Declaration

This thesis has been composed by myself and has not been submitted for any other degree or professional qualification. The research presented herein is my own, except where due acknowledgement is made by reference.

The data for Experiment 1 was collected in collaboration with Manabu Arai, but was primarily my responsibility. In turn, I contributed to the collection of data for the Experiments named A and B (see Chapter 2), for which he was responsible. All analyses reported in this thesis for Experiments 1-7 are original and were conducted by me.
Abstract

Although language usually occurs in an interactive and world-situated context (Clark, 1996), most research on language use to date has studied comprehension and production in isolation. This thesis combines research on comprehension and production, and explores the links between them. Its main focus is on the coordination of visual attention between speakers and listeners, as well as the influence this has on the language they use and the ease with which they understand it.

Experiment 1 compared participants’ eye movements during comprehension and production of similar sentences: in a syntactic priming task, they first heard a confederate describe an image using active or passive voice, and then described the same kind of picture themselves (cf. Branigan, Pickering, & Cleland, 2000). As expected, the primary influence on eye movements in both tasks was the unfolding sentence structure. In addition, eye movements during target production were affected by the structure of the prime sentence. Eye movements in comprehension were linked more loosely with speech, reflecting the ongoing integration of listeners’ interpretations with the visual context and other conceptual factors.

Experiments 2-7 established a novel paradigm to explore how seeing where a speaker was looking during unscripted production would facilitate identification of the objects they were describing in a photographic scene. Visual coordination in these studies was created artificially through an on-screen cursor which reflected the speaker’s original eye movements (cf. Brennan, Chen, Dickinson, Neider, & Zelinsky, 2007). A series of spatial and temporal manipulations of the link between cursor and speech investigated the respective influences of linguistic and visual information at different points in the comprehension process. Implications and potential future applications are discussed, as well as the relevance of this kind of visual cueing to the processing of real gaze in face-to-face interaction.
Acknowledgements

Having just declared that the thesis is all my own work, I would not have been able to produce it without the help, support, and encouragement of a large number of people.

First and foremost, I am extremely grateful to Martin Pickering and Sarah Haywood for bringing me to Edinburgh at all, and to John Henderson for joining us. Their complementary mix of skills made them an ideal supervisory team at all stages. I am particularly grateful to Martin for his open-door policy while I was writing up in the final half year – it even worked electronically!

Manabu Arai collaborated with me on setting up and running Experiment 1, and has been a mentor, stats advisor, and good friend ever since. That experiment would also not have been possible without the flexibility and conscientiousness of our confederates, Jacqueline Thomson and Kate Messenger. Manon Jones helped me develop the idea behind the gaze projection paradigm. Jiye Shen, Robin Hill, Tim Smith, Jens Apel, Robert Maier, Frances Wilson, Reinhild Glanemann, Nadine Kloth, and Mateo Obregón have all provided technical support and/or statistical advice at different times. I am also very grateful to Fernanda Ferreira for the use of her lab, and for welcoming me to her lab meetings. In fact, I could probably thank most of the members of the psycholinguistic community in Edinburgh, in one way or another. Writing the thesis has made me realise just how much I have learnt in the past 3-and-a-bit years, from reading groups and coffee meetings to world-class talks. At the same time, the companionship of the postgrad community has been invaluable. In particular, I am grateful to have had such great office mates in F23 (past and present) and for the friendship of Anne Finucane, Chris Burns, and Manon Jones. My family, as well as former flatmates, relatives, and several friends made it easy to move from Germany to Scotland and back again, by providing accommodation, transport, administrative or financial support. To them all, thankyou. Above all, danke an Fabian for not protesting when I decided to go off to Scotland, and for welcoming me back.
# Table of Contents

Declaration i  
Abstract iii  
Acknowledgements v  
Table of Contents vii  
List of Tables xv  
List of Figures xvii

## CHAPTER 1: GENERAL INTRODUCTION 1

1.1 Eye movements and what they can tell us 3  
   1.1.1 Eye movements and visual attention 4  
   1.1.2 Gaze control: Determinants of fixation location and duration 4  
   1.1.3 Terminology 8  
   1.1.4 The visual-world paradigm 9

1.2 Language Production 12  
   1.2.1 How to get from thoughts to articulation: A model of speech production 13  
   1.2.2 Empirical findings for the production of single words 15  
   1.2.3 Eye movements during the production of object names 16  
   1.2.4 Eye movements during the production of complex utterances 19  
   1.2.5 Conceptual preparation and starting points 20  
   1.2.6 The assumption of incrementality in production 23  
      1.2.6.1 How much of the sentence must be prepared at its onset? 23  
      1.2.6.2 How soon is the verb encoded? 25  
      1.2.6.3 Strategic incrementality 26  
   1.2.7 Summary: Language production 27

1.3 Language Comprehension 28  
   1.3.1 Critical features of auditory comprehension 29  
   1.3.2 How to get from hearing to understanding: A model of comprehension 30  
   1.3.3 Parsing models 32
1.3.4 The role of context in auditory sentence processing 36
  1.3.4.1 Eye movements during comprehension in a visual world 38
  1.3.4.2 The Coordinated Interplay Account 40
  1.3.4.3 The ‘good-enough’ approach to comprehension 42
1.3.5 Summary: Language comprehension 44

1.4 Dialogue and interactive language use 45
  1.4.1 Dialogue as joint action 46
    1.4.1.1 Differences between dialogue and monologue 46
    1.4.1.2 Shared responsibility for communicative success in dialogue 48
    1.4.1.3 Implications of dialogue for production and comprehension 49
  1.4.2 Interactive comprehension and production: A model of dialogue 51
  1.4.3 Perspective-taking in reference assignment 53
    1.4.3.1 The use of common ground by speakers and listeners 54
    1.4.3.2 Eye movements as early indicators of reference assignment 56
  1.4.4 Can eye movements reflect alignment? 60
  1.4.5 Summary: Dialogue 61

1.5 Overview of the experiments 62

CHAPTER 2: SYNTACTIC PRIMING AND EYE MOVEMENTS IN COMPREHENSION AND PRODUCTION – EXPERIMENT 1 65

2.1 Introduction 65
  2.1.1 Syntactic priming in an experimental context: Overview of findings 65
    2.1.1.1 Functional accounts of syntactic priming 66
    2.1.1.2 Syntactic priming in dialogue: The confederate-scripting paradigm 69
  2.1.2 Coordination vs. dissociation accounts of eye movements in dialogue 71

2.2 Eye movements during comprehension and production of related sentences 74
  2.2.1 Predictions regarding eye movements in comprehension and production 74
  2.2.2 Methods 75
    2.2.2.1 Materials 75
    2.2.2.2 Design 76
    2.2.2.3 Participants 77
    2.2.2.4 Apparatus and Confederates 77
    2.2.2.5 Procedure 78
  2.2.3 Analysis 79
CHAPTER 3: THE USE OF AN INTERLOCUTOR’S GAZE IN LANGUAGE PROCESSING

3.1 The role of gaze in face-to-face conversation
   3.1.1 Gaze detection and gaze following
   3.1.2 Functions of the addressee’s gaze in dialogue
   3.1.3 Functions of the speaker’s gaze in dialogue
   3.1.4 The role of joint attention

3.2 Gaze projection
   3.2.1 Precedents
   3.2.2 Advantages and limitations of the paradigm

3.3 Overview of the gaze-projection experiments
## CHAPTER 4: TESTING THE PARADIGM – EXPERIMENT 2

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Research question</td>
<td>127</td>
</tr>
<tr>
<td>4.2 Design and predictions</td>
<td>128</td>
</tr>
<tr>
<td>4.3 Methods and materials</td>
<td>130</td>
</tr>
<tr>
<td>4.3.1 Stimulus and item selection</td>
<td>130</td>
</tr>
<tr>
<td>4.3.2 Creating the gaze and speech stimuli</td>
<td>131</td>
</tr>
<tr>
<td>4.3.2.1 Recording procedure</td>
<td>131</td>
</tr>
<tr>
<td>4.3.2.2 Creating the gaze cursor and interest areas</td>
<td>133</td>
</tr>
<tr>
<td>4.3.2.3 Determining items from the speech recordings</td>
<td>133</td>
</tr>
<tr>
<td>4.3.3 Speaker results</td>
<td>134</td>
</tr>
<tr>
<td>4.3.3.1 Speakers’ item descriptions</td>
<td>135</td>
</tr>
<tr>
<td>4.3.3.2 Speakers’ fixation patterns</td>
<td>137</td>
</tr>
<tr>
<td>4.3.4 Listener methods</td>
<td>139</td>
</tr>
<tr>
<td>4.3.4.1 Participants</td>
<td>139</td>
</tr>
<tr>
<td>4.3.4.2 Procedure</td>
<td>139</td>
</tr>
<tr>
<td>4.3.4.3 Data processing and Analysis</td>
<td>141</td>
</tr>
<tr>
<td>4.4 Listener results</td>
<td>142</td>
</tr>
<tr>
<td>4.4.1 First fixation onsets</td>
<td>142</td>
</tr>
<tr>
<td>4.4.2 Click latencies</td>
<td>145</td>
</tr>
<tr>
<td>4.4.3 Memory tasks</td>
<td>146</td>
</tr>
<tr>
<td>4.5 Summary and Conclusions</td>
<td>148</td>
</tr>
</tbody>
</table>

## CHAPTER 5: VARYING INSTRUCTIONS ABOUT THE GAZE CURSOR – EXPERIMENT 3

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1 Research question and predictions</td>
<td>153</td>
</tr>
<tr>
<td>5.2 Methods</td>
<td>154</td>
</tr>
<tr>
<td>5.2.1 Design</td>
<td>154</td>
</tr>
<tr>
<td>5.2.2 Materials</td>
<td>155</td>
</tr>
<tr>
<td>5.2.3 Participants</td>
<td>156</td>
</tr>
<tr>
<td>5.2.4 Procedure</td>
<td>156</td>
</tr>
<tr>
<td>5.2.5 Data processing and analysis</td>
<td>157</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>---------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>5.3 Results</td>
<td>158</td>
</tr>
<tr>
<td>5.3.1 First fixation onsets</td>
<td>158</td>
</tr>
<tr>
<td>5.3.2 Click latencies</td>
<td>160</td>
</tr>
<tr>
<td>5.3.3 Questionnaire results</td>
<td>161</td>
</tr>
<tr>
<td>5.4 Conclusions</td>
<td>164</td>
</tr>
<tr>
<td>CHAPTER 6: METHODOLOGICAL ISSUES – EXPERIMENT 4</td>
<td>167</td>
</tr>
<tr>
<td>6.1 Research questions and predictions</td>
<td>167</td>
</tr>
<tr>
<td>6.2 Methods</td>
<td>169</td>
</tr>
<tr>
<td>6.2.1 Design</td>
<td>169</td>
</tr>
<tr>
<td>6.2.2 Materials</td>
<td>169</td>
</tr>
<tr>
<td>6.2.3 Participants</td>
<td>170</td>
</tr>
<tr>
<td>6.2.4 Procedure</td>
<td>170</td>
</tr>
<tr>
<td>6.3 Preliminary results</td>
<td>171</td>
</tr>
<tr>
<td>6.4 Problems</td>
<td>172</td>
</tr>
<tr>
<td>6.4.1 Speakers’ differences in descriptiveness</td>
<td>173</td>
</tr>
<tr>
<td>6.4.2 Differences in the accuracy of speakers’ eye movement records</td>
<td>174</td>
</tr>
<tr>
<td>6.4.3 Low listener-speaker ratio</td>
<td>174</td>
</tr>
<tr>
<td>6.4.4 Order of trial presentation</td>
<td>175</td>
</tr>
<tr>
<td>6.4.5 Was the lag manipulation too short?</td>
<td>176</td>
</tr>
<tr>
<td>6.4.6 Inaccuracies of listener eye movements</td>
<td>177</td>
</tr>
<tr>
<td>6.5 Conclusions and potential paradigm modifications</td>
<td>178</td>
</tr>
<tr>
<td>CHAPTER 7: REPLICATION AND EXTENSION – EXPERIMENT 5</td>
<td>181</td>
</tr>
<tr>
<td>7.1 Research question and predictions</td>
<td>181</td>
</tr>
<tr>
<td>7.2 Methods</td>
<td>182</td>
</tr>
<tr>
<td>7.2.1 Design</td>
<td>182</td>
</tr>
<tr>
<td>7.2.2 Materials: Selection of speaker recordings</td>
<td>182</td>
</tr>
<tr>
<td>7.2.3 Participants</td>
<td>183</td>
</tr>
<tr>
<td>7.2.4 Apparatus</td>
<td>183</td>
</tr>
</tbody>
</table>
7.2.5 Procedure

7.3 Results

7.3.1 First fixation onsets
7.3.2 Click accuracy and latencies
7.3.3 Questionnaire results

7.4 Conclusions

7.4.1 Contributions of speech and gaze cursor
7.4.2 Comparison with previous experiments

7.5 Summary and Discussion

CHAPTER 8: VARYING LAG TIMES - EXPERIMENTS 6 & 7

8.1 Experiment 6: Large lag variations and a misleading cursor
8.1.1 Research question and predictions
8.1.2 Methods
8.1.2.1 Design
8.1.2.2 Materials
8.1.2.3 Participants
8.1.2.4 Procedure
8.1.3 Results and Discussion
8.1.3.1 First fixation onsets
8.1.3.2 Click latencies
8.1.3.3 Summary of recall task results
8.1.4 Conclusions

8.2 Experiment 7: Small lag variations
8.2.1 Research question and Predictions
8.2.2 Methods
8.2.3 Results
8.2.3.1 First fixation onsets
8.2.3.2 Click latencies
8.2.4 Conclusions

8.3 Cross-experiment analyses
8.3.1 First fixation onsets and click latencies
8.3.2 Fix-to-click times
8.3.3 Correlation analyses 219
8.3.4 Questionnaire Data 220

8.4 Summary and Discussion 223

CHAPTER 9: GENERAL DISCUSSION 227

9.1 Summary of the experiments 227

9.2 Discussion of the gaze-projection experiments 230
  9.2.1 Limitations: A return to the ‘real gaze’ question 231
  9.2.2 Possibilities for further studies 232

9.3 From comprehension to production and back again: Coordinated attention 236

References 239

Appendix A. Eyepriming: Experimental items 265
Appendix B. Gaze projection photos and ROIs 267
Appendix C. Gaze projection: Item selection 270
Appendix D. Gaze projection: Example speaker description 272
Appendix E. Gaze projection: Post-experimental questionnaire 273
Appendix F. Gaze projection: Post-experimental recognition task 278
Appendix G. Gaze projection: Recall task (following Experiment 6) 281
# List of Tables

Table 2-1: Active and passive target descriptions depending on prime condition. .......... 85
Table 2-2: Factors affecting participants’ passive target structures. ................................. 86
Table 2-3: Overview of onset times by prime-target condition. ....................................... 87
Table 2-4: Analyses of the logit of gazes while listening to prime sentences following the N1 and N2 onsets. .................................................................................. 89
Table 2-5: Analyses of the logit of participants’ gazes immediately preceding the confederate’s speech onset. .................................................................................................................. 90
Table 2-6: Analyses of the logit of gazes preceding the N1 and N2 onsets for producing target sentences. .......................................................................................................................... 93
Table 4-1: Descriptive statistics per speaker...................................................................... 136
Table 4-2: Proportion of items with at least one fixation to the object currently being described. ................................................................................................................................. 137
Table 4-3: Number of correct responses in the memory tasks. ........................................ 147
Table 6-1: First fixation onsets and click latencies for the three cursor conditions......... 172
Table 7-1: Subcategories of clicks by condition................................................................. 189
Table 7-2: Descriptive statistics for rating questions. ....................................................... 192
Table 8-1: Descriptive statistics for comparing the (relatively) disconnected conditions of Experiments 5 and 6. ................................................................................................................. 204
Table 8-2: Descriptive statistics for the five speech-cursor lags: Onset times of first relevant gazes and click latencies. .................................................................................................................. 210
Table 8-3: Time (in ms) from first relevant fixation to correct click, for each lag. ............ 217
Table 8-4: Spearman’s rank correlations of first fixation onset and click latency for all lag conditions................................................................................................................................. 219
Table 8-5: Descriptive statistics for rating questions.............................................................. 221
Table 8-6: Participants’ post-experimental guesses of the condition they were assigned to. 222
List of Figures

Figure 2-1: Example of a prime or target image in Experiment 1. ........................................ 75
Figure 2-2: Gaze logits to the agent and patient during comprehension of active and passive
prime sentences. ......................................................................................................................... 88
Figure 2-3: Logit of gazes to the agent and patient, depending on prime-target condition.... 92
Figure 4-1: Example of experimental stimulus. ................................................................. 130
Figure 4-2: Regressive time graph of fixation proportions before name onset .................. 138
Figure 4-3: Mean onset of the first relevant fixation, depending on the presence of the gaze
cursor. ...................................................................................................................................... 143
Figure 4-4: Proportion of first gazes per cursor condition to the relevant object before and
after name onset .......................................................................................................................... 144
Figure 4-5: Click latency depending on the presence of the gaze cursor ......................... 145
Figure 4-6: Re-use of speakers’ original referring expression for listeners with and without
cursor, and for speakers.............................................................................................................. 148
Figure 5-1: Mean onset latency of the first relevant fixation, depending on instruction and
trial number. ............................................................................................................................... 159
Figure 5-2: Mean click latency depending on the gaze cursor instruction and trial number.
.................................................................................................................................................... 160
Figure 6-1: Median first fixation onsets for each cursor condition and speaker........... 173
Figure 6-2: Median first fixation onsets depending on the order in which each cursor
condition was experienced............................................................................................................. 176
Figure 7-1: Onset of the first relevant fixation, depending on the type of stimulus........ 186
Figure 7-2: Mean onset times of the first relevant gaze. ..................................................... 187
Figure 7-3: Mean latencies for unambiguously correct clicks depending on gaze cursor
condition and trial number.......................................................................................................... 190
Figure 8-1: Mean onset of the first relevant fixation, depending on gaze cursor condition and
trial number. ............................................................................................................................... 202
Figure 8-2: Mean latency to correct click depending on gaze cursor condition and trial
number. ........................................................................................................................................ 205
Figure 8-3: Boxplot showing onsets of first relevant gazes, depending on the gaze cursor lag. ................................................................. 209

Figure 8-4: Onsets of first relevant gazes depending on the gaze cursor lag and trial number . ................................................................. 210

Figure 8-5: Mean click latencies depending on gaze cursor lag and trial number. .......... 211

Figure 8-6: Experiments 6 and 7. Mean first fixation onsets and click latencies across time. 214

Figure 8-7: Fix-to-click measures for all lag conditions. ................................................. 217

Figure 8-8: Median time from first looking at an item to correctly clicking on it .............. 218
Chapter 1: General Introduction

Psycholinguistics, the psychological study of language, typically distinguishes between the two primary mental processes involved in language use, comprehension and production. However, while both areas have generated substantial amounts of research and many important findings, they have frequently been studied in isolation, and often even by different people. This practice may be justified for both theoretical and practical reasons at the level of the individual experimental study, but it ignores that fact that most language is produced and comprehended in an interactive context: “The most primordial and universal setting for speech is conversational, free interaction between two or more interlocutors” (Levelt, 1989, p. 29; see also Clark, 1996; Pickering & Garrod, 2004). Participants in such a conversation must alternate quickly between the two complementary tasks of preparing the message they wish to communicate, and making sense of the signal they have received. In fact, they will probably often do both simultaneously, for instance when planning their next contribution while following what their interlocutor is saying.

In such a situation, it seems highly likely that the comprehended utterance will influence the subsequent production, whether at the level of sentence structure, lexical selection, or with regard to the current conversational topic. Similarly, cooperative speakers will usually consider the comprehensibility of their utterances and try to make them relatively easy to understand. Again, considering the addressee’s needs can involve any level, from using a language he or she understands and speaking clearly, to selecting an unambiguous structure, employing reference terms that have been mutually established, and providing a sensible and interesting continuation to the previous discourse. Finally, the context in which the interaction is taking place and what has been said so far can be expected to influence both comprehension and production. For example, whether the object under discussion is in view and whether it has already been mentioned
will affect how the speaker refers to it, as well as the speed with which the addressee identifies it correctly.¹

The relevance of context for both comprehension and production, in addition to their reciprocal influence on each other, means that the study of interactive language use has become an essential aspect of psycholinguistics (cf. many of the contributions to Trueswell & Tanenhaus, 2005). Accordingly, this thesis combines research on comprehension and on production and explores the links between them. In particular, I am interested in the coordination of visual attention between speakers and listeners, as well as how it influences the language they produce and the ease with which they understand it. I have used two quite different paradigms for this purpose, though both involved tracking participants’ eye movements while they performed a language-based task.

The remainder of this chapter contains an introduction to each of the four areas of research involved, intended as a general background for the individual experiments. These are subsequently described in detail in the following chapters, including their rationale and theoretical motivation. As a result, the connections between the different areas may not always be obvious in the current chapter, though I emphasise those aspects that will become important later.

I begin by discussing eye movements and visual attention while viewing a scene (Section 1.1), and how this relates to language processing. I then summarise relevant psycholinguistic theories and findings individually for language production (1.2) and comprehension (1.3), as well as research concerning interaction and dialogue (1.4). Each of these three sections also considers what eye movements can tell us about the respective aspects of language processing. Finally, the chapter concludes with an overview of the experiments and the remainder of the thesis (1.5).

¹ Since this thesis discusses language comprehension and production as well as different forms of interaction, the term speaker is used for anyone producing spoken utterances. Comprehenders are referred to as addressees if they are able to interact with the speaker, and as listeners if they have no means of providing feedback. The term interlocutors designates partners in interaction without pre-assigned roles.
1.1 Eye movements and what they can tell us

Eye movements are one of the most frequent muscle movements made by humans, with gaze shifts occurring roughly three times per second, about 230,000 times during a waking day (Hoffman, 1998). The primary reason for shifting gaze is the fact that only the central fovea, comprising about 2° of visual angle, is able to gather visual information in high resolution. Parafoveal processing from the area surrounding the fovea (about 5° around the centre of fixation) is much less accurate; and while information from the periphery is still useful for selecting the next point of fixation, its acuity decreases further (Liversedge & Findlay, 2000). Most information is therefore acquired foveally, during the periods of steady fixation.

This means that frequent shifts of visual focus are necessary to acquire a clear image of the environment. They are achieved by very fast saccades, which reach a speed of over 500°/s (Rayner, 1998). Saccadic suppression ensures that practically no visual information is taken in during these brief ballistic movements, though processing of previously fixated items can continue (Matin, 1974; Irwin, 2004).

Which region of the image a saccade and the subsequent fixation are directed to is largely dependent on attention: precisely because focussing on an object allows them to gather detailed information about it, humans tend to move their eyes to whatever is most relevant to their goal at a particular moment in time (Hayhoe, 2000; Hayhoe & Ballard, 2005). In fact, this attention to task-relevant information was first demonstrated in pioneering work by Buswell (1935; cited by Henderson & Ferreira, 2004) and Yarbus (1967). By tracking participants’ eyes while they looked at paintings and photographs, they showed that eye movements are not random, but constitute an active exploration of the visual image; fixations concentrated on humans and on those regions which were relevant to the task the participants were performing.

The following sections examine the link between information processing, attention, and eye movements in more detail, as it is the basis for the insights that measuring eye movements can provide into processes as diverse as scene perception, reading, task planning and execution, decision making, and language processing (for overviews, see Rayner, 1998; van Gompel, Fischer, Murray, & Hill, 2007).
1.1.1 **Eye movements and visual attention**

If people are uncertain about the identity of an object, they tend to make it the focus both of their cognitive processing (i.e. attend to it), and of their current gaze (fixate it for detailed examination). The frequent depiction of attention as “a spotlight that enhances the efficiency of the detection of events within its beam” (Posner, Snyder, & Davidson, 1980, p. 172) captures this facilitatory influence on perception and recognition. To stick with this comparison, the spotlight’s beam must be moved around the room in order to obtain an adequate representation of the environment. Such shifts of attention can be controlled endogenously to comply with a goal or an expectation, or they can be triggered by exogenous stimuli, such as a sudden movement or noise.

In both cases, the shift of attention will typically be followed by an eye movement towards the object, unless this is deliberately avoided (Posner, 1980; Yantis, 1998; Johansson, Westling, Bäckström, & Flanagan, 2001; Findlay, 2004). An even stronger association between eye movements and attention holds in the opposite direction: a shift of fixation to a stimulus is always preceded by a shift of attention to that stimulus (Deubel & Schneider, 1996; Hoffman, 1998; Irwin, 2004). For this reason, eye movements provide a spatially and temporally accurate window onto attentional processing (though see Irwin, 2004, for a discussion of the limits of their usefulness).

