \in \|K\|$. Furthermore, let $\Gamma_Q$ be the context of scene $Q$, and let $w$ be the plot-world. A strategy $m$ is optimally plausible in $w$ iff $\arg\max_{K \in \kappa} Pr(w \in [K]|\Gamma_Q)$.

Assumption A.2: If some aspect of the plot-world is unknown, then it is assumed that that aspect is identical in the actual world.

Heuristic 6: Optimal markedness. Given a particular utterance, and a series of responses such that one response is made by each of the parties to a conversation, an operative interpretation of the utterance is optimally marked if it is most probable given all of the responses.

Definition A.6: Optimal markedness (from heuristic 6). Let $G^{n+1}$ be an iterated game constituted of $n + 1$ interpreted signalling games, such that $g^1$ has a specific Sender $i$, and such that for the following games $g^2, \ldots, g^n$, each player in $N$ (including $i$) is the Sender for exactly one game. Recall that $s^t$ denotes a strategy profile of length $t$. An interpretation $K \in \kappa$ is optimally marked if it solves

$$\arg\max_{K \in \kappa} Pr(K|m|s^{n+1}).$$ (A.1)

Heuristic 7: Convergence of repetition strategies. Given an intended interpretation and a series of utterances meant to convey this interpretation, a strategy of repetition achieves optimal convergence if and only if the intersection of the extensions of each of the utterances contains only the intended interpretation.

Definition A.7: Optimal convergence of a repetition strategy (from heuristic 7). Let $G^k$ be a composite signalling game constituted of $k$ subgames. A series of utterances $m_1, \ldots, m_k$, where each is produced in its own subgame, is optimally convergent as a repetition strategy with respect to a particular state of the world $w \in \mathcal{W}$ iff
(1) \( w \in [K_{m_1} \cup \ldots \cup K_{m_k}] \) where \( K_{m_i} \) denotes that \( m_i \in ||K_m|| \); and 

(2) there is no \( w' \in W \), such that if \( w' \neq w \), then \( w' \in [K_{m_1} \cup \ldots \cup K_{m_k}] \).

**Heuristic 8:** Optimal polarity. A response to an offer is optimally polar if it is unambiguously the case that what is intended to be conveyed by the offer utterance is taken up and subsequently accepted or rejected.

**Definition A.8:** Optimal polarity (from heuristic 8). Given an utterance \( m \), and an interpretation \( K \) such that \( m \in ||K|| \), an optimally polar response strategy solves either

\[
\arg \max_{m \in \mu} \Pr(K|m) \quad (A.2)
\]

if it is positive, or

\[
\arg \min_{m \in \mu} \Pr(K|m) \quad (A.3)
\]

if it is negative.

**Heuristic 9:** Default acceptance. A response strategy is by default positive.

**Heuristic 10:** An a priori actor subcontexts. For every subset of the set of actors involved in an improvised performance, there is always a subcontext that they share, such that the subcontext contains the information that

(1) the other agent is an actor;
(2) the activity they are performing involves constructing a plot;
(3) the structure of communication in the genre is that of iterated offer-response pairs; and
(4) there is an audience present.

**Heuristic 11:** Metanarrative function. An embedded meaning is optimally functional if the probability of one of the predicates proposes, accepts, or rejects being the embedded meanings, conditional on the actors’ shared subcontext, is greater than the respective probabilities of the other two.

**Definition A.9:** Optimal functionality, from heuristic 11. A stegoform \( s \) is optimally functional if there is an \( o \in \{\text{propose, accept, reject}\} \) such that, for each of the other functions \( o' \), \( \Pr(o|s) \geq \Pr(o'|s) \).