1.1.2 **Gaze control: Determinants of fixation location and duration**

But what determines which of the many spatial locations in the field of view is selected as the target of attention and fixation? Since the very early studies of scene viewing (e.g., Yarbus, 1967), it has been known that people avoid looking at empty or monotonous scene regions, preferring to fixate more varied and therefore presumably more informative regions. Henderson (2003) discusses two different ways in which the informativity of a scene region can be defined: through a bottom-up, stimulus-driven reliance on image statistics, or based on top-down knowledge stored in memory.
The former approach has its roots in the visual-search tradition, which assumes an initial pre-attentive state in which the whole field of view is rapidly processed more or less automatically and in parallel. The result is a division of the scene into objects, based on the perception of basic features such as colour, orientation, and of course location (for an overview, see Wolfe, 1998). Such basic features can be used to compute saliency maps, parsing the scene into background and regions of potential interest (Itti & Koch, 2000; 2001; see also Henderson, 2003; Underwood & Foulsham, 2006; Underwood, Foulsham, van Loon, Humphreys, & Bloyce, 2006). The details of how this might be achieved vary in different models, but the general idea is that scene regions are assigned saliency weights, and the region with the highest weight is fixated. Once it has been processed sufficiently, its saliency is reduced, and attention is reallocated according to an updated weight distribution.

In studies of the perception of natural scenes, computational models based on saliency maps provide useful baseline estimates of fixation probabilities. But although most authors acknowledge the fact that cognitive factors gain in importance as scene-viewing progresses, their models often struggle to incorporate influences from episodic memory about a particular scene, generic schema-inferences, and task-related knowledge. Probably the most striking illustration of the importance of viewers’ knowledge structures for the allocation of attention is the task-specificity of fixation patterns: the correlation of fixation sites with visual saliency is significantly less pronounced during active tasks than during free inspection of an image (Hayhoe, 2000; Land & Hayhoe, 2001; Shinoda, Hayhoe, & Shrivastava, 2001), and even the first fixations into a scene can be affected by the viewer’s goal (Bock, Irwin, Davidson, & Levelt, 2003; Underwood et al., 2006; Castelhano & Henderson, 2007; Glanemann, Zwitserlood, Bölte, Kreysa, & Dobel, 2009; Kreysa, Zwitserlood, Bölte, Glanemann, & Dobel, 2009).

Another finding which shows that visual saliency cannot be the only factor driving gaze is the fact that the gist of a scene can be extracted with surprising speed and accuracy, and this knowledge can then drive subsequent fixations. The term gist has not always been used consistently, but it generally refers to something like a broad label for the type of scene (e.g., kitchen, café, tutorial), including basic semantic or schema-knowledge associated with that situation, and an
approximation of its spatial layout (Henderson & Ferreira, 2004). Evidence from a variety of different sources and paradigms suggests that masked presentations lasting only a few tens of milliseconds can be sufficient to identify and correctly categorise a scene, and to improve subsequent processing (e.g., Potter, 1975; Biedermann, Mezzanotte, & Rabinowitz, 1982; Thorpe, Fize, & Marlot, 1996; Oliva & Schyns, 2000; Davenport & Potter, 2004; Rousselet, Joubert, & Fabre-Thorpe, 2005; Underwood, 2005; Castelhano & Henderson, 2007; Dobel, Gumnior, Bölte, & Zwitserlood, 2007b). However, rapid gist extraction does not usually include details or individual objects (see Tatler, Gilchrist, & Rusted, 2003).

Nonetheless, it is clear that gist is exceedingly useful in selecting the most informative locations for subsequent fixations. This has important consequences for choosing experimental materials as well. To borrow Henderson and Ferreira’s (2004) example, the speed with which participants can carry out an instruction to “point to the toaster” will be influenced by whether gist information is available. For instance, if the toaster is depicted in a typical kitchen scene, gist apprehension will allow them to reduce their search space to the counter tops where toasters are most likely to be found. In contrast, no such heuristic is available if the toaster is just one of nine disconnected objects in an array; in such a context, only the linguistic information, combined with the visual properties of the object itself can be used for identification.

In summary, two general classes of factors are involved in determining gaze locations: salient properties of the image attract fixations in a bottom-up manner, but cognitive knowledge structures also push gaze to potentially interesting locations in a top-down way. Henderson (2007) suggests two alternative ways of integrating the different sources of information, depending on the goal of the viewer. In some cases, for instance when looking for a clock on the wall, knowledge structures can directly modify the saliency map and filter it for relevant features, such as ‘something round and white’. In other situations, it will be necessary to generate an independent context-based map, highlighting regions likely to contain the target (e.g., a pedestrian is likely to be on the pavement but not in the sky). During visual search this contextual map can then be combined with a separate image-based saliency map (Torralba, Oliva, Castelhano, & Henderson, 2006).

Compared to the large number of theories, models, and experiments concerned with where people look in a scene, less is known regarding the duration
of individual fixations to the selected object. This is surprising considering the usefulness and importance of duration measures in reading research, where the time for which a word is fixated serves as an indicator of the relative ease (or difficulty) of processing it (see Rayner, 1998, for more on duration measures in reading).

In scene viewing, fixation durations have been shown to depend on features of the visual stimulus such as contrast, luminance, and size (see Henderson, 2003), but they are clearly affected by cognitive factors as well. For instance, Henderson, Weeks, and Hollingworth (1999) found that fixation durations were influenced by semantic consistency of the scene, while Meyer, Sleiderink, and Levelt (1998) showed increased viewing time when to-be-named objects were hard to identify. They also showed that participants’ gaze tended to shift away from the objects shortly before they were actually named. This finding is reminiscent of results for participants actively performing real tasks, such as preparing a sandwich, making a cup of tea, driving, or sorting blocks (Hayhoe, 2000; Johansson et al., 2001; Land & Hayhoe, 2001; Triesch, Ballard, Hayhoe, & Sullivan, 2003; Hayhoe & Ballard, 2005). In such situations, fixations and motor movements are performed in a consistent sequence: the eyes fixate the next relevant object just after the hands reach it, but before any action is performed on it. Interestingly, they also tend to shift away from the object shortly before the end of the action. In addition, fixations to objects that are not relevant for the current sub-task are extremely rare.

Hayhoe (2000) concludes that vision is based on an active sequence of routines which are used for just-in-time extraction of whatever information is needed for the current task. A series of fixations to the same location can thus serve completely different purposes. For instance, a slice of bread might first be fixated to guide the placement of the jelly, then to direct the knife in spreading it evenly across the whole slice, and finally to check the thickness of the resulting spread. Similarly, the duration of these looks would depend on the task being performed: in the first case, the bread would probably be fixated relatively briefly until the jelly was dropped, but then continuously throughout the spreading action. The final check for thickness might be limited to a single brief look.

Thus, gaze control – directing attention around the visual scene in the service of current processing – affects both fixation location and duration, since extracting
detailed information about an object and what is happening to it usually requires it to be fixated directly (Liversedge & Findlay, 2000). However, fixations may serve other functions as well. On the one hand, Hollingworth and Henderson (2002) argue that visual attention is necessary for spatial and semantic features of objects in a scene to be bound together and committed to short- and long-term memory. Similarly, they claim that refixation can facilitate the subsequent retrieval of this information (cf. Kent & Lambert, 2008). On the other hand, fixations also seem to function as deictic pointers, creating a link between incoming sensory data and the body’s internal coordinate system (Ballard, Hayhoe, Pook, & Rao, 1997). This makes it possible for motor actions to be planned and executed accurately in a moment-by-moment manner, despite simultaneous and independent changes of state, position, or perspective (Hayhoe & Ballard, 2005).

The previous sections have discussed a number of theoretically important connections between eye movements, attention, and cognitive processing. These supply the motivation for recording eye movements in experimental settings, using them as a window onto less overt operations such as action planning, memory retrieval, and language processing, which is the main focus of this thesis. The remainder of this chapter provides a brief theoretical introduction to the areas of language production, comprehension, and dialogue, as well as summarising previous eyetracking research in each field. First however, I briefly address a few practical issues with regard to eyetracking experiments in general, and to psycholinguistic ones in particular.

### 1.1.3 Terminology

So far, I have used fixations, gazes, and looks more or less synonymously. To the extent that I am referring to overt visual attention I continue to do so in the following as well. However, particularly with regard to duration measures, an important distinction must be made between fixations and gazes. A fixation is basically the period between two saccades in which the eye is relatively steady, usually lasting around 300 ms. In contrast, gaze refers to one or more consecutive fixations of the same object, word, or scene region. This measure thus includes any saccades within the region, up to when the first saccade is launched to an outside
target. Gaze durations can therefore range from a few hundred milliseconds to several seconds (e.g., the uninterrupted gaze to the bread while spreading the jelly).

Another term I have used frequently is scene, which is used in many psycholinguistic studies to denote any kind of visual display. However, Henderson and Ferreira (2004) caution against such indiscriminate use, pointing out that several properties of true scenes are missing from the ersatz scenes which usually function as stimuli. According to their classification, true scenes include backgrounds and general clutter, and depict objects on a realistic scale relative to each other. They also respect spatial constraints of gravity and dimension, as well as reflecting colour and light variations determined by the nature of the objects in view. For the rest of this thesis, the term scene will therefore be used to refer only to “a semantically coherent (and often nameable) human-scaled view of a real-world environment comprising background elements and multiple discrete objects arranged in a spatially licensed manner” (Henderson & Hollingworth, 1999, p. 244).

In contrast, any display that does not meet these criteria will be termed an ersatz scene. One important reason for distinguishing between true scenes and other forms of experimental display is evidence that a distinct brain area, the parahippocampal place area, is activated by natural scenes (Epstein & Kanwisher, 1998; Epstein, Harris, Stanley, & Kanwisher, 1999). This specialised processing may explain the superior performance in rapid categorisation tasks found for these scenes, compared to artificial stimuli (Braun, 2003; Underwood, Jebbett, & Roberts, 2004; Fei-Fei, VanRullen, Koch, & Perona, 2005; Rousselet et al., 2005). In addition, the possibility that different cortical mechanisms are involved in scene processing depending on the type of stimulus means that caution is required when generalising findings between studies and deriving general predictions (for a similar argument with regard to face perception, see Kingstone, Smilek, Ristic, Friesen, & Eastwood, 2003).

1.1.4 The visual-world paradigm

The distinction between true and ersatz scenes is particularly pertinent to what has become “the most commonly used paradigm in language-vision research”
(Henderson & Ferreira, 2004, p. 11), since studies using the so-called visual-world paradigm are often based on visually sparse displays. As will become evident in the following sections, the paradigm is popular in both comprehension and production studies, and it has been employed for dialogue research as well. Therefore, this section reviews the general idea behind the paradigm and the different forms it can take, before detailed results from all three fields are reported in the remainder of the chapter.

The first description of a visual-world paradigm experiment is a paper by Cooper (1974), who made use of corneal reflections to record participants’ eye movements while they listened to a story. Critically, the story contained both direct and implied references to a set of objects depicted as line drawings in an array in front of participants. The eye movement recordings confirmed the hypothesis that:

> When people are simultaneously presented with spoken language and a visual field containing elements semantically related to the informative items of speech, they tend to spontaneously direct their line of sight to those elements which are most closely related to the meaning of the language currently heard. (Cooper, 1974, p. 85)

Despite Cooper’s efforts to promote the usefulness of such a paradigm to explore language processing in real-time and without interrupting the flow of speech, it was practically ignored in the following 20 years until a groundbreaking article by Tanenhaus and colleagues (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995; see also Spivey, Tanenhaus, Eberhard, & Sedivy, 2002). In their setup, eye movements were tracked while participants followed instructions regarding a set of objects on the tabletop in front of them. In a one-referent condition, participants might see an apple on a towel, a pencil, an empty towel, and a box; while in the two-referent condition, they were confronted with the apple on a towel, another apple on a napkin, the empty towel, and the box. In both conditions, participants heard either a temporally ambiguous instruction, as in (1), or the disambiguated version in (2):

1. *Put the apple on the towel in the box.*
2. *Put the apple that’s on the towel in the box.*

Tanenhaus et al. compared the rate of fixations to the correct goal (the box) and to the competitor goal (the empty towel), as well as the latencies until they were
first fixated. Like Cooper, they found that fixations to the relevant object were time-
locked to the sentence participants were hearing, but also that the visual context
immediately affected how the linguistic input was interpreted. Interestingly,
participants were much more likely to look at the competitor in the one-referent
condition, interpreting “on the towel” as a destination rather than a modifier. In
contrast, in the two-referent context they initially looked at both apples equally,
trying to establish the correct referent, but “on the towel” was consistently
interpreted as a modifier, not as a destination.

These findings initiated a spate of subsequent studies on language
comprehension, all of which use fixation behaviour in a pre-defined visual context
as an online indicator of how the accompanying sentence is being interpreted.
Several variations of the paradigm are common: participants’ task can be to
manipulate real objects (e.g., Chambers, Tanenhaus, & Magnuson, 2004), to click on
a depiction of an object on the screen (Dahan, Magnuson, Tanenhaus, & Hogan,
2001b), or just to ‘listen and look’ (Huettig & McQueen, 2007). Visual displays
generally consist either of real objects (Hanna & Tanenhaus, 2004), of clip-
art pictures arranged to resemble a scene (Kamide, Altmann, & Haywood, 2003a), of
unconnected pictures displayed in a grid form (Boland, 2005), or else they display
an animated event (Knoeferle & Crocker, 2007; Papafragou, Hulbert, & Trueswell,
2008), or only a blank screen (Altmann, 2004). Finally, the accompanying sentences
can be designed to test many different aspects of linguistic processing, including the
resolution of syntactic and semantic ambiguities, thematic role assignment, and the
prediction of subsequent sentence components.

The idea that participants’ gaze can provide information on their current
language processing has been exported from comprehension to studies on
production (e.g., Meyer et al., 1998; Griffin & Bock, 2000) and dialogue (Brown-
Schmidt, Campana, & Tanenhaus, 2005; Brown-Schmidt & Tanenhaus, 2008). Many
of these are discussed in detail in the later sections. Critically however, any study
using the visual-world paradigm relies on the assumption of a strong link between
fixating an object and the internal processing of its referent. To some extent, the
prolific use of the paradigm and the many valuable and replicable insights gained
from it justify this assumption, particularly when experiments are designed in such
a way that fixation behaviour is compared between conditions which are identical
in all respects apart from the variable of interest. On the other hand, it must be borne in mind that the primary function of fixations is to extract more detailed information from an object in the field of view. For this reason, fixation patterns can always be influenced by bottom-up factors, such as visually salient aspects of the image (e.g., a particularly bright object), external events (the sudden appearance of a reflection on the screen), or by objects which are hard to identify correctly. As described above, top-down processes which are not connected to language use also influence gaze, and one important factor to bear in mind is the suggestiveness of the design: presenting a limited number of objects together with a sentence referring to them contains an implicit invitation to look at them, and makes it likely that participants will guess that this is what the experimenter would like them to do. These and other caveats are discussed by Tanenhaus, Magnuson, Dahan, and Chambers (2000), together with an appeal for researchers to formulate clear and testable linking hypotheses about the way that the measured eye movements might be affected by the experimental manipulation (see also Tanenhaus, 2007).

I return to some of these issues when discussing the individual experiments. For now, it is time to look at the different areas of language use in a bit more detail.

1.2 Language Production

Speech, as the ability to acquire and use a culturally specified inventory of sounds and rules to communicate our thoughts, serves many functions in everyday life: using different combinations of words, we can describe events and situations to others, communicate ideas, plans and emotions, sustain social relationships, warn of dangers, give orders, directions and requests, as well as enjoy language in the form of literature, jokes, poetry, and song.

Yet precisely this multitude of potential uses and topics has proved a challenge for the empirical study of language production, as it is hard to limit participants to producing a particular word or structure without affecting the processes with which they generate this expression (cf. Bock, 1996). Early insights into speech production were therefore based mainly on observational studies of impaired production, either from patients with localised brain damage, or as
spontaneously occurring errors in normal speech. Findings from these research traditions play a significant part in what we know about language production today, and continue to provide the background for many current theories and investigations (e.g., for a computational model linking the two traditions, see Dell, Schwartz, Martin, Saffran, & Gagnon, 1997).

However, as this thesis is experimental in nature, the following overview of how people produce sentences focuses primarily on what Levelt (1999a, p. 224) calls the *chronometric* tradition. This term alludes to the fact that most experimental paradigms in production involve comparisons of naming latencies between conditions, as in implicit priming or picture-word interference tasks. Only recently have these been supplemented by eyetracking (see in particular Sections 1.2.3 and 1.2.4 below) and brain imaging methods (Indefrey & Levelt, 2000). In the following section, I present a general model of the processes involved in generating speech (Section 1.2.1) and summarise empirical findings relating to the production of individual object names (1.2.2 and 1.2.3). Section 1.2.4 extends these results to the more complex task of formulating an entire sentence, and the usefulness of eyetracking paradigms for investigating it. Among other things, they have enabled researchers to address two important issues in sentence production, which I discuss in a little more detail: the factors which determine the starting point of a sentence (1.2.5) and the assumption of incrementality, which is a key feature of most models of production (1.2.6). Finally, a brief summary is provided in Section 1.2.7.

1.2.1 How to get from thoughts to articulation: A model of speech production

The most comprehensive theoretical account of what needs to happen before a speaker can speak was provided by Levelt (1989). The processes described in the following are therefore loosely based on his *blueprint of the speaker*, concentrating on those aspects which are relatively uncontroversial (see also Bock & Levelt, 1994; Levelt, 1999b; Levelt, Roelofs, & Meyer, 1999). Most versions of this and similar models assume the following stages of processing: conceptual preparation, grammatical encoding, morpho-phonological encoding, phonetic encoding, and articulation (the names and exact functions of components can vary). Suggestions as
to how some of these could be realised at the level of neuronal assemblies come from Pulvermüller and colleagues (e.g., Pulvermüller, 1999; 2000).

A few general properties of the model should be mentioned before looking at the components individually. Most importantly, it is both cascading and incremental. This means that the process of speech production always takes the same route in the same direction, beginning with conceptualisation and ending with articulation. Each processing component works autonomously, receiving its input from the component before it, and delivering specific output to the next one. Incremental processing is also usually taken to imply that later components can begin work before earlier processes have been fully completed: as soon as a fragment of suitable data is available, the next component can begin working in parallel. Finally, all the components are subject to more or less intensive error-monitoring via the comprehension system.

These three features – encapsulated processing, incrementality, and continuous self-monitoring – together ensure that meaningful speech acts are generally produced fluently and almost without unnecessary pauses or errors (Levelt, 1989). The whole process begins with a preverbal intention on the part of the speaker to communicate a piece of information. During conceptual preparation, procedural and declarative knowledge are both used to generate a preverbal message. This is a semantic representation of what is to be expressed and consists of lexical concepts selected to suit the present discourse situation. To use an example based on the experiment described in Chapter 2, imagine that a speaker wishes to describe a picture showing a clown chasing a policeman. The preverbal message would consist of these two concepts and the notion that both are running in the same direction, but also the speaker’s interpretation of the whole event: is the clown chasing or just following the policeman, or is the latter actually leading the clown? Which of the two characters is the listener more likely to be interested in?

Grammatical encoding can begin as soon as any of these details has been established. The first step is to select suitable lemmas to express the preverbal concepts. In the mental lexicon, each lexical concept is linked to one or more lemmas containing syntactic information about how to use the relevant word. For instance, the policeman-concept might be linked to the three competing lemmas policeman, cop, and bobby. All three identify themselves as a singular, masculine noun, but they
differ (among other things) in register. Hence, activation spreading from the message level makes it likely that policeman is selected over the other two, as the ‘official’ and thus most suitable term of reference in this context. Verb lemmas also specify the arguments they require, e.g., chase takes two animate arguments, a chaser and a chasee. Consequently, its activation makes it possible to assign grammatical functions: will the verb be in active or passive voice, and how are the roles of agent and patient allocated to the two characters? The result of grammatical encoding is a surface-structural representation, spelling out the left-right order of all the selected lemmas.

The process of morpho-phonological encoding uses this phrasal order to assemble the correct word forms (e.g., clown is preceded by a definite article but receives no plural marking; chase in active present progressive requires an is auxiliary and an ing-suffix). It also generates the syllabic and prosodic structure of the sentence. Phonetic encoding transforms this into an articulatory score, which is finally executed in overt articulation.

1.2.2 Empirical findings for the production of single words

Although Levelt (1989) explicitly states that his model is concerned with speech in the context of the sentence, the discourse, and the whole social interaction, most empirical tests of the theory have involved the naming of single objects. One important finding is that when speakers are asked to name pictures, the whole process from recognition to articulation takes between 600 and 1200 ms (Levelt, 1989, p. 222). Several factors operating at different levels of the model influence this speech-onset latency: for instance, material which is visually or semantically similar to the to-be-named object interferes with lexical selection, leading to longer response times (Schriefers, Meyer, & Levelt, 1990). Conversely, morphological and phonological similarity between target and distractor facilitates word-form retrieval and results in reduced naming latencies (Meyer & Schriefers, 1991), as does a high frequency target word (Jescheniak & Levelt, 1994). The order and timing of the

---

2 I have kept the description of the three later components brief, as they are less relevant for this thesis. More detail can be found in Levelt (1989; 1999b).
postulated processes leading to language production has been examined by varying the temporal relationship between target and distractor presentation (Glaser & Düngelhoff, 1984; Schriefers et al., 1990; see Bock, 1996, for a discussion of picture naming methods).

Despite the clever design of many of these studies, speech-onset latencies record only one single moment in time, the endpoint of all the cognitive processes which take place before speech initiation. For this reason, online measures such as eye movements and event-related potentials (e.g., van Turennout, Hagoort, & Brown, 1997; 1998) provide valuable additions to the classical picture-word interference studies of language production.

1.2.3 **Eye movements during the production of object names**

Recording eye movements during production has only become practical fairly recently, with the development of eye trackers which tolerate movements of the mouth and head. In one of the first investigations into speech production using eyetracking, Meyer, Sleiderink, and Levelt (1998) compared participants’ eye movements while naming two objects (e.g., “scooter and hat”) to those produced in a categorisation task of similar object pairs. Critically, they varied both a linguistic and a perceptual factor: the lexical frequency of the object names and the contour visibility of the image. As would be expected based on the findings reported above, frequency substantially affected naming latencies. It also affected viewing times in the production task, but not during categorisation, suggesting that viewing time depends on the time required for word form retrieval. Subsequent studies have corroborated and extended this finding, showing that gaze durations during speech are affected by very similar factors as speech-onset latencies: recognisability of the objects, codability and frequency of their names (Griffin, 2001), phonological priming (Meyer & van der Meulen, 2000), and name length (Zelinsky & Murphy, 2000; Meyer, Roelofs, & Levelt, 2003). It also seems that when people generate object names, they continue looking at the objects at least until they have retrieved the phonological form and probably also some of the articulatory commands (Griffin, 2001; Meyer & Lethaus, 2004).
Virtually all naming studies using eyetracking present participants with at least two to-be-named object pictures simultaneously (e.g., Meyer & van der Meulen, 2000; van der Meulen, Meyer, & Levelt, 2001; Meyer et al., 2003; Morgan & Meyer, 2005), in contrast to the single-image presentations in the picture-word interference paradigm. Overwhelmingly, speakers fixate each object in the order in which it is going to be mentioned, and look away (usually to the next object) shortly before the name onset (see Griffin, 2004, for an illustration of the respective time courses of gaze and speech). This pattern of gazes has been interpreted to reflect the incrementality of speech production:

Thus speakers generate noun-phrase conjunctions such as 'the cat and the dog' in a highly sequential fashion, starting to look at the second object just in time to be able to generate fluent speech, but not earlier. (Meyer & Lethaus, 2004, p. 197)

To avoid interrupting the speech stream, the encoding of the next unit must begin before the articulation of the preceding one has been completed, and it is these conceptualisation and lexical selection processes that the speaker’s fixation pattern is assumed to reflect (Meyer & Lethaus, 2004). However, two findings qualify the usefulness of this link between fixation and conceptualisation for research on language production. Firstly, Morgan and Meyer (2005) show that processing of an object that is about to be named does not begin at the onset of its fixation, but can start well before a saccade is made towards it. This means that gaze duration is not directly indicative of the time required for certain aspects of processing to be completed. Secondly, even when speakers don’t fixate an object, they still often name it correctly (van der Meulen et al., 2001; Bock et al., 2003; Meyer, van der Meulen, & Brooks, 2004; Dobel et al., 2007b). Yet if name-related fixations are not actually necessary, why do speakers usually look at the objects they are going to mention? In fact, not only do they fixate them, they even exhibit a strong temporal coupling between gaze and speech: the eye-voice span, i.e. the time from first fixating to first mentioning an object, is generally about 600-1000ms (Griffin & Bock, 2000; Meyer & Lethaus, 2004).

Plausible explanations for speakers’ propensity to make name-related gazes include the usefulness of gaze cues in social interactions, the idea that gazes help speakers to sequence the words correctly, or that they are made to avoid interference by other objects. However, Griffin (2004) finds no conclusive empirical
support for any of these theories. Alternatively, she suggests that name-related eye movements may merely be an automatic consequence of attention, reflecting a link between conceptual and spatial information about the object (see also Spivey, Richardson, & Fitneva, 2004; Altmann & Kamide, 2007; F. Ferreira, Apel, & Henderson, 2008; Hoover & Richardson, 2008). Thus, an activated element of the preverbal message might automatically trigger the spatial index associated with it, and this in turn would make it likely that an eye movement would be launched to that location. Following this explanation, the main reason for making name-related fixations might be that it is effortful to suppress them. In support of this idea, people usually look at to-be-mentioned entities even when they already know their names (van der Meulen et al., 2001) or when they are repeating each word verbatim (Griffin & Garston, 2003). So, despite the uncertainty regarding the function of name-related gazes, the fact that speakers have a strong tendency to make them means that their eye movements contain useful information on the sequence of processes leading up to articulation: “The choreography of the eyes and the mouth during speaking is a powerful means for investigating the relationship between perception and language” (Bock et al., 2003, p. 681).

Eyetracking is particularly useful for exploring the formulation of longer and more complex utterances than the production of object names discussed so far. Certainly, sentence production has been studied using speech-onset latencies as well (e.g., Lindsley, 1975; Schriefers, Teruel, & Meinshausen, 1998). But the insights to be gained by comparing single response measurements are limited, since the generation of a sentence must necessarily involve a whole series of processing stages with different factors playing a part at different times. In addition, sentence production includes some aspects of language processing which are fairly peripheral to picture naming: the issue of starting points (i.e. deciding what to speak about), the extent and limits of incrementality in the production system, and the generation of syntax. Sections 1.2.5 and 1.2.6 examine these aspects in a little more detail. First however, it is important to establish to what extent findings from eye movements recorded during picture naming can generalise to descriptions of the more detailed displays used to elicit sentences.
1.2.4 Eye movements during the production of complex utterances

In tasks where participants make more complex syntactic decisions than in picture naming, the tight coupling of fixating and pronouncing is, at first glance, less obvious. Examples for such experimental set-ups are naming a spatial array of objects (Meyer et al., 2004), the description of action scenes (Griffin & Bock, 2000), or the task of telling the time (Bock et al., 2003).

In such cases, it has been suggested that a brief appraisal or apprehension phase precedes the main process of formulation (Griffin & Bock, 2000; Bock et al., 2003; Meyer & Dobel, 2004). This proposal is based on the observation that within about 300 ms of display onset, speakers are already highly likely to be looking at the entity they are going to refer to first (Griffin & Bock, 2000; Bock et al., 2003). Thus it seems that speakers can often extract the information necessary to begin formulating a task-conform sentence without re-fixating the display at all, reminiscent of the rapid extraction of gist in scene perception (described in section 1.1.2 above). In addition, two studies which required participants to generate more complex utterances reported a similar and probably related phenomenon: van der Meulen (2001) and Meyer and Dobel (2004) both found an initial preview fixation into the display. In the former case, the task was to describe the spatial relationship of four objects, starting at the top; participants tended to begin trials with a fixation to the two bottom objects, as the sentence structure of the description depended on whether these were identical or different. Similarly, Meyer and Dobel had participants describe ditransitive events using either double-object or prepositional datives and found an overwhelming tendency for participants to fixate the theme object on either the first or the second fixation. A follow-up experiment, where participants named the agent, theme, and recipient in a list-like fashion, found no such initial preview fixation to the theme, suggesting that it was indeed linked with sentence production and served to “encode the action and select the verb” (Meyer & Dobel, 2004, p. 268).

Whether the apprehension phase is accompanied by a preview of the informative region or not, it is consistently followed by a slower, more incremental formulation phase, in which the order of fixation predicts the order of mention,
largely irrespective of the type of description produced. For instance, Griffin and Bock (2000) asked their participants to describe transitive events depicted in black-and-white line drawings, and compared their eye movements during the production of different sentence structures (e.g., *The mailman is being chased by the dog* vs. *The mailman is chasing the dog*). In contrast, participants in Bock et al.’s study (2003; see also Bock, Irwin, & Davidson, 2004) were eyetracked while they told the time from digital or analogue clock displays, using either absolute (e.g., *three-fifteen*) or relative expressions (e.g., *quarter past three*). Despite the very different tasks, both studies found that speakers tended to look at whatever they were about to mention approximately one second before they did so, and this eye-voice span was relatively constant across the position of mention in the utterance. Thus it seems fair to assume that gazes during the formulation phase in sentence production are linked to phonological retrieval in the same way as they are in naming studies (Meyer et al., 1998; Griffin, 2001).

The following section revisits the idea of an initial apprehension phase, which serves to extract the conceptual structure of what is to be described. This is of interest for studies of language production because this first impression may have consequences for the perspective and syntax of the resulting sentence.

### 1.2.5 Conceptual preparation and starting points

One issue that is rarely discussed in detail in models of language production is the basis for the preverbal message: what determines “the impetus for a production episode” (Bock et al., 2004, p. 273); which aspects of perception and thought processes does a speaker decide to communicate, and which of these are deemed most important? This section will discuss some proposals regarding such **starting points**.

If the first element of the sentence – usually the subject – is selected during the very brief apprehension phase mentioned above, it seems probable that particularly prominent concepts grab this position. Indeed, a wealth of primarily linguistic research suggests that sentential subjects are highly likely to be conceptually accessible: concrete, human (or at least animate), and the source or
cause of a described event (Bock & Warren, 1985; Branigan, Pickering, & Tanaka, 2008). Other features which have been discussed in connection with role assignment are perceptual salience and spatial properties (Tomlin, 1997; Chatterjee, 2001), discourse status (Prat-Sala & Branigan, 2000), structural incentives (Bock et al., 2004), and cultural, language-specific, or even individual preferences (Dobel, Diesendruck, & Bölte, 2007a).

Bock et al. (2004) provide a comprehensive discussion of theories on starting point selection. They point out that, at least in English and most European languages, the designation of a particular entity as the agent of an event and its assignment to subject status is usually confounded with its selection as the first sentence component. This can make it difficult to assess the relative importance of these properties for each of these roles. Again, eyetracking can be useful here, as it provides a direct measure of which aspects the speaker attends to before initiating a sentence.

In this context, Bock et al. highlight an interesting contrast between the factors known to influence eye movements during free viewing and those which have been found to affect word order: as we have already seen, perceptually salient features, such as abrupt onsets, colour, and visual complexity, are most likely to attract fixations during scene viewing, particularly on the initial fixations into a display. In contrast, while there is some evidence that perceptual factors can influence word order too (Flores d’Arcais, 1975, described in Levelt, 1989, p. 262; Tomlin, 1997), their effects are far weaker than conceptual influences (McDonald, Bock, & Kelly, 1993; Prat-Sala & Branigan, 2000; Papafragou et al., 2008) – in general, it seems that perceptual salience is not very important in the decision to mention an entity first (Griffin & Bock, 2000). This seems at odds with the fact that name-related gazes generally begin very early, within one or two fixations into the display, while at the same time first fixations must at least partly be driven by visual salience. Consequently, and based on a lack of correlation between the very first fixation location and the first sentence component in Griffin and Bock (2000) as well as in their own time-telling studies, Bock and colleagues argue that sentence structure is generally decided before the first name-related gaze. In this way, the apprehension phase may serve to establish the location of the entity which the speaker already plans to mention first:
Another way of putting this is that people begin with what it is they have something to say about, and they tend to avoid initiating utterances until they are fairly sure what the beginning of that something is. This implies that the rudiments of an utterance frame are present when an utterance is initiated. (Bock et al., 2004, pp. 270-271; see also Papafragou et al., 2008)

An alternative account states that visual attention can indeed affect word order and sentence structure, but that the manipulations used in previous studies were too weak to find such effects. Gleitman, January, Nappa, and Trueswell (2007) used a subliminal attention-capture manipulation just before display onset to pull participants’ attention to a particular depicted entity. The events participants had to describe were selected so as to allow semantically equivalent word-order variations in a variety of different constructions: the active/passive alternation, perspective predicates (e.g., the dog is chasing the man vs. the man is fleeing from the dog), symmetrical predicates (the policeman is meeting the construction worker vs. the construction worker is meeting the policeman), and conjoined noun phrases (the dog and the cat... vs. the cat and the dog are growling). Irrespective of the type of utterance participants were preparing, the attentional cue had a reliable effect and significantly predicted that the highlighted entity would be named first. In fact, even on trials where no cue attracted the speaker’s attention, the very first fixation was reliably related to the upcoming sentence structure. Based on these results, Gleitman and colleagues (2007) argue that apprehension and formulation occur in parallel: if attention is drawn to the cat in the example above (for instance because it appears on the left of the screen), this will lead to increased activation of the corresponding lemma, and in turn of its phonological form. Activated forms are likely to be slotted into early sentence positions (Bock, 1986a; Levelt, 1989), in this case potentially while the action and the dog are still in the process of being apprehended.

Both accounts have a certain plausibility, and the issue of the effect of visual attention on linguistic starting points is still far from resolved (for a response to Gleitman et al., 2007, see Kuchinsky, Bock, & Irwin, 2009). For the purpose of this thesis, Gleitman and colleagues’ account takes us back to the question of how incremental the language production system really is, and which processes can be assumed to be taking place in parallel.
1.2.6 The assumption of incrementality in production

Assuming that a suitable starting point has somehow been established, how does the speaker continue formulating the sentence, and how much more processing is required before the first component can be uttered? It seems obvious that speakers usually begin articulating the first word of a sentence before the rest has been fully prepared (Levelt, 1989). This makes sense in view of the fact that utterances have to be constructed on the fly and produced reasonably fluently to avoid losing the floor (Clark, 1996). At the same time, some concept of the intended end-result must be present before speech onset to ensure factors like appropriate case agreement and intonation.

The debate surrounding incrementality therefore focuses not so much on its existence, but on the amount of parallel processing and on which operations can be conducted simultaneously. For instance, as F. Ferreira and Swets (2002) point out, the very fact that several processes can operate at the same time (e.g., lemma selection and morpho-phonological encoding) implies that each individual piece of information must be handled fairly serially (e.g., phonological encoding of the word *sheep* cannot occur before its lemma is selected). Thus, several words can be processed in parallel at different levels, but earlier words must go through the whole process before later ones can.

In the following sub-sections, I discuss two connected aspects of incremental processing: the amount of processing occurring before sentence onset (1.2.6.1), and the issue of the encoding of the verb (1.2.6.2).

1.2.6.1 How much of the sentence must be prepared at its onset?

Different theories have been put forward regarding the processes which have to be completed before speech onset, agreeing only on the fact that phonological encoding of later sentence components occurs during speech (Meyer, 1996; Smith & Wheeldon, 1999; for a brief summary see Griffin, 2001). In an eyetracking study similar to the ones described above, Griffin (2001) asked speakers to describe three objects pictured on the screen, using the sentence frame “the A and the B are above the C”. The frequencies of the objects’ names and the number of alternatives (their codability) were varied systematically; both factors are known to affect naming
latencies. Indeed, speakers did spend more time looking at objects with low codability and low frequency names before naming them. Importantly however, speech onset (i.e. naming object A) was not affected by codability and frequency of objects B and C. It seems that speakers processed each object separately and in turn, naming A as soon as they had prepared a name for it, and before moving on to B and C (cf. Clark & Wasow, 1998).

Taking the opposite perspective and examining the factors that do influence eye-voice spans at each point during the formulation of an utterance, Bock and colleagues (2003; 2004) reached a different conclusion. They distinguish between linear and horizontal incrementality: linear incrementality assumes processing in a word-by-word or phrase-by-phrase fashion, such that each eye-voice span depends only on the complexity of the following unit of speech, as in Griffin (2001). In contrast, hierarchical incrementality would allow for additional processes to take place in parallel, such as the partial preparation of later components or the creation of some kind of sentence frame. Under hierarchical incrementality, the eye-voice span for the sentence-initial component could therefore be longer than later spans. In their time-telling task, Bock et al. (2003) found that while both the type of clock and the complexity of the expression affected eye-voice spans consistently throughout the utterance, the first eye-voice span was generally longer. In addition, it depended on the compatibility between the clock and the expression that was being prepared: eye-voice spans were reduced for combinations of a digital display with an absolute time, as well as for an analogue clock with a relative expression.

Bock and colleagues conclude that the language production system is characterised by hierarchical incrementality, suggesting that some form of ‘utterance groundbreaking process’ may be required before regular sentence formulation can begin. In a reversal of the wrap-up effects found in reading studies, which are thought to reflect the integration of the sentence content into the context (Just & Carpenter, 1980), they propose that in production such a groundbreaking process might serve to separate the individual message components out of the global context in order to retrieve suitable word forms.
1.2.6.2 How soon is the verb encoded?

Since different verbs allow for or even require different numbers and types of arguments, one potential form of groundbreaking critical to the selection of sentence structure might be the early encoding of the verb (for a theoretical model of syntactic production stressing verb importance, see F. Ferreira, 2000). In declarative English sentences, the conjugated verb obligatorily goes in second position, following the subject. Thus, there are two reasons which make it particularly interesting to determine how soon the verb is selected and encoded: on the one hand, the finding that all verb-related processing occurs after the onset of the subject noun phrase would imply a very strong sequentiality in sentence formulation, since the time available for preparing the verb would be limited to the time it takes to utter the subject. In Bock et al.’s (2003) terminology, this would be evidence of extreme linear incrementality. On the other hand, the selection of a particular verb and of the argument structure it requires provides a basic plan for the rest of the sentence. It would therefore be desirable to accomplish this as soon as possible, since such a plan could facilitate subsequent processing (F. Ferreira, 2000). Meyer and Dobel (2004) argue that this may be the motivation for the initial preview fixation they found in the study described above: participants fixated a region critical to the verb before the first name-related gaze to the agent.

A number of studies have investigated the time-course of verb encoding during sentence production, using a variety of different paradigms. For instance, Lindsley (1975) measured speech-onset latencies for four tasks: naming the agent of a transitive event (S), the action (V), the agent and the action (SV), or producing a whole active SVO sentence. The identity of the agent was either varied or kept constant across trials. In conditions where the subject was already known, speech onset latencies were longer for naming actions (e.g., “greeting”) than for producing whole SV- or SVO-sentences. This could be interpreted as evidence of linear incremental processing, as the planning of the verb clearly took place while the subject was being pronounced. When the subject was not known however, latencies for SV-utterances (“the man is greeting”) were longer than for naming the agent (“the man”), showing that some aspects of verb selection also occurred before the subject was mentioned. Certainly, both results show that the verb plays a special role in encoding a sentence.
Schriefers, Teruel, and Meinshausen (1998) used a version of the picture-word interference task to investigate the extent of advance grammatical planning for transitive and intransitive sentences. Speakers heard or saw semantically related or unrelated distractor verbs while they described a target picture. Importantly, the experiment was conducted in German, which made it possible to compare SOV- and VSO-sentences. No semantic interference was found for intransitive or for SOV-sentences, but speech onset of verb-initial transitive sentences was significantly delayed by semantically related distractors. Schriefers et al. conclude that verb-planning is not obligatory before speech onset. Instead, they propose that grammatical roles can be assigned in two different ways: either based on the verb’s syntactic stipulations, or through conceptual features of the to-be-described event, such as animacy or agency (see also Bock & Levelt, 1994). Which route is selected might depend on factors such as the order in which the input becomes available, and the extent to which a particular syntactic structure is pre-activated (Schriefers et al., 1998).

### 1.2.6.3 Strategic incrementality

The idea that the extent of incremental processing could be variable has been put forward elsewhere too. In a study comparing eye movements during reading and naming of complex number terms (house numbers and clock times shown either as Arabic figures or written words), Korvorst, Roelofs, and Levelt (2006) found that the amount of pre-speech planning (evidenced in gaze durations) depended critically on the utterance structure. Meyer, van der Meulen, and Brooks (2004) had participants produce descriptions such as “the chair next to the cross is brown”, where the colour of the first object was named only after the second object. For some participants, the first object or its colour was removed from the display while they were looking at the second object. Even when the colour information was no longer available, speakers usually refixated the first object region before producing the colour adjective. What differed between conditions was the time spent on these refixations; viewing times were shortest when the object had vanished, and longest when it remained unchanged. Under the assumption that viewing time for an object depends on the time required for lexical access (Griffin, 2001), this suggests that speakers who knew the colour information was transitory...
prepared the colour name beforehand, whereas those for whom it remained constant relied on being able to encode it later.

It seems then that speakers are able to decide how much incremental processing is necessary: in some situations they begin to speak as soon as the first phonological word has been formulated, preparing the rest of the utterance in parallel with their speech. At other times however, they may prepare most of the utterance before beginning articulation – presumably when they can allow themselves to relax, or conversely if they feel the need for more cautious processing. Thus, incrementality is not a hard-wired property of the language production system but is situation-dependent and subject to strategic control by the speaker (cf. F. Ferreira & Swets, 2002). However, this account also suggests that incremental processing requires effort, and is therefore avoided whenever possible:

The finding that speakers are in principle capable of speaking incrementally without any cost in accuracy suggests that, in fact, speakers prefer to plan more carefully than they absolutely need to. Indeed, ... people are actually more efficient when they speak incrementally: They accomplish more in less time. For whatever reason, however, most people are not inclined to speak in this manner. (F. Ferreira & Swets, 2002, p. 77)

Such a view of the traditional assumption of incrementality in the language production system raises new questions: rather than asking which production processes have to be completed before initiating a sentence and which can take place in parallel, it becomes important to look at tasks and situations which make speakers more or less likely to produce incrementally.

1.2.7 Summary: Language production

In short, sentence production describes the transformation of a speaker’s preverbal message into a corresponding sentence which can then be articulated. Each aspect of the preverbal message proceeds through a sequence of encoding steps, ensuring that contextually suitable lexical expressions are selected, slotted into a grammatically correct utterance, and verbalised comprehensibly. How much of this processing can take place in parallel seems to depend on the complexity of the utterance and the speaker’s preferences, and can be adapted to the current situation. This makes sense in view of the fact that even the naming of individual
objects has been shown to be affected by a number of factors, such as co-activated semantically related concepts, morpho-phonological similarities, and the frequency of object names.

Similar factors can also affect the time spent looking at an object before naming it – in fact, the relationship between looking and naming is so tight that viewing times can be used as indicators of word form retrieval. Although a completely satisfactory explanation for the tendency to gaze at to-be-mentioned entities is still missing, essentially the same pattern of eye movements has been found during the production of entire sentences as in single-object naming: speakers consistently look at the object they are next going to mention until just before they actually do. In fact, sentence production is an area where studies using eyetracking are particularly informative, as the speaker’s focus of attention can be followed both before and throughout the utterance, throwing light on the factors which affect linguistic phenomena such as starting points and sentence structure at different points in time.

1.3 Language Comprehension

Sentences are mostly produced to be understood by someone else, so we now turn to the logical complement of language production: comprehension. Comprehension occurs both in the auditory and in the visual modality, so I begin by reviewing some similarities and differences between listening and reading (Section 1.3.1), before presenting a basic model of the whole comprehension process (1.3.2). For the experiments presented later in this thesis, the way listeners understand sentences and longer text passages is of greater interest than how they retrieve the meaning of individual words. Core topics are therefore the general question of how sentences are parsed (1.3.3), and the ways in which the linguistic and visual context can influence this analysis (1.3.4). As in the review of language production, I focus particularly on how studies using eyetracking in a visual-world context have been able to contribute to the investigation of these areas, and provide a brief summary in the final section (1.3.5).
1.3.1 **Critical features of auditory comprehension**

At least from a monologue viewpoint, the listener’s task in comprehending a sentence is to reverse all the operations the speaker used to transform their thoughts into an acoustic signal: the signal must be mapped back onto words in the mental lexicon, so that the relationship between the underlying concepts can be extracted and the speaker’s intended meaning understood. Practically however, there are several differences between the speaker’s and the listener’s task. The sequentiality of the input is perhaps the most important: as we have seen, speakers can plan ahead and prepare several speech units simultaneously, even if they only utter one word at a time. In contrast, a listener has to work with those words which have already been spoken.

To a large extent, this sequentiality exists for reading too, since the eyes tend to move from left to right, and from one word to the next (see Rayner, 1998, for a review of reading research; and Cutler & Clifton, 1999, for a comparison of the auditory and visual modalities in comprehension). This similarity between reading and listening is important, as most research on sentence comprehension has been based on reading (e.g., exploring how local ambiguities affect fixation durations to the ambiguous and disambiguating regions of a sentence). Such studies of visual comprehension can use self-paced reading paradigms, or the eye movements can be recorded during free reading or with a moving window. Tasks more suited to exploring auditory comprehension are various forms of shadowing or monitoring (e.g., for a particular phoneme), as well as priming and sentence probes. Of course, eyetracking can be used for studying auditory sentence comprehension too (e.g., Tanenhaus et al., 1995), as can ERPs (Van Berkum, 2008).

Overall, most findings from reading research seem to generalise quite well to auditory comprehension, but some critical differences of the two signals should be borne in mind (Cutler & Clifton, 1999; see also Spivey et al., 2002; Snedeker & Trueswell, 2005). One is that words are not separated by spaces as they would be on a page, making it harder to segment the speech signal. On the other hand, the speaker’s prosody may be helpful for this task, but it is missing in written text. Also, printed material allows the reader to look back at difficult sections, while spoken sentences are transient – once they have been spoken, the only way to recover what
was said is to ask the speaker to repeat the utterance. At the same time, auditory comprehension frees up the listener’s eyes to look for clues to the speaker’s meaning in the world surrounding them. I will return to the issue of differences between reading and listening in Section 1.3.4.

Both the sequentiality of the input and the transience of the stimulus require auditory sentence comprehension to work in a strongly incremental fashion (Marslen-Wilson & Welsh, 1978). Processing of each fragment begins the moment it is heard, integrating it with provisional representations of the utterance structure and content as soon as possible (see Pickering, Clifton, & Crocker, 2000b). Although there may be situations where this incrementality is limited (e.g., the processing of syntactic ambiguities or anaphora resolution), different paradigms have shown it to be a pervasive characteristic of the comprehension system, stretching even to argument assignment before the verb is known in verb-final languages (Frazier, 1987b; Bader & Lasser, 1994).

1.3.2 How to get from hearing to understanding: A model of comprehension

Before focussing on different models of sentence parsing in more detail in Section 1.3.3, this section presents a brief overview of all the tasks required of a listener hearing a sentence and trying to make sense of it. Mostly it follows the blueprint of the listener provided by Cutler and Clifton (1999); detailed treatment of each individual processing step can be found in Friederici (1999).

In the first place, humans are exceptionally good at recognising speech and separating it from other auditory input, such as background noise and unrelated conversations. Their next task is to segment it into discrete phonemes, the building blocks of words, which in turn form the utterance. The problem here is that acoustic realisations of phonemes vary vastly, depending on who is speaking them and, in particular, which other phonemes they may be co-articulated with. A related problem is the segmentation of the speech stream into words, since assimilation and co-articulation also span word boundaries. However, segmentation can also benefit from incremental word recognition and utterance interpretation, while simultaneously providing a basis for these ‘later’ processes. Thus, prosodic factors
such as stress patterns, rhythm, tone, pitch accent, or morae (depending on the language in question) can all aid segmentation, as can vowel harmony and knowledge about legitimate phoneme combinations within a word or syllable (see Cutler & Clifton, 1999, for details).

Actual *word recognition* begins as soon as the sensory input makes contact with the lexicon (Frauenfelder & Tyler, 1987): most current models of word recognition assume that every entry compatible with the incoming speech is initially activated and competes for recognition.³ To use an example from Cutler and Clifton (1999): *steak, stay, take, ache, state, snake, stack, mistake, first acre* could all be activated at first, but would be ruled out one by one based on accumulating phonetic, syntactic, semantic, and contextual evidence. Ideally, only one word remains, allowing lexical access to its morphological structure, semantics, and syntactic requirements (though see van Turennout, Schmitt, & Hagoort, 2003).

Lexical ambiguities occur at this level of processing, and have been one of the most prolifically researched areas of comprehension. Their basis is the imperfect mapping between words and meanings. For instance, one meaning can be associated with several synonymous word forms (e.g., *bug* and *insect* both denote the same type of animal); equally, one word form can have a variety of separate meanings and even belong to different word classes (e.g., a *bug* can also be a problem or a recording device, while the verb *to bug* means either “to annoy” or “to hide a recording device”). Interestingly, all these senses are initially activated in the mental lexicon, usually irrespective of the frequency of their occurrence and their suitability in the semantic and syntactic context (Swinney, 1979; Tanenhaus & Donnenwerth-Nolan, 1984; see Tabossi, 1988, for evidence that strongly constraining contexts can reduce activation of non-dominant senses).

In addition to semantic meaning, lexical access also supplies syntactic information such as a verb’s subcategorisation frame. The incremental process of determining each word’s syntactic category and combining them to phrases is

---

³ Models of spoken word recognition basically come in two forms. Marslen-Wilson’s cohort model (Marslen-Wilson & Welsh, 1978) and its successors are bottom-up models, as is SHORTLIST (Norris, 1994). Connectionist models such as TRACE (McClelland & Elman, 1986) are interactive and allow feedback from higher processing levels.
Chapter 1: Introduction

referred to as *parsing* (covered in more detail in Section 1.3.3). Its aim is to create a coherent interpretation of the sentence, including the assignment of thematic roles.

Finally, the resulting sentence is integrated into the listener’s model of the current discourse. This contains a memory for what has already been said, inferences that may be drawn on this basis, as well as the listener’s beliefs about the speaker, and general sociolinguistic knowledge (e.g., that the speaker of the phrase *how are you*? is usually not expecting a detailed answer). Only when all these different aspects have been integrated into one representation of what the speaker meant can the listener be said to have comprehended the utterance – and even then, misunderstandings are fairly common in everyday conversation.

In the following section, I provide an overview of different kinds of parsing models. These prepare the ground for three recent, more general proposals of how comprehension is achieved, which will be introduced in Section 1.3.4.

### 1.3.3 Parsing models

Introducing a comprehensive review of models of the parsing process, Tanenhaus and Trueswell (1995) point out that temporary ambiguities are omnipresent in auditory sentence processing, precisely because the word-by-word nature of the input frequently makes it impossible to know how a sentence is going to continue, both syntactically and semantically. In contrast to global syntactic ambiguities such as (3), where the uncertainty about the meaning persists beyond the end of the sentence, local ambiguities are only ambiguous up to the point where a subsequent expression shows how the sentence ought to have been parsed.

(3) *Visiting relatives can be boring.*

Typical examples of local ambiguities are reduced relative clauses (4, 5), and ambiguities of argument assignment (6) and attachment (7):

(4) *The horse raced past the barn fell.*

(5) *The coach smiled at the player tossed the frisbee by the opposing team.*

(6) *After the child had visited the doctor prescribed a course of injections.*

(7) *John hit the girl with the wart.*
If the processor has been garden-pathed while hearing such a sentence, the disambiguating word (underlined in the examples) will be experienced as a grammatical violation until a revised analysis of the utterance is adopted. Such disruption can be measured in longer reading times of the disambiguating region, delayed reaction times, regressive eye movements, or specific ERP effects. This makes such measures informative about how the sentence was interpreted so far, and hence useful for distinguishing between predictions based on different models of sentence comprehension.

But what factors determine how a sentence is initially parsed? Pickering, Clifton, and Crocker (2000b) list available sources of knowledge which could be consulted for this purpose, among which syntactic category information (e.g., transitivity), grammatical features and rules, and facts about the usage frequency of a particular form play a central role. Semantic information is also exploited, at least in ruling out implausible analyses. For example, the fact that a transitive use of the verb to roll requires an animate agent means that sentence (8) is unlikely to be misanalysed, although formally it contains a temporal ambiguity.

(8) When the ball rolled the man jumped.

Other knowledge sources that can influence the processor are prosody and punctuation (adding a comma to sentence (6) disambiguates it), as well as awareness of the previous discourse and current focus. Thus, whether sentence (9) implies that I am seeing the man through the binoculars or that he is in possession of them can depend on how many men have been mentioned so far, and whether it is necessary to distinguish between them (Altmann & Steedman, 1988).

(9) I saw the man with the binoculars.

One of the main points of debate between conflicting models of sentence processing is how soon these different constraints affect the parse. In particular, restricted or two-stage models propose that autonomous syntactic information is prioritised in the first stage of processing, and that other levels of representation such as semantics and discourse play a secondary role. At all levels, processing is assumed to be informationally encapsulated.

The most prominent restricted model is the so-called Garden-Path model (e.g., Rayner, Carlson, & Frazier, 1983; Frazier, 1987a), which proposes that the processor
initially opts for the syntactically simplest parse. If and when this turns out to be incompatible with information from the syntactic, semantic or other processing modules, a complete reanalysis is necessary to create an alternative version. One important experimental finding supporting Garden-Path theory is that even strong manipulations such as the inanimacy of the potential subject in example (10) do not reliably reduce garden-pathing for reduced relative clauses (F. Ferreira & Clifton, 1986), though context can aid recovery from the misanalysis (F. Ferreira & Henderson, 1990; see also Trueswell, Tanenhaus, & Garnsey, 1994).

(10) The evidence examined by the lawyer turned out to be unreliable.

This suggests that the initial parse is indeed based only on syntax, and that the two forms of information are treated independently – conclusions which are borne out by lexical decision and naming results (O'Seaghdha, 1997), and by differential ERP effects for syntactic and semantic anomaly detection (Friederici, 2002; Hagoort, 2003; though see van Turennout et al., 2003; and Van Berkum, 2008, for the claim that semantic processing actually precedes syntax).

In contrast, unrestricted, constraint-based and interactive models hold that the different sources of information are rapidly and continuously integrated and can all contribute to parsing decisions at any time (e.g., Tyler & Marslen-Wilson, 1977; Boland, Tanenhaus, & Garnsey, 1990; MacDonald, Pearlmutter, & Seidenberg, 1994; Trueswell & Tanenhaus, 1994). Parsing is seen as a one-stage process, with no form of representation taking precedence on principle. Indeed, these models actually exploit the correlations that exist between the various forms of information associated with each word (e.g., the frequent co-occurrence of attributes such as being a noun, the subject of the sentence, and an animate agent). Since a word is not necessarily ambiguous at all these levels of representation, it makes sense for information about its status to flow between levels. In addition, the different potential readings of ambiguous forms are not equally frequent, and can be weighted by this probability.

Constraint-satisfaction models generally allow alternative hypotheses to be entertained in parallel. The analysis which receives most support across all the different constraints is most highly activated, so garden-path difficulties result whenever the correct parse of a local ambiguity has not received sufficient support
so far. Initial evidence for interactive accounts came from studies showing early effects of argument structure and other lexical properties of the verb on comprehension (e.g., Boland et al., 1990; Trueswell & Tanenhaus, 1994; Trueswell et al., 1994). More recently, visual-world paradigm experiments have shown that semantic and contextual factors can influence eye movements long before syntactic disambiguation is available (e.g., Eberhard, Spivey-Knowlton, Sedivy, & Tanenhaus, 1995; Tanenhaus et al., 1995; Chambers et al., 2004).

Another distinction between models of parsing concerns the construction of alternative parses. Serial accounts, such as Garden-Path theory, stipulate that only one analysis is available at a time. Parallel models allow several to be activated simultaneously, though total parallelism seems incompatible both with experimental results (cf. Pickering et al., 2000b) and with limitations of processing resources (F. Ferreira & Patson, 2007). Most parallel models therefore include a ranking mechanism between simultaneously activated analyses, with more probable analyses being foregrounded. Constraint-based models are of this type: competing alternative parses are weighted based on their compatibility with the available knowledge sources. In this sense, constraint-based models are ‘extended-parallel’ accounts, where the alternatives remain viable while the sentence unfolds (Pickering et al., 2000b). This distinguishes them from ‘momentary-parallel’ accounts, such as Referential Theory (Crain & Steedman, 1985; Altmann & Steedman, 1988). This ‘weakly interactive’ account combines parallel and serial aspects: alternative syntactic analyses are proposed in parallel at each stage of processing, but discourse and semantic information immediately reject any which are not contextually appropriate, thus eliminating the need to maintain large numbers of parallel analyses over time.

Somewhat similarly, the Unrestricted Race model (Traxler, Pickering, & Clifton, 1998; van Gompel, Pickering, & Traxler, 2001; van Gompel, Pickering, Pearson, & Liversedge, 2005) combines a two-stage design with unrestricted processing. It places no restrictions on the sources of information used to construct alternative parses, but rather than competing for activation, the alternatives engage in a race, with the speed determined by the amount of supporting evidence. Whichever analysis is constructed fastest is selected as the only surviving interpretation of the
fragment in question, so any inconsistencies with later information require complete and costly reanalysis.

A detailed comparison of the many different proposals for the architectures and mechanisms underlying sentence processing is well beyond the limits of this thesis. Having briefly introduced some of the main points of discussion, the following section concentrates on contextual factors affecting online comprehension.

1.3.4 The role of context in auditory sentence processing

Generalising across the different models presented in the previous section, what can we conclude regarding factors which influence how sentences are comprehended? In a recent review of the evidence for and against the different accounts, Pickering and van Gompel (2006) conclude that “semantic processing is delayed relative to syntactic processing... consistent with the predictions of modular accounts” (p. 468). Similarly, while referential context is able to influence syntactic ambiguity resolution, this seems to be limited to cases where there is no strong a-priori preference for one of the two interpretations, at least in reading-based studies (Britt, Perfetti, Garrod, & Rayner, 1992). Other knowledge sources, however, seem more likely to influence early syntactic processing: prosodic information, for instance, is used very rapidly for disambiguation, despite the fact that speakers don’t even provide it consistently (Snedeker & Trueswell, 2003).

Most interesting for this thesis is the finding that although the influence of linguistic context on sentence reading is somewhat volatile, the visual context surrounding a listener exerts a strong and consistent effect even on the earliest moments of syntactic processing, as shown by studies investigating eye movements in the real world (Tanenhaus et al., 1995; Spivey et al., 2002; Chambers et al., 2004). In one such experiment, participants followed temporarily ambiguous instructions such as “pour the egg in the bowl over the flour”. The objects available to them varied in their compatibility with the two readings of the sentence, i.e. in their affordance of the pouring action: eye movements were compared between a condition with one egg in liquid and one in solid form, or both eggs liquid (the display also contained an empty bowl and a heap of flour). Whether listeners
initially misinterpreted “in the bowl” as the destination rather than as a modifier
depended on the number of appropriate eggs: they were consistently garden-pathed
when only one egg could be poured, but much less likely to look at the empty bowl
when both eggs were in liquid form and the instruction therefore required
additional specification (Chambers et al., 2004). In short, listeners’ assessments of
how the instruction could be carried out given the constraints of the current
situation influenced their syntactic analysis of the sentence as it unfolded.

Using a very different task, Knoeferle, Crocker, Scheepers, and Pickering
(2005) also showed that the visual context (in this case a depiction of the verb, e.g.
*painting*) had a rapid effect on the assignment of agent and patient roles in German
SVO- and OVS-sentences, where the status of the first noun phrase was initially
ambiguous between subject and object. In contrast, the visual display was ignored
for verb-final constructions, where only linguistic cues (case-marking and adverbial
bias) allowed role assignment before encountering the verb. This condition is
reminiscent of the situation during reading, where linguistic factors are often the
only available clues to the meaning of the sentence. It suggests that sentence
processing may actually be more interactive in auditory real-life contexts than
during reading. Interestingly, Snedeker and Trueswell (2004) argue that children’s
parses also initially rely only on linguistic constraints, with the adult strategy of
using any form of available information developing gradually. Similarly, Pickering
and van Gompel point out that while traditional two-stage accounts such as
Garden-Path theory fail to appreciate the range of sources informing initial
processing, syntax does often seem to be prioritised: “ Whereas research has shown
that discourse context, plausibility, and frequency play important roles during
sentence processing, they often do not entirely override basic preferences for
particular types of structure” (Pickering & van Gompel, 2006, p. 486).

To summarise, adult listeners in the real world seem to be skilled at rapidly
making use of any kind of information that helps to disambiguate the sentence they
are trying to comprehend, including the properties of objects surrounding them and
their knowledge of the previous discourse. However, when such information is
reduced or unavailable, the default strategy is to rely on the structural properties of
the sentence.
At the same time, it seems that studies measuring eye movements during reading may underestimate the importance of context, due to the nature of the stimuli they employ. The investigation of real-time situated language allows for greater consideration of contextual factors, and the corresponding change of focus in methodology and research questions has generated novel accounts of language comprehension, emphasising the goal of understanding-in-context more than the details of moment-by-moment processing. Three of these proposals are presented in the following.

1.3.4.1 Eye movements during comprehension in a visual world

In fact, studies using the visual-world paradigm have shown that people are extremely quick to integrate the visual context into the developing representation of the speaker’s meaning: listeners generally look at a visual entity as soon as they have accessed its referential expression. For example, Allopenna, Magnuson, and Tanenhaus (1998) showed that within 200 ms of the onset of the word beaker, people were more likely to look at the target than at a phonologically unrelated object (e.g., carriage). However, they were also more likely to fixate a cohort competitor (e.g., beetle) and – slightly later – a rhyme competitor (e.g., speaker), suggesting that eye movements are continuously mapped with referential expressions, in line with lexical access (see also Dahan, Magnuson, & Tanenhaus, 2001a; Dahan et al., 2001b).

However, phonological identification is not the only factor driving fixations during comprehension: Huettig and Altmann (2005) recorded their participants’ eye movements while they listened to sentences such as (11):

(11) Eventually, the man agreed hesitantly, but then he looked at the piano and appreciated that it was beautiful.

Three conditions varied the objects displayed in a 2 × 2 grid: these either included a piano, a trumpet, or both a piano and a trumpet, each presented among distractors. Fixations to the piano rose rapidly when the word “piano” was mentioned, but so did fixations to the trumpet (in the trumpet-only condition). Clearly, semantic category information was activated almost as soon as the word could be identified, and it was strong enough to drive eye movements (see also Yee & Sedivy, 2006). Similar findings have been reported for shape competitors (Huettig & Altmann, 2007) and even objects whose use involves similar body movements.
(e.g., piano and typewriter; Myung, Blumstein, & Sedivy, 2006). To illuminate the time course with which different kinds of information influence eye movement behaviour during comprehension, Huettig and McQueen (2007) varied the temporal presentation with respect to the onset of the target word of displays containing phonological, shape, and semantic competitors. They concluded that fixations signal the detection of an overlap between the utterance and the display at any of these levels of processing.

However, all these studies looked only at factors affecting single-word comprehension in limited displays of unconnected objects. In contrast, participants in a classic study by Altmann and Kamide (1999) heard sentences like The boy will eat the cake or The boy will move the cake while looking at an ersatz scene which depicted a boy, a cake, and several other objects (e.g., a toy train). Directly on hearing the verb eat, participants looked at the cake more often than on hearing move. Importantly, these eye movements occurred before the onset of cake: given the context of the sentence, participants were able to use the meaning of the verb to narrow down the list of potential referents to something edible. In this way, they were able to predict what was likely to be mentioned next, rather than fixating what had already been referred to.

Subsequently, such predictive eye movements have been shown to depend on the listener’s interpretation of the prior context, and not just on the meaning of a single word such as eat (e.g., Kamide et al., 2003a; Kamide, Scheepers, & Altmann, 2003b; Knoeferle et al., 2005). In fact, Altmann and colleagues argue that they reflect incremental processing, applying all constraints available from any source of information to each sentence fragment as it unfolds. For each new word, this involves representing everything that has been heard so far, beginning (in the example) with syntactic assignment of the boy to the subject role and will eat as the future-tense verb. Additionally, the emerging sentence representation can include the semantics of the two words, and ultimately the anticipation that something that will be eaten is going to be mentioned postverbally, and that this must refer to something the depicted boy would be able to eat (Kamide et al., 2003a).

Interestingly, predictive eye movements do not require the referents to be visible (Altmann, 2004; Altmann & Kamide, 2004), and can also reflect the consequences of a described event (Altmann & Kamide, 2007; Knoeferle & Crocker,
2007). For instance, participants heard either “the man will drink...” or “the man has drunk...” while looking at an ersatz-scene including a full glass of beer and an empty wine glass. The empty wine glass attracted more saccades during the sentence in past tense than in the future version, and vice versa for the full beer glass (Altmann & Kamide, 2007). Altmann and Kamide conclude that eye movements during comprehension are not just a direct reaction to each word as it is encountered. Instead, they reflect the unfolding representation of the language in the context of the constraints and affordances implied by the scene, the discourse, and any other contributing factors: “Anticipatory eye movements do not reflect the unfolding language; they reflect an unfolding (mental) world” (Altmann & Kamide, 2007, p. 515). In their linking hypothesis between language processing and visual attention, the greater the conceptual overlap between a visuo-spatial representation of an object and its affordances on the one hand, and the language-mediated representation of the description or event on the other, the more likely people are to launch an eye movement to the region of the visual world currently associated with that object (Altmann & Kamide, 2007; 2009; see also Altmann & Mirković, 2009).

1.3.4.2 The Coordinated Interplay Account

A related explanation for the interaction between visual context and auditory comprehension is the Coordinated Interplay Account of situated sentence processing (Knoeferle & Crocker, 2006; 2007; Crocker, Knoeferle, & Mayberry, in press). According to this proposal, the comprehender uses both semantic knowledge (e.g., stereotypes) and the surrounding visual world (e.g., depicted actions) to establish reference and assign thematic roles, though visual information is preferred if the two are in conflict (Knoeferle & Crocker, 2006). This priority of the visual context is explained through the close temporal relationship between comprehension and the allocation of visual attention: the incremental interpretation of the unfolding sentence drives attention in the visual world, grounding and even anticipating events and referents. This results in an ‘utterance-mediated’ increase in salience of the attended objects, and in turn makes them more likely to be used for disambiguating and interpreting the remainder of the utterance.

For instance, in one of Knoeferle and Crocker’s experiments (conducted in German), participants saw a display containing a pilot, a wizard with a spy-glass,
and a detective holding out a plate of food (Knoeferle & Crocker, 2006). On hearing the initial sentence fragment “the pilot\text{ACC}”, they began to construct a representation in which the pilot was the target of some future event, so they directed their attention to fixate him. At this stage, they could also consult the scene to discover what was happening with the pilot: he was being spied on, as well as being served food. Thus, as soon as they heard the next fragment, “is being spied on”, this event was integrated into the unfolding sentence representation and they could now try to predict the upcoming agent of the event. However, two different knowledge sources provided conflicting evidence about the spy-er: stereotypically, a detective is more likely to be spying than a wizard, but the visual context showed exactly the opposite. Anticipatory eye movements to the wizard revealed that participants prioritised the visual information and used it to understand the sentence before it had even been fully heard. Thus, the very fact that visual attention is initially mediated by the utterance allows an efficient extraction of those aspects of scene information which can contribute towards the sentence interpretation.

In the most recent version of the Coordinated Interplay account, which includes a computational model with linking hypotheses to behavioural and electrophysiological data (Crocker et al., in press), the authors argue that sentence interpretation, visual attention, and scene apprehension have co-evolved to provide an online grounding mechanism during comprehension in the real world. However, despite the strong interconnection between the three processing systems, they remain distinct from each other, and situations where they might not interact are conceivable (e.g., in tasks demanding divided attention). This is a critical difference compared to Altmann and colleagues’ account, where a single representational substrate is assumed to underlie the processing of linguistic and non-linguistic information and the allocation of attention. There is no fundamental separation between understanding language and understanding the world (for a more detailed comparison see Altmann & Mirković, 2009). Nonetheless, the two accounts resemble each other in many other points, not least in the incremental development of the sentence representation as the utterance unfolds, and in the substantial influence of the visual context on this representation. In addition, both consider prediction to play a central role in sentence processing.
This is in line with an increasingly widely held view in language comprehension research: evidence of predictive processing has been reported not just in the visual-world experiments described above, but also in reading paradigms (e.g., Frisson, Rayner, & Pickering, 2005; Staub & Clifton, 2006) and in ERP studies (DeLong, Urbach, & Kutas, 2005; Van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005). It has been suggested that this kind of prediction might be achieved through the comprehender employing their own production system, in a sense ‘re-producing’ or emulating the very same procedures that the speaker is employing (Pickering & van Gompel, 2006; Pickering & Garrod, 2007). I return to this issue of the interface between production and comprehension in Section 1.4.1.3.

1.3.4.3 The ‘good-enough’ approach to comprehension

To conclude this discussion of how listeners come to understand spoken sentences, in this final sub-section I briefly introduce a rather different account, which cuts across many of the previously mentioned issues. All the models of parsing reviewed above are based on the idea that comprehenders process each unit of the unfolding sentence step-by-step and integrate it into the whole, ultimately arriving at a detailed and generally accurate representation of its meaning. The ‘good-enough’ approach to sentence processing (F. Ferreira, Bailey, & Ferraro, 2002; F. Ferreira & Patson, 2007) challenges this basic assumption by providing evidence that the final sentence representation – however it has been derived – is frequently incomplete or even inaccurate. For example, given a sentence context such as (12), comprehenders often fail to detect the semantic anomaly (Barton & Sanford, 1993; Bohan & Sanford, 2008).

(12) The plane crashed right on the border. Wreckage was equally strewn in France and Spain. The authorities were trying to decide where to bury the survivors.

They also tend to misinterpret implausible passive sentences like (13) in the way suggested by the content words (i.e. the dog bit the man, cf. F. Ferreira, 2003).

(13) The dog was bitten by the man.

Finally, the meaning of the initial garden path in sentences such as (14) has been shown to remain activated even after disambiguation, and may influence the final representation of the sentence: questioned whether Anna dressed the baby,
participants are highly likely to answer “yes”, although the sentence does not contain this information (Christianson, Hollingworth, Halliwell, & Ferreira, 2001; see also van Gompel, Pickering, Pearson, & Jacob, 2006).

(14) While Anna dressed the baby played in the crib.

Ferreira and colleagues argue that comprehenders frequently rely on heuristics to obtain the meaning of a sentence, rather than incrementally computing a complete syntactic representation (F. Ferreira et al., 2002; F. Ferreira, 2003; F. Ferreira & Patson, 2007). This strategy is particularly likely to be adopted if the content words match the context the sentence occurs in:

The interpretation of the sentence seems fine to the comprehender (even though in reality it is unlicensed by the input), and so he or she judges that processing has been successful and that further operations are not necessary. (Christianson et al., 2001, p. 400)

At first glance, the ‘good-enough’ strategy seems risky, especially as most of the supporting evidence comes from situations where it leads to an error. However, Ferreira and Patson (2007) point out that it is actually very efficient to normalise implausible sentences so that they appear reasonable. After all, speakers make errors too, and don’t always say exactly what they mean. Similarly, a lingering garden path as in sentence (14) will rarely even be noticeable, since it is actually quite plausible that Anna did dress the baby, as well as herself. Consequently, it makes sense for listeners to economise on processing resources and arrive at the most probable representation as soon as possible (cf. Gigerenzer, 2008, on the use of heuristics in cognitive processing generally). In particular, this strategy seems likely to be adopted in real-world interactions, which differ from the situation in lab experiments with regard to the number of simultaneous demands for attention, as well as in the amount of additional information available. Intriguingly, Ferreira et al. (2002) suggest that good-enough processing is particularly relevant for dialogue interactions, though this seems likely to depend on the task:

Usually when people talk to one another, turns are not separated by gaps. Therefore, comprehension and production processes must operate simultaneously. The goal of the comprehension system might be to deliver an interpretation that is just good enough to allow the production system to generate an appropriate response; after all, it is the response that is overt and that determines the success of the participants’ joint activity. (F. Ferreira et al., 2002, p. 14)
Language use in dialogue forms the final topic of this introductory chapter, which includes the presentation of a model and a summary of findings from studies exploring how this joint activity is conducted. First, however, I summarise the main points made regarding language comprehension.

1.3.5  **Summary: Language comprehension**

As we have seen, sentence comprehension involves a series of steps which ultimately lead to a representation of what the speaker (or writer) meant to communicate. Models of sentence processing and of comprehension in general differ with regard to the order in which the operations leading from input to understanding are performed, the knowledge sources deemed critical for this process, and the quality of the ultimate meaning representation. However, they generally agree that comprehension is an inherently incremental process, with incoming information being dealt with rapidly, efficiently, and usually fairly effortlessly. There is also a growing consensus that much of this processing is predictive, including the construction of likely sentence continuations and thus the anticipation of the speaker’s meaning.

The interplay of various different forms of information affecting comprehension is reflected in listeners’ eye movements. On the one hand, listeners fixate objects that are being referred to (or spatial locations associated with them), reminiscent of the same tendency during language production. On the other hand, the temporal relationship between fixation and mention is much looser than for name-related gazes in production: eye movements during comprehension can both follow (e.g., Allopenna et al., 1998) or precede referring expressions (e.g., Altmann & Kamide, 1999), presumably depending on how complete and accurate a representation the listener has succeeded in constructing so far, and how well this corresponds to what they see around them.

Arguably, this can make eye movements in comprehension harder to interpret than in production. At the same time, well-designed studies using variants of the visual-world paradigm – where eye movements are related to the linguistic structure and content of the unfolding language as well as to the static or dynamic
properties of the surrounding world – can provide moment-by-moment insights into which sources of information are exploited as people make sense of an utterance and its visual context, and how they integrate the two based on their current interests. Or, as Van Berkum optimistically asserts regarding the use of ERPs for studying sentence processing: “What we see is an opportunistic, proactive brain at work” (Van Berkum, 2008, p. 379).

1.4 Dialogue and interactive language use

One important issue for the models of comprehension outlined in the previous section is the extent to which the listener’s sentence representation accurately reflects the speaker’s meaning, and what happens when subsequent information reveals that the initial analysis was wrong. This debate can obscure the fact that in everyday conversation a mis-parse is rarely critical; for an addressee experiencing difficulties it is often easiest simply to ask for clarification. Even more frequently, the speaker will realise the problem and reformulate the utterance unambiguously, or provide other forms of assistance. After all, both interlocutors have a vested interest in the success of their communication: conversation is a collaborative enterprise, in which speakers and addressees cooperate to achieve mutual understanding (Clark & Wilkes-Gibbs, 1986; Schober & Clark, 1989; see also Grice, 1975).

The possibility for interlocutors to adapt both the form and content of their own utterances as well as the interpretation of their partner’s to the current situation is just one way in which dialogue interactions differ from monologue experimental designs. Consequently, the final part of this introductory chapter discusses some of the characteristics of dialogue, and the extent to which they can be and have been investigated using experimental methodology. Section 1.4.1 considers the differences between dialogue and monologue, which provide a basis for Clark’s assertion that language in conversation is a collaborative product, for which all interlocutors share responsibility (e.g., Clark, 1996). This means that comprehension and production frequently take place simultaneously, making it important to study the influence of these two tasks on each other.
In fact, Garrod and Pickering (2004) assume that the same representations are accessed in the comprehension and in production, and that this is one of the reasons for the comparative ease of interactive communication. Their Interactive Alignment model (Pickering & Garrod, 2004) is presented in Section 1.4.2, together with some experimental evidence. Next, Section 1.4.3 discusses what has arguably been the most intensely debated area of dialogue research: the extent to which interlocutors take each other’s perspective. Experimentally, studies using the visual-world paradigm have played a prominent role in examining this question, mainly with regard to referential ambiguity and its resolution. Indeed, Tanenhaus and Trueswell (2005) claim that recording eye movements provides a promising way of investigating interactive language use in general. One such application is to compare interlocutors’ gaze patterns across time as a way of quantifying the amount of coordination between them (1.4.4). A brief summary is provided in Section 1.4.5, which also addresses some methodological problems in developing useful paradigms.

1.4.1 Dialogue as joint action

1.4.1.1 Differences between dialogue and monologue

In formal and written language, sentences are usually complete and grammatically correct, ellipsis and repetition are avoided, and turns are at least a sentence long – in fact, an entire book containing several thousand sentences can count as one turn. Spoken language in conversation is very different, as the following transcript reveals. It is part of a conversation between two students performing a joint spot-the-difference task, comparing photos of a cluttered kitchen (Note: square brackets indicate overlapping speech).

1. A: ...and then two pots, maybe for sugar, I don’t know, two white pots...
2. B: ... are yours spotty?
3. A: No, they’re just plain white.
4. B: Oh, I’ve got spotty ones.
5. A: OK. And then directly above the pots there should be a plug socket, and the orange light’s on.
6. A: Ok, to the right of the pots there should be fairy and a scrubbing brush...
7. B: Yeah, a blue scrubbing brush?
8. A: ... [uh...]
9. B: [well, blue-black,] looks like.
10. A: Yeah.
11. B: And green [fairy]
12. A: [Yeah.]
13. B: and then the sink
14. A: Mhm, with a white tray inside it
15. B: Yeah, and a chopping board just behind the sink
16. A: Yes. Yeah, we're going good here!
17. A: So, above the fairy, on the window sill, I can see two wine bottles in a wine rack?
18. B: Yeah. Are they on the upper ... [ones?] 
19. A: [No, they're on the lower...]
20. B: Oh, mine are on the upper!
21. A: ok. Yeah, [and they've got red...tops.]
22. B: [...]and they're pointing down, yeah?]
23. A: And then to the right, I think there's a watering can.
24. B: Yeah.
25. B: Yeah, with a plant pot behind?
26. A: ... Yes.
27. B: Yeah? ... and then out of the window, can you see trees?
28. A: Yeah. (laughs) ... be interesting if you could move those! 4

Despite the comparatively strong internal structure imposed by the task they are performing, the two participants regularly produce elliptical and fragmentary utterances that would make little sense on their own (e.g., 9, 11, 13, 14, 20). The excerpt also features interruptions (2, 19, 22), overlapping speech (10/11, 21/22), disfluencies (1, 8, 18), and it is quite repetitive. However, the interlocutors appear to be satisfied with the conversation (16). Their success at discovering two differences in this short time also testifies to the fact that they understand each other.

As this example reveals, interlocutors often simultaneously comprehend an utterance and plan their own response, which they may subsequently be obliged to modify on the fly. Thus, dialogue requires constant spontaneous task-switches between the roles of speaker and addressee, and between the modalities of production and comprehension. Nonetheless, participating in a conversation is surprisingly easy – generally much easier than typical monologue skills such as

4 This data was collected for a pilot study which is not included in this thesis, though it did inform the stimuli used in Experiments 2-7 (see Chapter 4, Section 4.3.1, and Appendix C).
giving a speech or writing a paper, which many people never learn (Garrod & Pickering, 2004).

In contrast, dialogue requires no formal training. In fact, face-to-face conversation is the form in which we first acquire language at all, supporting the claim that dialogue is the primary and most basic form of language use (Fillmore, 1981; Clark, 1996; Garrod & Pickering, 2004). This suggests that the human language processor is best adapted to the characteristic features of face-to-face conversations: for instance, such interactions are immediate, with both interlocutors in the same location. No technology is required to transmit the speaker’s signal to the hearer, but the signal is transient, with no record of what was said remaining, apart from what the interlocutors managed to commit to memory (for a more detailed discussion of basic features of dialogue, see Clark & Brennan, 1991). Consequently, Clark (1996) suggests that whenever some of these basic features are missing from a given interaction, successful communication becomes more effortful to achieve. This is obvious when comparing the likelihood of misunderstandings during a telephone call or email exchange with those during a personal chat. It is also an important consideration with regard to many psycholinguistic research paradigms, where participants are presented with linguistic stimuli in isolation (cf. Schober & Clark, 1989).

1.4.1.2 Shared responsibility for communicative success in dialogue

According to Clark (1996), dialogue is easy because interlocutors collaborate in generating the conversation and in achieving mutual understanding. He views dialogue as a joint activity, comparable to ballroom dancing, playing tennis, or completing a commercial transaction: coordinated actions of two or more participants advancing step by step towards a common goal. A successful interaction of this type does not mean that participants perform the same moves at the same time; rather, their moves must be carefully coordinated to complement each other. In conversation, this means that each utterance must be constructed appropriately for the addressee at that particular point, which in turn requires the speaker to pay constant attention to any form of feedback (e.g., whether a particular term has been understood). For example, A’s hesitation at (8) caused B to modify and reformulate her potentially misleading description of the scrubbing brush.
Hence, both participants were equally responsible for reaching the understanding signalled in (10): B was trying to produce an adequate description for A, but in order for her to accommodate his needs, he let her know when he encountered difficulties, and when they were resolved. In this way, even an extensive conversation can be viewed as “a chain of paired actions” (Clark & Bly, 1995, p. 385), consisting of many individual contributions negotiated between the interlocutors.

1.4.1.3 Implications of dialogue for production and comprehension

The continuous task-switches between listening and speaking in dialogue would be facilitated if some of the processes involved could be conducted in parallel and/or if representations were shared between modalities. As Pickering and van Gompel (2006) point out, many models of language production (e.g., Levelt, 1989) do in fact assume a self-monitoring process which works through the comprehension system. By contrast, comprehension accounts rarely foresee any role for production. Instead, the two processes of encoding and decoding are assumed to be largely autonomous, one preceding the other. Yet, given sufficient sentence context, comprehenders frequently anticipate the final words in a spoken sentence, or at least make a very good guess about their syntax and semantics (see above; also DeLong et al., 2005; Van Berkum et al., 2005), which might be viewed as a form of production itself (Pickering & Garrod, 2007). An advantage of such prediction is that it allows comprehenders to begin processing the speaker’s meaning without having to wait until they have heard the whole utterance. In addition, prediction is useful in compensating for noisy or ambiguous input.

Pickering and Garrod (2007) suggest that comprehension makes use of continuously evolving predictions, which are covertly generated by the production system. Roughly, the idea is that the comprehender’s production system imitates the speaker’s production, simulating the processes it would go through if it were formulating an overt utterance. Similar forms of covert imitation have been shown for non-linguistic tasks such as motor perception too, where perceiving another’s movement leads to activation of one’s own motor areas. It is assumed that this activation reflects internal simulation of the perceived action (for an introduction to this literature, see Barsalou, 2008). For motor movements, Wilson and Knoblich
(2005) propose that such imitation fulfils very similar functions to those discussed above for linguistic prediction:

Such internal modeling allows the perceiver to rapidly interpret the perceptual signal, to react quickly, to disambiguate in situations of uncertainty, and to perceptually complete movements that are not perceived in their entirety. (Wilson & Knoblich, 2005, p. 468)

Correspondingly, being able to use the language production system to anticipate upcoming to-be-comprehended material could dramatically affect the speed, accuracy and necessary effort of decoding the speaker’s utterance, and aid the process of generating an adequate response.

A related proposal at the level of individual linguistic representations, rather than for the entire language processing system, also comes from Pickering and Garrod (2004; Garrod & Pickering, 2004). One of the key assumptions of their Interactive-Alignment account (see Section 1.4.2) is the parity of representations between comprehension and production (see also Liberman & Whalen, 2000). In other words, hearing a particular sentence leads to increased activation of its syntactic structure and word forms. Because comprehension and production are assumed to draw on the same representations, such pre-activation can explain the tendency to repeat material from previously heard or produced sentences, i.e. various forms of priming. An advantage of this recycling of utterance components is that they do not need to be encoded from scratch, thus economising on cognitive resources and again facilitating quick and flexible turn-taking in conversation.

In short, although these two accounts differ in scope and experimental support, both use the assumption of a bi-directional link between production and comprehension to explain the comparative ease of dialogue interaction, as opposed to monologue. At the same time, they reveal the limits on autonomous accounts of comprehension and production, in that the two modalities clearly influence each other to a considerable extent.
1.4.2 Interactive comprehension and production: A model of dialogue

Pickering and Garrod’s (2004) Interactive-Alignment account provides an alternative, collaborative description of language use in dialogue settings. It assumes that the goal of any interaction is the alignment of situation models (Zwaan & Radvansky, 1998) for those aspects of the world which are under discussion. In other words, communication is successful if the interlocutors develop a shared understanding of critical aspects, such as who they are speaking of, the time and place they are referring to, and potential implications and consequences. Studies of interactive language use show that such alignment of representations is achieved largely without conscious negotiation (Clark & Wilkes-Gibbs, 1986; Garrod & Anderson, 1987). Instead, Pickering and Garrod assume that it is based on automatic priming processes. I have already alluded to the fact that interlocutors in dialogue tend to prime each other at various linguistic levels, e.g., choice of words, grammatical form, aspects of meaning, and even articulation. For instance, hearing a picture described as “the priest is offering a cake to the painter” can have consequences for the syntax and semantics of subsequent utterances, as well as for other levels of representation: it becomes more likely that following sentences will have the same prepositional object structure (rather than the double object), but also that the words priest, offer, and painter will be reused (instead of e.g., monk, give, artist) – indicating syntactic and lexical alignment, respectively. Importantly, alignment is assumed to spread not just horizontally within one level of representation, but also vertically. In Pickering and Garrod’s words, it “percolates” between levels of representation: alignment at one level generates alignment at other levels, right up to and including the level of situation models, which is the key to successful communication. As already mentioned, this interactive generation of alignment relies on the assumption of representational parity between comprehension and production. For instance, because the same representations are accessed during speaking and listening, the activation of a prepositional dative sentence structure caused by hearing such a sentence may affect which structure is selected for subsequent production, or how a sentence is preferentially parsed in subsequent comprehension.
Experimental support for interactive alignment exists for many different levels of language representation. The requirement that participants be partaking in a dialogue interaction has meant that most studies have used variations of Krauss and Weinheimer’s (1964; 1966) referential communication task; generally, interlocutors play some kind of collaborative game, often navigating mazes or describing ambiguous pictures to each other. Often, they are assigned specific roles, such that one plays a director and the other a matcher; the use of a confederate of the experimenter in one of these roles is also frequent.

In one of the first studies of this kind, Garrod and Anderson’s (1987) participants had to assist each other in navigating through a computerised maze, where success depended on their ability to communicate their positions unambiguously. Players employed a number of different strategies to do this, e.g., describing a path (see the bottom right, go two along and two up), supplying coordinates (I’m at C4), line description (I’m on the second row from the bottom at the end on the right), or naming landmarks (I’m at the right-turn indicator). Interestingly though, pairs tended to converge on one of these strategies without explicit negotiation and usually stuck with it, which suggests that the members of each pair rapidly developed the same situation model of the maze.

Individual referring expressions were also frequently reused within a pair, for example the word level to denote a horizontal row in the line description strategy. In three experiments exploring the nature of such lexical entrainment, Brennan and Clark (1996) showed that if a stimulus picture was initially named using a complex expression such as pennyloafer or docksider, pairs of interlocutors tended to reuse this term to designate that particular shoe, even in new contexts which did not require the additional specification. According to Brennan and Clark, such mappings of objects to particular names constitute conceptual pacts, tacitly agreed on by pairs of interlocutors. Clark and Wilkes-Gibbs (1986; see also Wilkes-Gibbs & Clark, 1992) provide further evidence of lexical alignment.

Reports of syntactic alignment between interlocutors are also numerous (see Pickering & Ferreira, 2008, for a review, and Chapter 2 of this thesis for a detailed summary); indeed, the effects of syntactic priming can be stronger in dialogue settings than in monologue (see Branigan, Pickering, & Cleland, 2000). Finally, there is also evidence of alignment at one level affecting alignment at another. For
instance, syntactic priming is enhanced if open-class lexical expressions in the two sentences are repeated or related: more participants produce “the sheep that’s red” (rather than “the red sheep”) if they have previously heard a description such as “the goat that’s red” than after “the book that’s red” (Cleland & Pickering, 2003). This spread of alignment across multiple levels of representation provides a convincing reason for the wide-spread repetition found in dialogue (Garrod & Pickering, 2004). Priming of this kind means that the desired alignment at the level of the situation model can develop as an automatic consequence of alignment at lower levels.

1.4.3 Perspective-taking in reference assignment

From automatic processes which create coordination between interlocutors, we turn now to an aspect of conversation which may be under more conscious control: the extent to which interlocutors take their partner’s perspective into consideration. Certainly, social norms of politeness suggest that adopting the viewpoint of one’s partner is the correct thing to do whenever possible (Mainwaring, Tversky, Ohgishi, & Schiano, 2003). Indeed, Tversky and Hard (2009) show that the mere presence of a person on a photo can lead participants to describe objects in front of him from his perspective, rather than their own.

In a dialogue study, where a director had to describe the location of one of two circles so that the matcher could select it correctly, Schober (1993) found that overall, directors adopted the matcher’s perspective more often than any other strategy of description. However, he also found that directors speaking to a real matcher used egocentric descriptions more often than directors who recorded their instructions for an absent matcher. Presumably, they were relying on the matcher’s ability to question the description. Schober concludes that spatial perspective-taking depends on many factors:

Speakers do more than take into account their own egocentric bias, their addressee’s relative position, and the nature of the objects they are describing. They also take into account what their addressee gives as evidence of having understood, and how their addressees have already described locations. (Schober, 1993, p. 20)
Yet many researchers would argue that speakers do not routinely “take into account” all these factors, since maintaining all these separate representations is extremely resource-intensive (e.g., Keysar, Barr, & Horton, 1998b; Keysar & Barr, 2005). Similarly, interactive alignment means that interlocutors do not need to model each other’s perspective extensively. If situation models are aligned – and, as we have seen, there is a strong tendency for this to occur – communication will be successful even if each interlocutor simply uses their own knowledge state as a proxy for their partner’s (Pickering & Garrod, 2004).

Most of the debate on the extent of egocentric processing in dialogue has focussed on whether speakers avoid producing ambiguous utterances, and whether addressees make use of common ground during reference resolution. The idea is that definite references such as the small candle, or indeed pronouns such as he, she, and it are only felicitous when the referent can be uniquely identified in the current context. Therefore, investigations into which objects are considered by speakers and addressees to be potential competitors (thus requiring further specification) provide insights into the extent to which they take each other’s perspectives into account when formulating and interpreting utterances. Since these studies also constitute the majority of work to date using eye movements in dialogue, I will first sketch out the two sides of the argument (Section 1.4.3.1), before reviewing the empirical evidence (1.4.3.2).

1.4.3.1 The use of common ground by speakers and listeners

Common ground refers to the totality of information that is shared between participants in a conversation. Such shared knowledge accumulates over the course of the interaction, as each new contribution becomes grounded. For example, interlocutors tend to develop their own names to refer to objects, such as the ice-skater for an ambiguous tangram figure (Clark & Wilkes-Gibbs, 1986). In subsequent turns, this term can be reused frequently and without further specification, on the assumption that all interlocutors have tacitly agreed that it refers to that particular figure. Common ground of this kind, which grows out of the collaborative grounding of contributions, is based on linguistic copresence (Clark & Marshall, 1981). Another form of common ground particularly relevant to visual-world studies is physical copresence, i.e. objects that are visible to all interlocutors and can
therefore be referred to easily. Finally, common ground can be based on community membership; for instance, I can assume that anyone who has lived in Edinburgh knows what Arthur’s Seat refers to, and that members of the psychology department know where the coffee room is.

According to Clark and colleagues’ Restricted Search hypothesis (e.g., Clark & Carlson, 1981; Clark & Marshall, 1981), common ground limits the set of objects which are considered as potential referents. Specifically, both speakers and addressees are assumed to follow the principle of optimal design (Clark, Schreuder, & Buttrick, 1983):

The speaker designs his utterance in such a way that he has good reason to believe that the addressees can readily and uniquely compute what he meant on the basis of the utterance along with the rest of their common ground.

Working backwards, the addressee can assume he has been given enough information and can therefore reason through to the speaker’s meaning. (Clark et al., 1983, p. 246)

If interlocutors follow this principle, they will comprehend and produce utterances that are consistent with common ground, and this will ensure that they understand each other (Clark, 1996). However, Keysar and colleagues argue that, rather than explicitly considering which facts are shared, interlocutors initially assign reference only from their own perspective (Horton & Keysar, 1996; Keysar, Barr, Balin, & Paek, 1998a; Keysar, Barr, Balin, & Brauner, 2000). Their two-stage Perspective-Adjustment account proposes that addressees’ processing begins with a fast, unrestricted search for any suitable referent. Common ground is only considered later, during a cognitively demanding monitoring process, which serves to correct obvious errors in the initial interpretation. In fact, if one’s own knowledge is often a good proxy for common ground in dialogue (Pickering & Garrod, 2004), the risk of occasional misinterpretation due to such an egocentric heuristic may be outweighed by the benefits of speed and efficiency, particularly in an ongoing conversation.

Between these two conflicting accounts of Restricted Search and Perspective-Adjustment sits a third approach, claiming that interlocutors consider common

5 A similar Monitoring & Adjustment account by the same authors assumes a comparable processing sequence during production.
ground whenever it is convenient to do so, for instance if time and resources allow (Horton & Gerrig, 2005a; 2005b; Bard et al., 2007).

### 1.4.3.2 Eye movements as early indicators of reference assignment

To determine whether comprehenders behave egocentrically, a frequently-used paradigm is to provide the comprehender with information that the speaker cannot be aware of. In one of the first published studies using eyetracking in dialogue, Keysar et al. (2000) monitored the eye movements of an addressee who was following instructions to move everyday objects such as a car, a candle, and an apple. The objects were positioned in a vertical lattice between the addressee and a confederate director, but the addressee could see that some slots were occluded from the director’s view. These occluded slots contained the critical items; e.g., while the addressee might see three candles of different sizes, the smallest candle would be hidden from the director. Thus, if the addressee adopted an egocentric interpretation of the instruction “move the small candle beside the bottle”, she would be expected to fixate the smallest candle, considering it as a potential referent. In contrast, if she immediately used common ground to restrict the potential referents, she should ignore the candle the speaker did not know about and consider only the middle-sized one, i.e. the smallest one that he could be aware of. Indeed, before hearing the critical utterance, addressees looked almost only at shared objects, suggesting that they concentrated on objects in common ground. On hearing the instruction however, addressees in the test condition fixated the hidden object earlier and for longer than in the control case, where the smallest candle was replaced with a toy monkey; they frequently even reached out for it. So despite being aware of the different status of shared and hidden objects, their eyes were still drawn to the hidden distractor, in accord with the Perspective-Adjustment account (see also Keysar et al., 1998a).

However, Hanna, Tanenhaus, and Trueswell (2003) pointed out that the potential referents in Keysar et al.’s study were not jointly entered into common ground, since common ground was based on physical copresence only. In their experiment, the speaker (again, a confederate) instructed the addressee to arrange shapes on a board, allegedly to make their boards match. This mention as part of the discourse meant that all potential referents were grounded via linguistic copresence.
Each trial began with the addressee being handed a ‘secret shape’ by the experimenter. In the test condition, this was identical to the object the speaker would refer to in the critical utterance, e.g., a red triangle for “now put the blue triangle on the red one”. The visual context for this instruction could include either one or two red triangles, of which one was secret or both were shared. The critical question was whether a same-colour competitor would compete with the target as strongly when it was ‘secret’ as when it was in common ground. This was not the case: when the competing shapes were both in common ground, addressees showed frequent shifts of fixation between them and often asked for clarification. In contrast, when one was shared and one secret, initial looks were more often directed to the (shared) target than to the (secret) competitor, and the proportion of looks to the target was higher than to the competitor throughout the duration of the trial. Two conclusions can be drawn from this result: on the one hand, the competing referent in privileged ground was not wholly ignored, so common ground did not completely restrict the domain of reference. However, there was also no evidence for an initially egocentric stage of processing; common ground contributed to reference resolution from the very start.

Hanna et al. (2003) therefore suggested that alternative interpretations are assessed in parallel and based on multiple cues. In this view, common ground is just one kind of contextual constraint, the strength of which depends on various factors (for similar conclusions, see Nadig & Sedivy, 2002, demonstrating children’s consideration of common ground; Hanna & Tanenhaus, 2004, using an innovative real-world design based on cooking).

A similar experiment-based exchange of arguments between the two camps has addressed the issue of whether conceptual pacts are interlocutor-specific (i.e. part of common ground), or whether they transfer to other conversational partners (see Brennan & Clark, 1996; Barr & Keysar, 2002; Metzing & Brennan, 2003; Kronmüller & Barr, 2007). Yet although the question of reference resolution is clearly one to which the perspective of both speaker and addressee are critical, none of these studies examined unscripted conversation between interlocutors. At best, one of the two was a confederate (see Lockridge & Brennan, 2002, for a discussion of the risks of using confederates; Kronmüller & Barr, 2007, for arguments in support of pre-recorded instructions).
As an alternative, Brown-Schmidt and Tanenhaus (2008; Brown-Schmidt, Gunlogson, & Tanenhaus, 2008) have recently proposed using so-called targeted language-games to study referential communication. These are designed to elicit high proportions of the kind of utterance under examination (e.g., *wh*-questions), while allowing pairs of naïve participants to interact with each other naturally, and to focus on the collaborative task they are performing. This also reduces the risk of strategic eye movements or those affected by the availability of only a very limited number of entities to fixate. Another advantage is that existing computational models of dialogue tend to be task-oriented too, opening up future possibilities of comparisons between experimental and computational investigations.

In one study of this kind, participants were completely unscripted in how they described complimentary patterns of blocks to each other (Brown-Schmidt & Tanenhaus, 2008). The results differed from comparable studies in a monologue setting in an interesting way: similar to Allopenna et al. (1998), Brown-Schmidt and Tanenhaus looked at the time point when participants first fixated certain critical blocks, relative to speech. The pictures on these blocks were designed to create temporary phonological competition (e.g., a *cloud* and a *clown*). Nonetheless, addressees in dialogue interactions looked at the correct picture (e.g., *cloud*) well before their partner reached the point of disambiguation. This contrasted with a control condition, in which the experimenter provided similar descriptions: in this case, as in Allopenna et al., the addressee did not look at the correct picture until after disambiguation.

A second experiment revealed that speakers frequently produced ambiguous descriptions (e.g., *the green piece* when there was more than one green block on the board), yet listeners almost always seemed to know which block they were talking about, fixating only the correct target. The explanation is that speakers considered only certain areas of the board as belonging to the relevant referential domain. Therefore they tended only to include additional specification in their descriptions when two conflicting blocks were actually within the same sub-area – and the fact that addressees were usually not troubled by the ambiguity shows that they had narrowed their referential domain in the same way. This result can be interpreted as a strong effect of common ground on both formulation and interpretation, but it
also presents striking evidence for the alignment of the interlocutors’ perspectives, in a way which could not be shown in monologue studies.

Brown-Schmidt, Gunlogson, and Tanenhaus (2008) actually compared such a targeted language-game, involving pairs of naïve participants manipulating real-world objects, with a more controlled version of the same task, where the (scripted) experimenter gave instructions concerning items in a computer display. These experiments provide an interesting extension of the common-ground discussion: the authors reason that if interlocutors share a joint goal requiring the exchange of information, it is actually appropriate (rather than egocentric) to consider privileged information when preparing to answer a question. After all, in a realistic setting, it is unlikely that the speaker would ask a question if he already knew the answer. Indeed, almost all recorded instances of wh-questions referred to information that was available only to the addressee (e.g., what’s above the cow with the shoes?). And in both experiments, addressees rapidly used their knowledge of both visual and linguistic common ground to direct their attention to the relevant privileged information.

Thus, Brown-Schmidt et al. conclude that there is no simple distinction between the strategies of egocentric processing or considering common ground. Instead, in situations where interlocutors share joint goals and can exchange information freely, “...an addressee who is sensitive to the perspective of the speaker will attend to both shared and private information, at different times and in different linguistic contexts” (Brown-Schmidt et. al.2008, p. 1133).

In sum, common ground clearly does play a rapid role in reference resolution, but at the same time numerous studies have shown that interlocutors are by no means limited to optimal design considerations. As this section has illustrated, reference resolution is one area where findings and methods from monologue studies – in particular the visual-world paradigm – have been successfully transferred to more interactive and world-situated contexts. These, in turn, have revealed just how finely tuned the coordination between interlocutors in the joint activity of dialogue can become.

However, such studies use eye movements only as indicators of what the speaker or listener is currently attending to. The following section summarises
research using eye movements in dialogue for a rather different purpose: instead of considering how soon a particular object is fixated depending on the experimental condition, comparisons of the overall pattern of looks between interlocutors reflect the extent of their alignment.

1.4.4 Can eye movements reflect alignment?

The reviews of eye movement research in production and comprehension (in Sections 1.2 and 1.3 above) revealed that eye movements are in general closely linked to the utterance being produced or understood. Consequently, if dialogue is a joint activity – comparable to ballroom dancing or playing a duet – it seems plausible that the coordination of linguistic moves between interlocutors would be reflected in their eye movements as well.

Richardson and Dale (2005; Richardson, Dale, & Kirkham, 2007) propose a novel form of analysis aimed at quantifying the coordination of visual attention between individuals. Thus, Richardson and Dale (2005) recorded eye movements of speakers describing an array of characters from familiar TV programs, The Simpsons and Friends. Using a cross-recurrence analysis method (Eckmann, Oliffson Kamphorst, & Ruelle, 1987), these gaze recordings were compared with recordings of listeners who heard the description while looking at the same array. Across time, listeners’ patterns of fixations were most similar to the fixations that the speakers had made about two seconds earlier, suggesting that listeners’ eye movements did indeed repeat the speaker’s pattern of gazes.

Interestingly, applying the same method of analysis in a subsequent study where two interlocutors sat in different eyetracking labs and communicated via telephone, Richardson, Dale, and Kirkham (2007) found the greatest overlap of fixations without any lag at all. They suggest that this is because participants in dialogue take turns at speaking and listening, and correspondingly, their eye movements switch between leading and following their partner’s, creating an average lag of zero. An alternative explanation, in line with the Brown-Schmidt and Tanenhaus (2008) data discussed above, is that precisely because participants are
interacting in real-time, they are able to restrict their referential domains and coordinate more closely.

These studies provide a fascinating insight into a further form of alignment between interlocutors in dialogue, and to my knowledge, they are the only investigations to date of how joint attention develops over time. However, they also leave many interesting questions unanswered, in particular with regard to whether and how such coordinated attention relates to task success – indeed, Richardson and Dale (2005) argue that it is a causal factor in achieving understanding. However, they also point out that previous – and potentially shared – knowledge must also play a role, as must the available visual world. I return to some of these questions in later chapters.

1.4.5 **Summary: Dialogue**

If the Richardson et al. (2007) study reveals one thing clearly, it is just how tightly linked interlocutors can become in dialogue – this coordination affects not just their word choice, syntax, and conceptualisation, but also the allocation of attention and presumably many other levels of cognitive processing as well (Pickering & Garrod, 2004). In the previous sections, I hope to have shown that the constant give and take between interlocutors is due to the fact that they are collaboratively engaged in a joint activity and share responsibility for its success (Clark, 1992; 1996).

This means that in order to advance our knowledge of how humans use language, it is essential to supplement traditional monologue research on production and comprehension with investigations of more interactive language use. Recording participants’ eye movements during real-time interactions seems to be a method that is well-suited to such investigations. On the one hand, visual-world studies have shown that fixation behaviour is tightly time-locked to speech and sensitive to underlying cognitive processes, while on the other hand, modern recording systems are fairly unobtrusive and unlikely to interfere substantially with communication (Tanenhaus & Trueswell, 2005). Some of the paradigms outlined in
the section on reference resolution (Section 1.4.3) constitute an initial move in this direction.

Yet this kind of research also faces serious methodological challenges. Very few studies to date have actually recorded eye movements from two (or more) interlocutors in unscripted conversation – partly for technical reasons, but also because of the lack of experimental control and resulting difficulties with analysing the data. At the same time, this is not as critical as it may at first appear, since dialogue and monologue should not be seen as categorically different modes of language. Instead, language is used in many different ways, including casual conversation, formal interactions, and written narratives. The problem is that traditional monologue studies all investigate one extreme form. In contrast, the research reviewed in the previous sections can be situated at different points along this continuum, thus providing insights into language use in contextualised settings that more closely resemble the real world. Correspondingly, my primary aim in the experiments described in the remainder of this thesis has not been to investigate real-time interactive conversation, but rather to isolate certain aspects of language and attention in comprehension and production, and to study their respective effects on the complementary task. The final section of this chapter outlines these experiments.

### 1.5 Overview of the experiments

Broadly speaking, this thesis is concerned with eye movements related to spoken language during both production and comprehension. Moreover, I am interested in the role played by visual attention in mediating between the two modalities in interactive contexts. For instance, to what extent are fixation patterns similar for speaking and hearing analogous sentences, and can the allocation of visual attention during comprehension influence gaze or sentence structure in production?

These issues are explored in Chapter 2: in Experiment 1, participants’ eye movements were recorded while they performed a syntactic priming task with a scripted confederate (cf. Branigan et al., 2000). This setup made it possible to
compare the factors affecting fixation patterns while listening to a prime (e.g., *the policeman is following the waitress*) with those when participants subsequently produced a very similar sentence (e.g., *the soldier is following the sailor*). Eye movements were not particularly similar in the two situations, but there was an effect of priming: they were somewhat more similar when the prime and target used the same sentence structure – i.e. if interlocutors were aligned at a syntactic level – than when they did not.

Experiments 2-7 explored the interplay of vision and language from the other direction: whether comprehension can be facilitated by knowing where a speaker was looking during production (all speakers were unscripted). To introduce this *gaze-projection paradigm*, Chapter 3 first provides a brief introduction to how gaze information is used in face-to-face conversation, and how this transfers to studies where gaze is represented by an on-screen cursor. The first experiment in this series, Experiment 2, is described in detail (Chapter 4), since the materials and general procedure were reused in subsequent experiments. It revealed that seeing a gaze cursor does indeed assist comprehension, so the following experiments examine how this facilitation is achieved, and what affects it. Experiment 3 (Chapter 5) manipulated participants’ belief about what the cursor represented. Experiment 4 threw up a number of methodological issues, discussed in Chapter 6. Using a modified version of the paradigm, Experiment 5 then compared the separate and combined usefulness of the visual and linguistic information (Chapter 7). Finally, Experiments 6 and 7 (Chapter 8) varied the precise time-lag between speech and gaze, as well as addressing the question of how disturbing an unconnected ‘random’ cursor would be. They provide evidence that participants are able to choose whether they attend to the bright moving dot, benefitting from the information it contains when it relates to what is being said, but able to ignore it if it isn’t helpful.

In Chapter 9, the concluding discussion begins by summarising and evaluating the results of the gaze-projection experiments. It addresses both the merits and the limitations of the paradigm, and suggests possibilities for future applications. The scope of the chapter then broadens to integrate the findings of the priming study, and to draw some more general conclusions about the interplay of speech and visual attention in interactive language use.
Chapter 2: Syntactic Priming and Eye Movements in Comprehension and Production – Experiment 1

2.1 Introduction

The experiment described in this chapter formed part of a larger investigation into the coordination of eye movements in a syntactic priming task (Arai, Kreysa, Haywood, & Pickering, 2009). The experiment is included in this thesis because of the light it throws on factors that influence eye movements in comprehension and production. In this context, the syntactic priming manipulation should be regarded mainly as a means of eliciting similar sentence structures in production to the ones participants heard during comprehension.

Nonetheless, I begin the chapter by revisiting the phenomenon of syntactic priming in a little more detail. In particular, Section 2.1.1.2 introduces the confederate-scripting paradigm. I then briefly outline two plausible accounts of how eye movements of speakers and listeners might relate to each other in dialogue situations (2.1.2). This is followed by the predictions of primary interest for the thesis, and a description of the actual experiment (2.2). Section 2.3.2 then briefly summarises and compares the results of the other two experiments in the project. Although these do not form part of the thesis, the summary is necessary for the conclusion in Section 2.4, which takes into account the findings of all three studies.

2.1.1 Syntactic priming in an experimental context: Overview of findings

The phenomenon of syntactic priming (also known as structural priming or structural persistence) has already been mentioned in Chapter 1 (Section 1.4.2). It
Chapter 2: Eyepriming - Experiment 1

refers to a widespread tendency to re-use recent sentence structures, and to the facilitated processing resulting from such repetition. Observational and corpus studies (e.g., Tannen, 1987, and Gries, 2005, respectively) show that sentence structures are repeated frequently. Experimentally, syntactic priming has been shown to occur from production to production (e.g., Bock, 1986b), from comprehension to comprehension (Arai, van Gompel, & Scheepers, 2007; Ledoux, Traxler, & Swaab, 2007; Thothathiri & Snedeker, 2008a; Traxler & Tooley, 2008), as well as between the two modalities (see Branigan, Pickering, & McLean, 2005, for experiments testing various directions of priming). It has been demonstrated in a number of different tasks, in written as well as in spoken language (e.g., Hartsuiker, Bernolet, Schoonbaert, Speybroeck, & Vanderelst, 2008; Kaschak & Borrego, 2008), and even based on single word primes (Melinger & Dobel, 2005). Syntactic priming can affect many different syntactic structures, for example the dative alternation (Branigan et al., 2000), active vs. passive voice (Bock, Loebell, & Morey, 1992), attachment of relative clauses and prepositional phrases (e.g., Haywood, Pickering, & Branigan, 2005; Desmet & Declercq, 2006; Traxler, 2008), as well as adjective modification (Cleland & Pickering, 2003). Finally, syntactic priming has been shown for a variety of different languages, and for different groups of speakers, such as children (Huttenlocher, Vasilyeva, & Shimpi, 2004; Thothathiri & Snedeker, 2008b), bilinguals (Schoonbaert, Hartsuiker, & Pickering, 2007), and aphasic patients (Hartsuiker & Kolk, 1998). Clearly, this is a prolific area of research, for which Pickering and V. S. Ferreira (2008) provide an up-to-date overview.

In the following, I focus on those aspects of syntactic priming research most relevant to this thesis. I begin by contrasting current proposals concerning the function and mechanism of syntactic priming. The subsequent discussion of the theoretical and practical importance of syntactic priming for dialogue research then provides an opportunity to introduce the confederate-scripting paradigm, which was used in Experiment 1.

2.1.1.1 Functional accounts of syntactic priming

According to V. S. Ferreira and Bock’s (2006) discussion of conflicting theoretical accounts of syntactic priming, several alternative explanations for the phenomenon can be distinguished. According to one account, syntactic priming
stems from activation (Pickering & Branigan, 1998; Pickering, Branigan, Cleland, & Stewart, 2000a; Cleland & Pickering, 2003; see also Traxler & Tooley, 2008) or alignment (Pickering & Garrod, 2004). Others see its primary purpose in promoting fluency and reducing production effort (Levelt & Kelter, 1982; Smith & Wheeldon, 2001; Corley & Scheepers, 2002), or describe it as a form of implicit learning (Bock & Griffin, 2000; Chang, Dell, Bock, & Griffin, 2000; Chang, Dell, & Bock, 2006).

The activation account extends Levelt, Roelofs and Meyer’s (1999) model of lexical access in language production, which was described in Chapter 1 (Section 1.2.1). Pickering and Branigan (1998) propose that lemma nodes are connected not just to conceptual and word-form nodes, but also to combinatorial nodes. These are shared between lemmas and become activated when a word is used in a particular construction. For instance, hearing or producing a sentence such as (15) would activate both the lemma give, as well as the combinatorial node associated with the double-object construction. Later, residual activation of this node would make it more likely that the double-object structure would be selected for producing a subsequent sentence such as (16), compared to the prepositional dative in (17).

(15) The man gives the dog a bone.

(16) The woman gives the boy a present.

(17) The woman gives a present to the boy.

In addition to incorporating the explanation of syntactic priming in a more general theory of language production, one of the main strengths of the activation account is that it can easily account for the phenomenon of lexical boost. It is a frequent finding that although syntactic priming occurs in the absence of lexical repetition, re-using prime words in the target sentence generally enhances the priming effect, especially if the verb is repeated (Branigan et al., 2000; Gries, 2005; Arai et al., 2007; Schoonbaert et al., 2007; Thothathiri & Snedeker, 2008b; Traxler & Tooley, 2008). Following Pickering and Branigan (1998), this boost occurs because of residual activation from two complementary sources: in the example above, from the lemma give itself, as well as from the combinatorial node for the double-object construction.

According to Pickering and Garrod’s (2004) Interactive Alignment model, this kind of priming serves a communicative function. As described in Chapter 1
(Section 1.4.2), Pickering and Garrod claim that interlocutors in dialogue align their linguistic representations at many different levels, and that this is a key to communicative success. According to their model, priming is an important and automatic mechanism for creating such alignment across different levels. In fact, following Pickering and Garrod (2007), comprehension-to-production priming could also be useful in predicting upcoming linguistic material, thereby providing a basis for the rapid switches between the two tasks in conversation.

A different facilitatory function of syntactic priming is proposed by the *effort-reduction* hypothesis. This explanation was first suggested by Levelt and Kelter (1982; see also Tannen, 1987) and holds that the “function of syntactic persistence is to reduce the processing costs of the speaker” (Smith & Wheeldon, 2001, p. 159). In essence, re-using a previous structure allows speakers to cut down on the time and cognitive resources necessary for syntactic planning before they begin articulation, as shown in reduced speech-onset latencies for primed structures (Smith & Wheeldon, 2001).

A rather different view of syntactic priming is held by Chang and colleagues (e.g., Chang et al., 2000; Chang et al., 2006). They contend that syntactic priming reflects *implicit learning* of structure-meaning relationships. For instance, a particular structure (e.g., a prepositional dative) is encountered in contexts where an agent transfers an object to a recipient, and with verbs expressing such events. Frequent co-occurrence of event representation and construction (relative to the frequency with which they do not co-occur) promotes the acquisition of a strong link between the two, which in turn can result in syntactic priming. In this account, one of the primary functions of syntactic priming is in language acquisition: basically, a child acquiring this kind of form-meaning mapping would easily learn to comprehend and produce novel sentence structures correctly (Savage, Lieven, Theakston, & Tomasello, 2006). In addition, an implicit learning explanation seems likely for findings of persistent syntactic priming over extended periods of time (e.g., Bock, Dell, Chang, & Onishi, 2007; Hartsuiker et al., 2008; Kaschak & Borreggine, 2008). In contrast, one phenomenon not easily explained by implicit learning is lexical boost.

V. S. Ferreira and Bock (2006) therefore conclude that the different accounts may in fact be complimentary (see also Branigan, 2007; Hartsuiker et al., 2008). In
their view, syntactic priming could have multiple cognitive bases and fulfil several functions:

In particular, it may be that structural priming has both a longer-term manifestation, critical to its function as a reflection of implicit learning, and a shorter-term manifestation, critical to its function in alignment. (V. S. Ferreira & Bock, 2006, p. 1023)

2.1.1.2 Syntactic priming in dialogue: The confederate-scripting paradigm

The experiment described here is concerned only with the latter, short-term form of syntactic priming in a dialogue context. As described in Chapter 1, interlocutors in dialogue show a strong tendency to coordinate their utterances at the semantic and conceptual level (Garrod & Anderson, 1987; Schober, 1993; Brown-Schmidt & Tanenhaus, 2008), as well as with regard to the lexical expressions they use (Clark & Wilkes-Gibbs, 1986; Brennan & Clark, 1996). Evidence of assimilation between interlocutors has even been found at the level of articulation (Pardo, 2006).

Following Pickering and Garrod (2004), this alignment at multiple levels is no coincidence, since alignment at one level is likely to enhance alignment at another, even without deliberate effort by the interlocutors. Often, this is achieved through priming:

In this case, hearing an utterance that activates a particular aspect of a situation model will make it more likely that the person will use an utterance consistent with that aspect of the model. (Pickering & Garrod, 2004, p. 173)

Obviously, syntax is an important linguistic level of representation; so if priming is so important for dialogue and so widespread at and between other levels, syntactic priming should be particularly frequent in interactive contexts. This is also suggested by the high incidence of structural repetition in natural conversations (Weiner & Labov, 1983; Tannen, 1989) and a tendency for studies of syntactic priming between interlocutors to find larger effects than in monologue tasks (cf. Branigan et al., 2000).

The first systematic investigation of syntactic priming in dialogue was a picture description task reported by Branigan, Pickering, and Cleland (2000). To ensure that participants were confronted with equal numbers of PO- and DO-prime sentences – highly unlikely in unscripted dialogue between naïve interlocutors – they employed a confederate of the experimenter. She was treated as if she were a
real participant\(^6\) in every respect, so the actual participants received the impression they were taking part in a picture-matching task with another naïve student. Due to a screen between them, they could not see that the confederate had a script specifying the structure she should use on each trial. Confederate and participant took turns describing pictures to each other (e.g., the cowboy offering the banana to the burglar) and selecting the corresponding picture from their cards, with confederate’s utterances always constituting the prime and participants’ utterances the target sentences. Branigan et al. (2000) found 55 % more repeated than non-repeated structures when the same verb was used for prime and target (lexical boost), and 26 % more for different verbs. In both cases, the amount of priming was greater than in a similar monologue task (Pickering & Branigan, 1998: 17 % more coordination for repeated verbs, 4 % for different; but note that this was a sentence completion task rather than picture description).

In order to assess the importance of the participant’s role, Branigan, Pickering, McLean, and Cleland (2007) extended this confederate-scripting paradigm by including a second confederate. They found greater syntactic priming if participants had been personally addressed in the previous turn than if they had just listened to the utterance as a side-participant. Branigan et al. account for this finding by suggesting that addressees encode more deeply than side-participants. This could be because addressees are required to collaborate with the speaker to achieve understanding (Schober & Clark, 1989; Wilkes-Gibbs & Clark, 1992), or because they know that they will have to respond to the speaker’s utterance and so attend more closely. It seems likely that similar motivations contribute to the generally higher frequency of repeating syntactic structure in dialogue contexts compared to monologue tasks. These findings clearly have interesting theoretical implications in supporting the Interactive Alignment model. At a practical level, they also allow the use of the confederate-scripting paradigm to create an interactive context, thereby increasing the strength of the priming effect under examination (e.g., Cleland & Pickering, 2003; Haywood et al., 2005; McDonough, 2006).

---

\(^6\) Unless otherwise specified, in this chapter the term participant denotes ‘real’ participants who are naïve to the purpose of the experiment, in contrast to scripted confederates.
In this thesis, Experiment 1 compared participants’ eye movements during production and comprehension of active and passive sentences. To ensure that participants produced a sufficient number of passive sentences for a meaningful comparison, comprehension trials were spoken by a scripted confederate and consisted of an equal number of active and passive primes. The participants’ task was to determine whether the sentence matched a picture displayed on their screen, and then to describe another picture back to the confederate, thereby producing the target sentence. The primary purpose of using both the confederate-scripting paradigm and repeated verbs (for lexical boost) was to increase the priming effect and obtain eye movements from a reasonable number of passive production trials (for most verbs with human agents, active is the unmarked sentence structure and therefore preferred over passive; Greenberg, 1966; Slobin & Bever, 1982; Weiner & Labov, 1983; cf. Bock, 1986b). Before outlining the specific hypotheses for this experiment, in the next section I present the rationale of the larger research project that it contributed to.

2.1.2 Coordination vs. dissociation accounts of eye movements in dialogue

Regarding the general question of how people’s eye movements relate to what they say and hear, a series of three experiments (conducted in cooperation with Manabu Arai) served to distinguish between a coordination and a dissociation account.

The assumption that eye movements in comprehension and production would be coordinated is based on the eye-mind assumption (e.g., Just & Carpenter, 1980), according to which eye movements are moment-by-moment indicators of cognitive processing. If what people look at is closely related to what they are thinking about, they could be expected to make very similar eye movements when describing a scene and when listening to a comparable description of that scene.
Roughly speaking, according to the *coordination account*, eye movements in listening should be a delayed reflection of eye movements in speaking.\(^7\)

An additional argument for a coordination account comes from the Interactive Alignment model (Pickering & Garrod, 2004). As described above, this model assumes that dialogue leads to alignment at many different levels of linguistic representation, with alignment at one level promoting alignment at others. Although a visual-attentional level of representation is not itself part of communication and therefore does not feature in the original model, the link between visual attention with semantic and situational representations would make it a plausible additional candidate for alignment. For example, it is easy to see how interlocutors would be more likely to agree on which maze opening was “to the left” or to name an ambiguous tangram figure an “ice-skater” if both of them were simultaneously looking at the same location. Visual coordination therefore seems likely to increase coordination at other levels. Similarly, interlocutors sharing representations at other levels might be expected show an increase in visual coordination. Support for this account comes from the Richardson and Dale studies described in Chapter 1 (Section 1.4.4). In a monologue setting, listeners’ gaze most closely resembled where speakers had looked about two seconds earlier, relative to speech (Richardson & Dale, 2005). However, during interactive telephone conversations about the same stimuli, the greatest overlap of fixations was found without any lag at all (Richardson et al., 2007). On the basis of these results, the coordination account predicts listeners’ eye movements to overlap with speakers’ either simultaneously, or with a delay of up to two seconds.

Alternatively, the *dissociation account* of eye movements in comprehension and production states that while coordinated visual processing may support linguistic alignment, visual attention does not itself constitute a linguistic level of representation and thus is not critical to the success of a conversation. Therefore, eye movements in production and comprehension could frequently dissociate: even when a speaker and an addressee process the same description of the same scene and achieve the same understanding of it, differences in their communicative goals

\(^7\) The delay is due to the general lag in production between looking at an entity and describing it (the *eye-voice span*) and the corresponding lag in comprehension between hearing the description of an entity and looking at it (the *voice-eye span*).
may mean that they look at the scene in a different way. In addition, while speakers fairly consistently look at every entity just before describing it (Griffin & Bock, 2000; but see also van der Meulen et al., 2001), there is good evidence that listeners’ fixations do not just follow what is being mentioned, but frequently go to the entity they predict will be described next (Altmann & Kamide, 1999; Chambers & San Juan, 2008).

In order to differentiate between the two accounts, we conducted three experiments in which people comprehended and produced descriptions of events. By using a confederate-scripting paradigm, we were able to record participants’ eye movements during comprehension and production trials within the same experiment, and in response to the same stimuli. Two of the experiments investigated syntactic priming of the dative alternation (verbs were repeated in Experiment A, but differed between prime and target in Experiment B)\(^8\), while Experiment 1 considered priming of active or passive voice with repeated transitive verbs.

As described in Section 2.1.1.2, the syntactic priming manipulation provided some control over the relationship between comprehended and produced utterances and their syntactic structure, as well as making it possible to assess the extent of linguistic alignment. Visual coordination during those trials in which syntactic priming occurred would suggest that visual representations are implicated in interactive alignment, as suggested by the coordination account.

A secondary question addressed in these experiments was whether there would be an effect of priming on the eye movements themselves. Supporting the theoretical position that syntactic priming reduces processing effort and promotes fluency (see Section 2.1.1.1), there is some evidence that syntactic priming can facilitate speech production (e.g., Smith & Wheeldon, 2001). We therefore also compared speech onset times and first fixation latencies for coordinated and non-coordinated target utterances.

\(^8\) Experiments A and B do not form part of this thesis, so they are referred to by capital letters instead of numbers. They are summarised in Section 2.3.2.
2.2 Experiment 1: Eye movements during comprehension and production of related sentences

While an arbitration between the two accounts of speakers’ and listeners’ eye movements in dialogue formed the general theoretical motivation for Experiment 1, the study is included in this thesis primarily because of the relevance it bears to eye movements in production and comprehension. The following sections will therefore concentrate on these aspects of the study.

2.2.1 Predictions regarding eye movements in comprehension and production

Active and passive sentences are semantically similar, though they differ in focus and emphasis to a greater extent than, for instance, prepositional and double object sentences (cf. Bock & Griffin, 2000). More importantly for this experiment, they contain only two arguments (agent and patient) and their word order differs from the beginning of the sentence. In comprehension, participants cannot know which structure is going to be used until they hear the first entity mentioned. Eye movements relating to the first noun phrase should therefore be reactive, probably reaching this entity shortly after the onset of the first noun (N1). However, as soon as the referent of the first noun is clear, so is the second, since there are only two entities on the screen. Eye movements related to N2 may therefore be anticipatory, beginning between N1 and N2 onsets.

In production, by contrast, eye movements are expected to consistently precede the mention of each entity, in line with phonological retrieval (Griffin, 2001). Thus, active sentences should show increased fixation of the agent before speech onset, followed by fixations to the patient, while passive sentences should follow the opposite pattern. However, if factors not directly related to the production process can influence these name-related gazes, differences in fixation onset or fixation likelihood could emerge due to priming, structural preference, or display characteristics (cf. Gleitman et al., 2007). For instance, if primed structures
are generally easier to produce, pre-speech fixations to the N1 might be earlier or shorter.

2.2.2 Methods

2.2.2.1 Materials

Figure 2-1 provides an example of the prime and target images, which were adapted from the black-and-white line drawings used by Branigan et al. (2000). For a list of experimental items, see Appendix A.

![Figure 2-1: Example of a prime or target image in Experiment 1.](image)

All prime and target images depicted actions which could be described using one of twelve transitive verbs, with two easily nameable cartoon characters in the roles of agent and patient. Each verb was used for a total of four images in the course of the experiment, involving different pairs of characters. Verbs were always repeated between prime and target, but the entities performing or undergoing the action were not. Each prime picture was paired with one of two types of prime sentence: an active description such as (18) and a passive description such as (19).

(18) The policeman is chasing the swimmer.

(19) The swimmer is being chased by the policeman.
The 24 prime-target pairs were supplemented by 72 pairs of filler pictures, which could always be described using an intransitive structure. For fillers, verbs were only repeated in one third of trials (24 prime-target pairs). Half of all pictures matched the descriptions they were paired with, as participants’ task while listening to prime sentences was to judge whether they accurately described the picture on the screen. The mismatching picture items were created by reversing the roles of agent and patient (50 % of mismatching items) or by replacing the agent or patient with a different character (25 % each). This could lead to sentences such as (20) or (21), respectively, for describing Figure 2-1:

(20) The swimmer is chasing the policeman.

(21) The waitress is being chased by the policeman.

Each picture also came in a mirror-image version, which was created by flipping it horizontally, reversing the positions of the agent and patient. In this way, the agent appeared on the left side of the screen in exactly half the prime trials. This counterbalancing is important because of a general bias to conceive of action events as occurring from left to right, at least in cultures with left-right writing systems (Maass & Russo, 2003; Dobel et al., 2007a). However, in an effort to encourage participants to produce passive sentences, the orientation of target images was kept constant, such that the patient always appeared on the left. All fillers appeared in a mirror-reversed version too, though in this case all that changed was the direction in which the character was facing (it was usually depicted roughly in the centre of the screen).

The black line drawings on white backgrounds were stored in jpeg format. They measured 1024 × 768 pixels and were presented at a screen refresh rate of 75 Hz on a grey screen background.

2.2.2.2 Design

Two experimental lists were created, with an equal number of participants assigned to each list. On each list, half the prime pictures matched their description, half did not; half showed the agent on the left, the other half on the right. As mentioned above, the verb was always repeated between prime and target in experimental trials. The actual item order was pseudo-randomised for each
experimental session, with the constraint that at least two pairs of fillers intervened between experimental items.

The main manipulation was the type of prime structure spoken by the confederate and comprehended by the participant (active or passive). Additional independent variables were whether prime picture matched the prime description, and its orientation (agent left or right); all were varied within participants. The only between-participants variable was the identity of the confederate.

2.2.2.3 Participants

Twenty members of the University of Edinburgh student community were paid £6 to take part in Experiment 1, which was conducted at the same time as Experiment A (see Section 2.3.2). Both experiments used the same instructions, confederates, and experimenters. Participants who signed up to take part were assigned to the two experiments alternately. All were native speakers of English with normal or corrected-to-normal vision. Their mean age was 20.5 years; six out of 20 were male. Invalid data from two further participants had to be replaced, one due to experimental error and one because she suspected the purpose of the experiment.

2.2.2.4 Apparatus and Confederates

The experiment took place in a quiet basement lab (the Edinburgh Joint Eyetracking lab, JEL). It consisted of a sound-proof container fitted with two 22’ cathode ray tube monitors set back-to-back (resolution 1024 × 768 pixel, 75 Hz refresh rate). Each was equipped with a height-adjustable chair at about 90 cm viewing distance, a keyboard, a clip-on microphone, a Microsoft Sidewinder® gamepad, and an EyeLink 2 head-mounted eyetracker (SR Research, Canada). This tracker is designed to allow a certain amount of head movement – important for language production studies – while maintaining a sampling rate of 500 Hz and spatial resolution less than 0.01°. For monocular tracking, each participant’s dominant eye was determined using a simple parallax test at the start of the experimental session, though viewing was binocular. Stimulus presentation was controlled by Experiment Builder (SR Research), as were the recordings.
This setup meant that the confederate could arrive with the participant, sit at the second computer screen in the same room, and be fitted with an eyetracker (though confederates’ eye movements were not actually recorded). It also meant that the confederate could interact naturally with the participant before and after the main experiment, thus increasing the plausibility of the cover story. The two confederates were both female postgraduate students, one Scottish and one English, and they participated in ten valid experimental sessions each, roughly alternately.

### 2.2.2.5 Procedure

Participants were told that the experiment would investigate eye movements while switching between the tasks of speaking and listening. Together with the confederate, they were asked to take turns describing pictures to each other, and to decide whether the other’s description matched the image they could see on their screen. For their own descriptions, they were instructed to use the verbs presented on-screen just before the start of a trial. Seeing such a verb would indicate that it was their turn to describe the next image. They received no further instructions with regard to what to say.

The setup of the eyetrackers and a nine-point calibration routine was followed by four practice trials, all using intransitive verbs. An automatic drift correction was performed at the start of each trial, which made it possible to recalibrate if recording accuracy was reduced at any time. Full calibration was repeated halfway through the experiment. This also gave participants the opportunity for a short break.

In prime trials, the confederate saw a script of the prime sentence displayed below the prime image. The image simultaneously appeared on the participant’s screen, but since the confederate then began speaking, the participant would assume that the verb had appeared on her screen (speech onset by the confederate was not controlled, but occurred fairly soon after the appearance of the picture). The participant therefore pressed one of two keys on the gamepad to indicate whether the description matched the image. In the subsequent target trial, the verb was displayed to the participant for 1500 ms before the target image appeared; in response to the description, the confederate also pressed a button on her gamepad. The presentation of the images was synchronised between the confederate’s and the
participant’s screens, and the recording of the participant’s eye movements began with the appearance of the image in all trials. The image was removed as soon as a response was registered by the gamepad of the person in the listening role. If no response was made, image presentation terminated automatically after seven seconds.

After the experiment, participants were asked to fill in a questionnaire. In addition to some demographic questions, this was intended to assess whether they had noticed the purpose of the experiment. The whole procedure typically lasted about one hour.

2.2.3 **Analysis**

The experimental trials produced two types of data: audio recordings of confederates’ prime and participants’ target descriptions, as well as participants’ eye movement data recorded during comprehension and production.

2.2.3.1 **Data treatment and coding**

Sentences were coded as active if the first constituent was an agent followed by a main verb and the second constituent was a patient. They were coded as passive if the first constituent was a patient followed by a passive verb (be and past participle verb) and a by-phrase containing an agent. Four prime trials could not be used because of confederates’ errors or outside noise. This also led to the exclusion of the corresponding target trials. In addition, 25 target trials were excluded because participants failed to respond, or because their descriptions didn’t meet our criteria for active or passive structures. Of the remaining 451 target utterances, 326 were active and 125 (28%) were passive descriptions. Participants generally judged accurately whether their image matched the confederate’s utterance (90.5%).

In order to analyse the eye movement data with reference to the timing of each speech utterance (whether comprehended or produced), all audio files were transcribed and segmented manually to mark speech onset, verb onset, and the onsets of the first and second noun, respectively. Speech onset was defined as the
onset of the determiner of the subject argument, meaning that any expressions produced before the determiners were ignored (e.g., “uh”, “um”).

For the eye movement data, the analysis software (EyeLink DataViewer, SR Research) automatically classified each fixation depending on which of three manually predefined regions of interest (ROI) it fell in. Two of these ROI corresponded to the agent and patient, extending around their contours by about 30 pixels; the third ROI captured all looks to the screen background.

Analyses were run on empirical logits of gazes, with a gaze defined as the accumulation of all successive fixations to a particular ROI before a fixation to another ROI (for more detail on the logit transformation, see Section 2.2.3.2 below). Gazes to each ROI were assessed in time slices of 20 ms from the relevant auditory onset, and aggregated into larger time windows for the inferential analyses.

In comprehension, eye movements were expected to respond to the unfolding referring expression, so gazes were primarily analysed in the two time windows between 200 ms and 600 ms following the onsets of the N1 and N2 in each trial. The starting point of this interval was chosen because it takes approximately 200 ms to program a saccadic eye movement (Matin, Shao, & Boff, 1993; though see also Altmann & Kamide, 2004). Thus, the referring expression is unlikely to have much effect on eye movements during the first 200 ms following its onset – any effects that are visible at this point can therefore be attributed to anticipatory processing. Similarly, the end point of the window ensured that effects were unlikely to be affected by following words (the mean interval between the N1 and verb onsets was 543 ms). To account for developments over time, the four 100 ms time bins within each 400 ms window were included in the analysis in the form of a covariate. This made it possible to examine whether participants were already looking at an entity at the start of a window (implying an anticipatory effect), or whether they looked at it increasingly over time after hearing its name. The latter finding would be a rate effect, reflected in the covariate and its interactions with other factors (see Barr, 2008a).

---

9 Note that a further 200 ms could be added to this interval before eye movements would be affected by the verb (Matin et al., 1993).
Similar considerations motivated the selection of the time window ranging from -600 ms to -200 ms before the onset of each noun for the analysis of the production data. Clearly, it made sense to use an interval of the same length as for comprehension to be able to compare the two tasks. In addition, previous research has shown that speakers tend to look at items roughly 1000-800 ms before mentioning them in descriptions, with such name-related looks peaking around 600 ms before mention (Griffin & Bock, 2000). Typically, they then shift their eyes away again about 300-100 ms before mention (Meyer et al., 1998; Meyer & van der Meulen, 2000). Gazes to to-be-mentioned objects in the window between -600 and -200 ms should thus show a consistent decrease, fitting the linearity assumptions of linear mixed effects (LME) models, the form of analysis used for Experiment 1. Finally, the mean interval between the two nouns in the target sentences was always substantially longer than 1000 ms, making it unlikely that eye movements in the 600 ms before N2 onset would still be affected by the articulation of N1.

All analyses were conducted using linear mixed effects models. Since this is still a fairly novel form of analysis in psycholinguistic and eyetracking studies, Section 2.2.3.2 provides a brief general introduction to using and interpreting such models. This is followed by the actual results, beginning with the language-based effects of the priming manipulation on the structure and timing of participants’ target utterances (2.2.4.1). Next, I present separate examinations of their eye movements during comprehension (i.e. prime trials) and production (i.e. target trials) in sections 2.2.4.2 and 2.2.4.3, respectively.

2.2.3.2 Linear mixed effects models and their interpretation

In visual-world paradigm experiments such as the one reported here, eye movements have traditionally been analysed by aggregating proportions of fixations to the ROIs within pre-defined time windows, and then using analysis of variance to test for effects of the manipulations. Recently, this strategy has been questioned for several reasons (e.g., Barr, 2008a; Jaeger, 2008; Mirman, Dixon, & Magnuson, 2008; Quené & van den Bergh, 2008).

---

10 In fact, the study reported here was initially analysed using ANOVA as well, with broadly similar results.
One important criticism is the fact that the assumptions for parametric tests such as ANOVA are blatantly violated by the characteristics of eye movement data. Violations based on the properties of individual datasets are fairly common in psycholinguistic research and are rarely deemed critical (e.g., unbalanced cell frequencies, non-normal distributions, lack of sphericity and homoscedasticity; for an extensive discussion of the importance of these factors see Rietveld & van Hout, 2005). However, ANOVA was originally developed to assess effects of categorical factors on continuous dependent variables, whereas visual-world paradigm studies typically investigate effects of a continuous variable (time) on categorical data (fixation or no fixation to ROI). This creates a violation of assumptions based on the dependent variable itself: observations of the same person’s eye movements over time are clearly not independent, as they are much more likely still to be looking at the same ROI after 20 ms than after 1020 ms.

An additional, more practical criticism of using ANOVA on visual-world paradigm data is the loss of information resulting from the practice of collapsing over time, which leads to a confound of anticipatory and rate effects or even conceals them entirely. This will be discussed in more detail below and when describing the results.

Some solutions which have been put forward for these problems are fitting parametric curves to capture developments over time, transforming the data to log-odds scale to avoid violating assumptions and to deal with multiplicative effects (Barr, 2008a; but see also Tanenhaus, Frank, Jaeger, Pier Salverda, & Masharov, 2008), and including crossed random effects (Baayen, Davidson, & Bates, 2008; Jaeger, 2008). The latter procedure makes it possible to assess the effects of participants and items separately but within the same analysis, thereby avoiding the need for performing and interpreting separate $F1$ and $F2$ analyses (Rietveld & van Hout, 2005; Brysbaert, 2007).\footnote{An alternative solution is to use hierarchical loglinear regression (Knoeferle et al., 2005; Altmann & Kamide, 2007; Knoeferle & Crocker, 2007).}

For these reasons, LME models of the logarithm of odds of a particular occurrence (e.g., a passive target, or a fixation of the agent) were used in all the analyses reported here, following Barr (2008a). As mentioned in Section 2.2.3.1,
gazes were aggregated into larger time windows of 100 ms, and the empirical logit of gazes to each ROI was calculated using Equation 2-1.

**Equation 2-1: Calculation of the empirical logit of gazes:** \( n \) is the sum of all gazes to all ROIs, including the background; \( y \) is the sum of gazes to a particular ROI (Barr, 2008a).

\[
\eta' = \ln\left(\frac{y + 0.5}{n - y + 0.5}\right)
\]

Unlike a probability scale, the logit scale can be infinitively positive or negative. The measure yields a score of zero if gazes to the ROI under examination \( y \) account for half the total gazes compared to all other ROIs. A positive value would mean that more looks were made to that ROI than to all other objects.

The predictor variables entered into the model will be explained separately for each analysis, but the general procedure was to begin with a fully specified model. This model included the factors of interest and any predicted interactions, as well as all binary interactions of factors with a significant main effect. The model was simplified using the Akaike Information Criterion (AIC), a measure for optimising model fit with reference to the amount of variance and the number of degrees of freedom (cf. Demberg & Keller, 2008): in each step, one interaction or simple factor which did not contribute significantly to the model was removed. If this procedure led to a reduction of the AIC, contributions of the remaining factors were recalculated, until all factors remaining in the model improved its fit significantly. Significant coefficients of the remaining factors were then interpreted in a similar way to significant effects in an analysis of variance: depending on how the factor levels were originally contrast-coded, the sign of each coefficient reveals whether more or less gazes occurred to the entity of interest (e.g., the agent) than to all other ROIs.

Such models allow several different kinds of effect to be distinguished, which I will describe using the example of gazes to the agent (logit-transformed) during comprehension of the first noun (for a more detailed description, see Barr, 2008a). For the analysis of gazes, passive primes and targets were always coded as -0.5, actives as +0.5. With this form of contrast coding, the coefficients’ sign illustrates the effect of that factor when an active structure is heard and/or produced, as the following examples will illustrate. Similarly, an agent on the left was coded as +0.5, an agent on the right as -0.5.
The first component of the optimised model is always the Intercept. This provides information on the extent to which a particular ROI was fixated at the start of the time period, irrespective of any other factors. It can therefore only be affected by what has been processed previously, making it indicative of participants’ expectations and anticipatory processing. In our example, a significant positive Intercept for gazes to the agent would mean that participants were more likely to fixate the agent than any other region even before the onset of the prime sentence.

In contrast, coefficients of simple factors reflect overall preferences for fixating a particular entity when an active is heard/produced, across the time period under examination. For example, a significant positive coefficient for \textsc{Prime Type} would show that during the mention of an agent N1 (i.e. an active prime sentence), participants were more likely to look at the agent than during the mention of a patient N1. Such main effects could be affected by processing occurring during the period of analysis, but the fact that they extend throughout the time period implies that they are generally the result of a correct prediction, which caused participants to shift their gaze to the entity before the onset of its name.

The coefficient for the time window covariate forms a special case, as this comprises the four consecutive 100 ms bins within the larger time period and can therefore signal changes in the rate of fixation, potentially influenced by the content of the unfolding utterance. In our example, a significant positive coefficient for \textsc{Time Window} would reveal that fixations to the agent increased while hearing the N1. Such an increase of gazes to a just-mentioned entity will be termed a \textit{responsive} fixation pattern. It could indicate that participants were not able to predict the upcoming referent, so they fixated it as soon as it was unambiguously mentioned. By contrast, a decrease of gazes to the just-mentioned entity would be \textit{reactive}; it would suggest that the referent had been sufficiently processed, so attention could be directed to the next relevant entity. The \textsc{Time Window} covariate is particularly useful when it interacts with a simple factor, since such an additional interaction can reveal temporal differences in fixation behaviour between conditions. For example, a negative interaction of \textsc{Prime Type} and \textsc{Time Window}, the strength of which exceeds the simple effect of \textsc{Time Window}, would show that fixations to the agent decreased while hearing an active prime.
Finally, interactions between factors can be interpreted in a similar way, by comparing the coefficients for the respective main effects and the interaction. Thus, a positive interaction of PRIME TYPE and IMAGE ORIENTATION would mean that while hearing an active prime, participants fixated the agent more if it appeared on the left side of the screen than on the right.

2.2.4 Results and Discussion

Having explained how the data was analysed and how to interpret the output, the results are presented in the following sections: Section 2.2.4.1 examines the linguistic effects of priming on participants’ speech production, while Sections 2.2.4.2 and 2.2.4.3 are concerned with eye movements during comprehension and production, respectively.

2.2.4.1 Syntactic priming and word onsets

To assess the extent to which participants reproduced the structure used by the confederate, Table 2-1 summarises the number of active and passive target descriptions produced for each prime condition.

<table>
<thead>
<tr>
<th>Prime Condition</th>
<th>Active descriptions</th>
<th>Passive descriptions</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active Prime</td>
<td>221 (97 %)</td>
<td>8 (3 %)</td>
<td>229</td>
</tr>
<tr>
<td>Passive Prime</td>
<td>105 (47 %)</td>
<td>117 (53 %)</td>
<td>222</td>
</tr>
<tr>
<td>Total</td>
<td>326 (72 %)</td>
<td>125 (28 %)</td>
<td>451</td>
</tr>
</tbody>
</table>

The fully specified LME model of passive target descriptions contained the following fixed factors, as well as two random effects for participants and items: PRIME TYPE (active vs. passive), PICTURE MATCH (whether the confederate’s description matched the image seen by the participant), PRIME IMAGE ORIENTATION (agent left vs. agent right), and CONFEDERATE. Table 2-2 summarises the effects of factors remaining in the optimised model, once any that did not contribute significantly had been excluded (as described in Section 2.2.3.2 above).
Table 2-2: Factors affecting participants’ passive target structures, including coefficients and their significance levels in the optimally minimised model.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Coefficient</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-2.10</td>
<td>***</td>
</tr>
<tr>
<td>Prime Type (Active-Passive)</td>
<td>4.05</td>
<td>***</td>
</tr>
<tr>
<td>Confederate</td>
<td>-0.26</td>
<td>-</td>
</tr>
<tr>
<td>Confederate * Prime Type</td>
<td>-0.65</td>
<td>*</td>
</tr>
</tbody>
</table>

*** p < .001; * p < .05; - p > .1

With the contrast-coding scheme used here (-0.5 vs. +0.5), the regression coefficients reported in the table are equal to the differences between the marginal means, and the Intercept corresponds to the overall mean in the log-odds scale (J. Cohen & Cohen, 1983; Barr, 2008a). Its significant negative coefficient reflects a general preference for the active structure (it would be zero if the number of passive descriptions equalled the sum of all other kinds of description); overall, as demonstrated by Table 2-1, participants were more than twice as likely to produce an active than a passive target sentence. In addition, the strong positive effect of PRIME TYPE reflects the fact that participants produced more passive descriptions following passive primes than following active primes. Neither PICTURE MATCH nor PRIME IMAGE ORIENTATION had any effect on participants’ sentence structures. There was no simple effect of CONFEDERATE either, though the extent of priming was influenced by which confederate participants heard – it seems that one of the two primed stronger than the other. Although this finding was not specifically predicted, individual differences in priming are plausible. Hence, CONFEDERATE will be included as a factor in the following gaze analyses.12

Syntactic priming has been reported to reduce the speech onset latency of the target utterance, in addition to its effect on structural selection (Smith & Wheeldon, 2001; Corley & Scheepers, 2002), though the effects have not been replicated reliably (see Pickering & Ferreira, 2008). Therefore, the N1 latency and the time between N1 and N2 onsets were also analysed using LME, although these analyses were limited to active descriptions because of the scarcity of passive targets following active

12 CONFEDERATE did not significantly affect the likelihood of active targets, for which the corresponding optimal model contained only the predictor PRIME TYPE.
primes \((N = 8)\). Table 2-3 shows mean onset times for the remaining three prime-target conditions.

Table 2-3: Overview of onset times by prime-target condition (APAT: active prime & target; PPAT: passive prime, active target; PPPT: passive prime & target), in ms from picture onset.

<table>
<thead>
<tr>
<th>Prime-target condition</th>
<th>Speech onset</th>
<th>N1 onset</th>
<th>is</th>
<th>verb</th>
<th>N2 onset</th>
</tr>
</thead>
<tbody>
<tr>
<td>APAT ((N = 221))</td>
<td>M</td>
<td>1312</td>
<td>1482</td>
<td>1922</td>
<td>2044</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>383</td>
<td>472</td>
<td>490</td>
<td>505</td>
</tr>
<tr>
<td>PPAT ((N = 105))</td>
<td>M</td>
<td>1340</td>
<td>1488</td>
<td>1913</td>
<td>2041</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>367</td>
<td>409</td>
<td>443</td>
<td>454</td>
</tr>
<tr>
<td>PPPT ((N = 117))</td>
<td>M</td>
<td>1388</td>
<td>1532</td>
<td>1924</td>
<td>2285</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>419</td>
<td>456</td>
<td>487</td>
<td>485</td>
</tr>
<tr>
<td>Total ((N = 443))</td>
<td>M</td>
<td>1339</td>
<td>1497</td>
<td>1920</td>
<td>2107</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>389</td>
<td>453</td>
<td>478</td>
<td>498</td>
</tr>
</tbody>
</table>

Only the effect of PRIME TYPE was of interest in these analyses, so no other predictors were included in the models, apart from participants and items as random factors. The analysis of the N1 onset showed an Intercept of 1483 ms, reflecting mean onset time across the two prime conditions, and no effect of PRIME TYPE (Coefficient = 5.84, \(t = 0.16, p > .9\)). Similarly, the time between the N1 and N2 onset was not affected by PRIME TYPE (Coefficient = -3.12, \(t = -0.09, p > .9\)), with an Intercept of 1147 ms.

In short, a substantial effect of syntactic priming was found, with participants’ choice of syntactic structure affected by the prime, as well as by an overall preference for active descriptions. This is in accord with previous studies of syntactic priming in dialogue, such as Branigan et al. (2000). Visual factors such as picture orientation did not interact with the effect of priming, nor did the prime sentence influence the timing of the target sentence. More important for this thesis, however, is the relationship between prime and target type and participants’ eye movements, which will be discussed in the following sections.

---

13 The onset of N1 was used instead of speech onset, because participants frequently extended the determiner “the” while apparently still planning the utterance. However, the same analysis based on the actual onset of speech (generally the determiner) showed very similar results.
2.2.4.2 Eye movements in comprehension: Gaze analysis

Comprehension eye movements were recorded while participants heard the prime sentences. Only trials where the image on the participant’s screen matched the prime description were included in the analysis, to avoid eye movements caused by a mismatch between display and speech. Figure 2-2 shows the logit of gazes to the agent and patient in 20 ms time bins for 2000 ms following the onset of N1, for each prime type.

![Gaze logits to the agent and patient during comprehension of active and passive prime sentences, beginning at N1 onset. The vertical bars show the average onsets of the active and passive N2 separately (M = 1056 ms and M = 1400 ms, respectively, both SD = 130). A logit of zero reflects equal proportions of gazes to the two entities.](image)

The initial LME model of the logit of gazes to the two entities contained PRIME TYPE as the main predictor. In addition, it included the factors IMAGE ORIENTATION and CONFEDERATE, as well as the TIME WINDOW covariate, comprising four 100 ms bins, and all binary interactions. Table 2-4 reports the coefficients that remained in the minimally optimised models (AIC criterion), with their respective significance levels. In the following, these will be described separately for the time periods following N1 and N2.
Table 2-4: Analyses of the logit of gazes while listening to prime sentences in the 400 ms following the N1 and N2 onsets: Coefficients remaining in minimised models and their significance levels.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Coefficient</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Following the N1 (200ms - 600ms after N1 onset)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>0.20</td>
<td>-</td>
</tr>
<tr>
<td>Prime Type</td>
<td>-0.14</td>
<td>-</td>
</tr>
<tr>
<td>Agent logit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time Window</td>
<td>-0.27</td>
<td>***</td>
</tr>
<tr>
<td>Image Orientation</td>
<td>0.66</td>
<td>***</td>
</tr>
<tr>
<td>Prime Type * Time Window</td>
<td>0.27</td>
<td>*</td>
</tr>
<tr>
<td>Patient logit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-0.61</td>
<td>**</td>
</tr>
<tr>
<td>Prime Type</td>
<td>0.14</td>
<td>-</td>
</tr>
<tr>
<td>Time Window</td>
<td>0.25</td>
<td>***</td>
</tr>
<tr>
<td>Image Orientation</td>
<td>-0.59</td>
<td>***</td>
</tr>
<tr>
<td>Prime Type * Time Window</td>
<td>-0.30</td>
<td>*</td>
</tr>
<tr>
<td>Following the N2 (200ms - 600ms after N2 onset)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agent logit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-0.05</td>
<td>-</td>
</tr>
<tr>
<td>Prime Type</td>
<td>-3.68</td>
<td>***</td>
</tr>
<tr>
<td>Time Window</td>
<td>-0.05</td>
<td>-</td>
</tr>
<tr>
<td>Prime Type * Time Window</td>
<td>0.66</td>
<td>***</td>
</tr>
<tr>
<td>Patient logit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-0.20</td>
<td>-</td>
</tr>
<tr>
<td>Prime Type</td>
<td>3.48</td>
<td>***</td>
</tr>
<tr>
<td>Time Window</td>
<td>-0.08</td>
<td>-</td>
</tr>
<tr>
<td>Image Orientation</td>
<td>0.47</td>
<td>***</td>
</tr>
<tr>
<td>Prime Type * Image Orientation</td>
<td>0.29</td>
<td>*</td>
</tr>
<tr>
<td>Prime Type * Time Window</td>
<td>-0.62</td>
<td>***</td>
</tr>
</tbody>
</table>

*** p < .001; ** p < .01; * p < .05; - p > .1

In the interval following the N1, the significant Intercept for the patient logit is an interesting finding. This is an anticipatory effect, showing that at the very start of the relevant time period, participants generally looked less at the patient than at any other region. In contrast, a general preference for looking at the agent before the onset of the confederate’s description was confirmed by a similar analysis on logits of gazes in the period from -600 ms to -200 ms before speech onset, summarised in
In this period, the odds of participants looking at the agent were influenced only by whether it was positioned on the left of the screen, in which case it was fixated more. At the same time, the patient received significantly fewer pre-speech gazes than the agent overall, and particularly if it was on the right, dispreferred side of the screen.

Table 2-5. Analyses of the logit of participants’ gazes immediately preceding the confederate’s speech onset: Coefficients remaining in minimised models and their significance levels.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Coefficient</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preceding N1 (-600ms to -200ms before speech onset)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agent logit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>0.12</td>
<td>-</td>
</tr>
<tr>
<td>Image Orientation</td>
<td>0.57</td>
<td>***</td>
</tr>
<tr>
<td>Patient logit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-1.61</td>
<td>***</td>
</tr>
<tr>
<td>Image Orientation</td>
<td>-0.72</td>
<td>***</td>
</tr>
<tr>
<td>Time Window</td>
<td>0.19</td>
<td>***</td>
</tr>
<tr>
<td>Confederate</td>
<td>0.32</td>
<td>***</td>
</tr>
<tr>
<td>Image Orientation * Time Window</td>
<td>0.20</td>
<td>***</td>
</tr>
</tbody>
</table>

*** p < 0.001; - p > .1

As already mentioned, the initial preference for the agent over the patient persisted in the primary period of analysis following N1 (Intercept in Table 2-4). By contrast, the main effect of TIME WINDOW for both agent and patient looks is a rate effect, showing that looks to the agent generally decreased over time, while looks to the patient increased. More interestingly, an interaction of PRIME TYPE and TIME WINDOW was found for both agent and patient looks. The coefficients suggest that this rate effect was responsive: when the N1 referred to an agent (i.e., in an active prime sentence), looks to the agent stayed roughly constant (note the equal coefficients for TIME WINDOW and its interaction with PRIME TYPE), whereas looks to the patient decreased further. Finally, the main effect of IMAGE ORIENTATION for both agent and patient looks implies that there were more looks to the agent if it was on the left side of the picture and less looks to the patient if it was on the right side, i.e. a general initial preference for the entity on the left.

---

A high proportion of fixations to the background at this early point in time is the reason why this preference does not show up in the Intercept for agent logits following N1 onset.
In the interval following the N2, neither Intercept reached significance, demonstrating that participants didn’t generally favour the agent over the patient. There was a main effect of PRIME TYPE and an interaction between PRIME TYPE and TIME WINDOW for both agent and patient looks; participants generally looked more at whichever entity had just been mentioned (e.g., the patient in an active sentence), but these looks decreased over the 400 ms period. Arguably, the main effect is indicative of predictive processing, as participants who had heard the agent N1 mentioned in an active sentence were able to shift their gaze quickly to the patient N2, reaching it before the onset of the period of analysis and maintaining it throughout this time. A main effect of IMAGE ORIENTATION was found for patient looks only, as well as an interaction of IMAGE ORIENTATION and PRIME TYPE; participants looked at the patient more when it was on the right side of the screen, and especially if they had just heard the patient mentioned.

To summarise, the eye movements observed in comprehension shortly after hearing the first and second noun were affected by the unfolding utterance, but conceptual and visual factors were important predictors as well. In the period following N1, the strongest determinant of gazes was their visual location on the screen (main effect of IMAGE ORIENTATION). Additionally, patients were dispreferred, which may have been due to anticipation of an active description – which is more frequent in everyday language, and even in this experiment – or to the conceptual salience of agenthood (Segalowitz, 1982; Bock, 1986b; Kreysa et al., 2009). The only significant effect related to the actual sentence was the interaction of PRIME TYPE and TIME WINDOW; this reflected a reactive increase in looks to whichever entity had just been mentioned.

Following N2, the most substantial effect was sentence-based: overall, participants tended to fixate the most recently mentioned character. This effect can be interpreted as evidence of predictive processing, since the consistent main effect of PRIME TYPE implies that the referent for the N2 was identified and fixated before the onset of the time window. Again there was a significant interaction of PRIME TYPE and TIME WINDOW, though this time it signalled a gradual decrease of these looks, presumably because the end of the sentence had been reached and participants were preparing their response. However, visual factors also continued
to influence gaze patterns, at least with respect to fixations of the patient (main effect and interaction of IMAGE ORIENTATION).

Accordingly – at least for this type of stimuli and in these time windows – gaze patterns in comprehension seem to be influenced by many different factors simultaneously: sentence-related gazes can be both reactive and predictive, but gaze patterns are also affected significantly by visual and conceptual features.

2.2.4.3 **Eye-movements in production: Gaze analysis**

We now turn to the eye movements recorded during production, i.e. while participants were describing the target pictures. A further 24 trials were excluded from these analyses because participants made either speech corrections or long pauses in the middle of their utterances (> 2 s). Figure 2-3 displays the logit of gazes for each 20 ms time bin over a total time period of 3000 ms, separately for gazes to agent and patient in each prime-target condition. Since practically no passive targets were produced following active primes (see Section 2.2.4.1 above), this condition was not included in Figure 2-3.

![Figure 2-3: Logit of gazes to the agent and patient, depending on the prime-target condition (APAT: active prime & target; PPAT: passive prime, active target; PPPT: passive prime & target). The first vertical bar shows the N1 onset, around which the graph is centred; the second bar shows the mean N2 onset for active, the third for passive descriptions.](image)
Analyses were again conducted using LME, with the following factors and their binary interactions included in the initial model: PRIME TYPE, TARGET TYPE (active vs. passive), PRIME IMAGE ORIENTATION (agent left vs. agent right; note that target image orientation was kept constant), as well as TIME WINDOW as a covariate. Originally, the factors PICTURE MATCH and CONFEDERATE were also included, but they did not contribute significantly to any model. For the optimised models of gaze logits in the relevant time windows, Table 2-6 summarises the coefficients and their significance levels.

Table 2-6: Analyses of the logit of gazes for producing target sentences in the 400 ms time windows preceding the N1 and N2 onsets: Coefficients and their significance levels for optimally minimised models.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Coefficient</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preceding the N1 (-600ms - 200ms before N1 onset)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agent logit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-0.18</td>
<td>-</td>
</tr>
<tr>
<td>Prime Type</td>
<td>0.11</td>
<td>-</td>
</tr>
<tr>
<td>Target Type</td>
<td>3.82</td>
<td>***</td>
</tr>
<tr>
<td>Time Window</td>
<td>-0.06</td>
<td>.08</td>
</tr>
<tr>
<td>Prime Type * Target Type</td>
<td>0.38</td>
<td>*</td>
</tr>
<tr>
<td>Target Type * Time Window</td>
<td>-0.19</td>
<td>**</td>
</tr>
<tr>
<td>Patient logit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-0.34</td>
<td>*</td>
</tr>
<tr>
<td>Prime Type</td>
<td>-0.32</td>
<td>***</td>
</tr>
<tr>
<td>Target Type</td>
<td>-3.70</td>
<td>***</td>
</tr>
<tr>
<td>Time Window</td>
<td>0.05</td>
<td>-</td>
</tr>
<tr>
<td>Target Type * Time Window</td>
<td>0.22</td>
<td>***</td>
</tr>
<tr>
<td>Preceding the N2 (-600ms - 200ms before N2 onset)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agent logit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-0.31</td>
<td>**</td>
</tr>
<tr>
<td>Target Type</td>
<td>-3.50</td>
<td>***</td>
</tr>
<tr>
<td>Patient logit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-0.17</td>
<td>-</td>
</tr>
<tr>
<td>Target Type</td>
<td>3.95</td>
<td>***</td>
</tr>
<tr>
<td>Time Window</td>
<td>-0.04</td>
<td>-</td>
</tr>
<tr>
<td>Target Type * Time Window</td>
<td>-0.19</td>
<td>**</td>
</tr>
</tbody>
</table>

*** p < .001; ** p < .01; * p < .05; - p > .1

A substantial main effect of TARGET TYPE was found for both agent and patient looks before speech onset. Before producing an agent name (i.e. an active target), participants looked more at the agent, but substantially less at the patient.
TARGET TYPE also interacted with TIME WINDOW; looks to the next-to-be mentioned agent decreased over the following 400 ms, while looks to the patient increased.

Interestingly, PRIME TYPE, i.e. the sentence structure heard during comprehension, influenced eye movements in production as well. For agent gazes, it interacted with TARGET TYPE, with additional agent looks occurring if an active target followed an active prime sentence. For gazes to the patient, a main effect of PRIME TYPE indicated that fewer such gazes occurred following an active prime. However, this probably just reflects the fact that passive targets – presumably generating sentence-initial patient gazes – practically only occurred following passive primes. Similarly, the significant negative Intercept for patient looks is probably due to the smaller number of passive target sentences overall.

In the interval preceding the N2, the overall predominance of active sentences is reflected by the negative Intercept for agent looks, since the N2 in an active sentence is always the patient. More importantly, the main effect of TARGET TYPE for both agent and patient looks persisted; participants again looked predominantly at the entity they were about to mention. The interaction between TARGET TYPE and TIME WINDOW, signalling a reduction of these looks to the to-be-mentioned entity at the end of the sentence, was significant for patient looks only.

To conclude: eye movements during sentence production were very strongly affected by the target sentence that was being prepared and uttered, a pattern of results consistent with established findings (e.g., Meyer et al., 1998). Participants routinely looked at referents before they mentioned them, and these looks usually decreased in the 600 ms before the onset of the noun.\textsuperscript{15} Taken on its own, this finding would suggest that eye movements in production are almost entirely driven by the preparation of the referent nouns. As described in more detail in Chapter 1, such fixation behaviour has been linked to the formulation phase in language production (Griffin & Bock, 2000; Bock et al., 2003): the eyes move consistently from one referent to the next in the order of their mention in the sentence, reaching each one a little less than a second before the onset of its name.

\textsuperscript{15} Calculation of the eye-voice span for each referring expression corroborated this finding: Participants generally looked at the entities a little less than 1 s before they mentioned them.
In this experiment, however, participants’ eye movements in production were not uniquely determined by the formulation of their own utterance. Crucially, they were also affected by having previously heard a particular prime structure. Interpretations of this novel finding will be put forward in the concluding section, following a comparison of these results with those obtained in the other two studies.

2.3 Summary of findings

The main purpose of this section is to compare the results of Experiment 1, summarised in the following, with those of Experiments A and B (Section 2.3.2). This should help to clarify which findings are due to characteristics of the individual experiments, and allow me to draw some more general conclusions with regard to eye movements in comprehension and production.

2.3.1 Summary of Experiment 1

Experiment 1 replicated previous studies using a dialogue setup (e.g., Branigan et al., 2000) by showing an effect of syntactic priming from comprehension to production. Participants’ gaze patterns during the two tasks also echoed established findings. Fixations during comprehension were influenced by a variety of factors, with the structure of the sentence increasing in importance as it unfolded (Eberhard et al., 1995; Yee & Sedivy, 2006; Huettig & McQueen, 2007). In contrast, fixations in production were tightly linked to upcoming sentence content, with the target structure as the main source of influence throughout (Meyer et al., 1998; Griffin & Bock, 2000; Gleitman et al., 2007).

In addition, Experiment 1 extended previous research by examining participants’ eye movements in a syntactic priming situation. The prime sentence structure was found to affect eye movements during target production, over and above its influence on the choice of target sentence structure: formulation-related eye movements to the upcoming entity were enhanced if prime and target used the same (active) voice.
2.3.2 Summary of related Experiments A and B

As mentioned in section 2.1.2, two additional experiments similar to Experiment 1 were conducted at roughly the same time and in comparable conditions. While Experiment 1 used transitive verbs to look at priming of active and passive voice, Experiments A and B investigated syntactic priming of double object (DO, see example (22)) and prepositional object datives (PO, (23)).

(22) The nun is giving the cowboy an apple.
(23) The nun is giving an apple to the cowboy.

In Experiment A, verbs were repeated between prime and target (as in Experiment 1), but Experiment B used different verbs for prime and target trials, so after hearing a prime like (8), the participant might have to use the verb throw for a picture of a sailor throwing a ball to a monk. In both experiments, the orientation of the target images was counterbalanced, with the agent appearing on the left in exactly half the target trials. In all other respects, Experiments A and B were comparable to Experiment 1, though the three arguments required by ditransitive verbs altered some procedural aspects of the analysis: with regard to the first noun phrase, gazes to the agent were analysed in the time window of 200-600 ms following N1 onset in comprehension, and preceding N1 onset in production (note that the agent was always the first-mentioned subject, regardless of PO or DO structure). For the corresponding time periods relating to N2 and N3, the dependent variables were the empirical logits of gazes to the recipient and the theme.\textsuperscript{16}

The important results of these two experiments are summarised in detail in the following. Unless otherwise stated, significance values refer to the relevant coefficients in both experiments (detailed analyses will be found in Arai et al., 2009).

2.3.2.1 Syntactic priming in Experiments A and B

A substantial priming effect was found in both experiments, with participants less likely to produce a PO target after a DO than after a PO prime, and

\textsuperscript{16} Themes consistently received fewer fixations than recipients (significant Intercepts in all analyses). This could be due to many factors, including the small size of their depiction relative to the agent and the recipient, their position in the centre of the screen, or their inanimacy.
vice versa. In Experiment A, the priming effect also received a lexical boost, with 54% more PO target descriptions following PO primes than following DO primes (relative to all target descriptions), compared to 11% in Experiment B.

2.3.2.2 **Eye movements in comprehension: Gaze analysis**

In both experiments, the odds of fixating the agent at the start of a comprehension trial were affected only by its position on the screen, with agents on the right receiving fewer fixations than those on the left ($p < .01$). Interestingly, the Intercept was not significant ($p > .1$), implying that the agent was fixated no more than the other two ROIs together.

Following the onset of the sentence, gazes to the recipient and theme in comprehension were affected primarily by the prime sentence structure. In the 400 ms following the onset of the N2, a consistent interaction of PRIME TYPE and TIME WINDOW reflected a gradual reactive reduction of gazes to the recipient, after hearing it mentioned in a DO description ($p < .01$). Correspondingly, gazes to the next-to-be-mentioned theme increased ($p < .001$).

Following the N3, an overall effect of PRIME TYPE reflected a reduction in gazes to the recipient while hearing the N3 theme of a DO sentence, while the theme itself was fixated more (all $p < .001$). This finding is comparable to the predictive effect of PRIME TYPE following N2 in Experiment 1, and will be discussed in more detail later. PRIME TYPE again interacted with TIME WINDOW, though this time gazes to the recipient increased, while those to the theme decreased (both $p < .05$).

In short, eye movements in comprehension were largely based on the unfolding sentence structure, with evidence for both prediction (e.g., the increase of gazes to theme following N2 recipient mention) and reaction (e.g., the overall preference for looking at the mentioned entity during N3).

2.3.2.3 **Eye movements in production: Gaze analysis**

A similarly strong link with the current sentence was found in production, as the following paragraphs will illustrate. In the time window before speech onset, for instance, there was a general preference for fixating the agent, irrespective of its position ($p < .001$). At the same time, IMAGE ORIENTATION affected pre-onset looks to the agent in production as well ($p < .05$), with fewer looks to agents on the right
side of the screen. There was also an effect of TARGET TYPE, with a reduction in looks to the agent if participants were preparing a DO target sentence ($p < .01$). This could be due to the fact that looks to the next-to-be-mentioned recipient were already increasing in this condition, though this hypothesis was not tested.

Before producing the N2, i.e. the recipient in a DO sentence, gazes in production were primarily affected by the structure of the target sentence: the recipient was fixated considerably more than any other ROI, the theme much less (both $p < .001$). In addition, TARGET TYPE interacted significantly with TIME WINDOW, such that these looks to the recipient decreased over time, while looks to the theme increased (both $p < .001$). PRIME TYPE affected gaze patterns in this time window as well, over and above its influence on the choice of target structure; participants who had heard a DO prime were more likely than others to be looking at the theme ($p < .05$). However, it seems likely that this finding was caused by the overall low proportion of DO targets, particularly after PO primes and in Experiment A. In addition, participants in Experiment A were less likely to look at the theme if the recipient was on the left ($p < .001$).

In contrast, gaze patterns preceding N3 mention were affected only by TARGET TYPE, i.e. by the identity of the upcoming referent. Participants preparing a theme name in a DO sentence were much less likely to be looking at the recipient and more likely to fixate the theme (both $p < .001$). These theme looks continued to increase over the time period, while those to the recipient decreased further (both $p < .01$).

To summarise, eye movements during sentence production were largely determined by the sentence structure: fixations were consistently directed to the next to-be-mentioned entity and usually decreased just before the onset of its name. However, other factors did also play a role; particularly at the beginning of the sentence, gazes were also affected by the orientation of the target image, as well as by which prime structure had been heard.
2.3.3 Comparison of results between experiments

2.3.3.1 Syntactic priming effect

All three experiments consistently found an effect of syntactic priming on the structure of the target sentence. With repeated verbs, the size of the effect was roughly equivalent across prime structures (active vs. passive voice in Experiment 1, and DO vs. PO datives in Experiment A: 49 % and 54 % priming, respectively). In both cases, the dispreferred structure was hardly ever produced as a target if it hadn’t been preceded by a prime with this structure. This was different in Experiment B, where verbs were never repeated between prime and target. Here, although the prime structure did affect target production, both possible target structures occurred after both kinds of prime. Overall, the reliable priming found in all three experiments makes it possible to compare gaze patterns during comprehension and production of similar sentences, as well as between interlocutors with aligned syntactic representations.

2.3.3.2 Eye movements in comprehension

Eye movements in comprehension were affected by a number of different factors in all three experiments. At the very start of the trial, visual characteristics of the scene layout played the most important role, with participants initially preferring to fixate the left-most entity on the screen. This visual influence continued to influence gaze patterns throughout the entire trial in Experiment 1, and towards the end of the sentence in Experiments A and B.

Nonetheless, in all three experiments the unfolding sentence also gained in importance in the course of the trial. For instance, following the N1 in Experiment 1, the initial interaction of PRIME TYPE and TIME WINDOW only enhanced the general tendency to fixate whichever entity had just been mentioned. By contrast, a similar interaction following the mention of the N2 in Experiments A and B signalled a reaction to the sentence content, such that participants predicted the upcoming N3 and moved their gaze to look at it. Finally, PRIME TYPE exerted an independent effect on fixations following the final noun in all three experiments, showing that participants had correctly predicted its identity and shifted their gaze toward it.
The main difference between experiments in comprehension concerns the very start of the trial. In Experiment 1 (transitive verbs), participants were considerably less likely to fixate the patient than the agent, irrespective of its position on the screen. No such effect was found in Experiments A and B (ditransitive verbs), where the agent was initially fixated to roughly the same extent as the other ROIs (the individual odds of fixating the other entities were not calculated, but in view of the results in the later time windows it seems likely that the theme was fixated less than the recipient).

2.3.3.3 Eye movements in production

During production trials, eye movements were much more tightly locked to the sentence that was being prepared. A strong tendency for fixations to whichever entity was going to be mentioned next was found in all three experiments and in every time window. There was also a pervasive interaction of TARGET TYPE and TIME WINDOW, with the sign of the coefficient always the opposite of the coefficient for the simple TARGET TYPE effect. This means that fixations to the next-to-be-mentioned entity consistently decreased just before the onset of its name, while fixations to the following entity increased.

Nonetheless, gaze patterns in production were not exclusively determined by the target sentence. Experiments A and B, which used displays with three entities, found that the factor IMAGE ORIENTATION also contributed to fixation patterns, particularly in the early part of a trial. Related to this finding, early visual attention to a particular entity on the screen has been reported to influence sentence structure (Myachykov, Tomlin, & Posner, 2005; Myachykov & Garrod, 2006; Gleitman et al., 2007). Experiment 1 would have been well-suited to address this question, since the first-mentioned entity automatically determined whether an active or passive sentence would be produced. However, to maximise the number of passives, all target images in Experiment 1 depicted the agent on the right, so this hypothesis could not be tested.

Finally, perhaps the most interesting finding of Experiment 1 with regard to production was the influence of comprehended PRIME TYPE on gaze patterns before the N1 onset. In particular, the interaction of PRIME TYPE and TARGET TYPE for the logit of agent gazes reflected a clear increase in fixations to the agent whenever
Chapter 2: Eyepriming - Experiment 1

prime structure was repeated in the target. This finding confirms the hypothesis that syntactic coordination can lead to coordination of visual attention as well. This result was not unambiguously replicated by Experiments A and B; although PRIME TYPE exerted a main effect on gaze patterns here too, it did not interact significantly with TARGET TYPE. Thus, due to the confound that DO targets were mainly produced following DO primes, whereas PO targets occurred after both kinds of prime, it may be that the effect of the prime structure on eye movements can be explained entirely through its effect on the choice of target structure (and consequently through the formulation-related eye movements while producing it). In contrast, the interaction found in Experiment 1 occurs for identical target sentences following different primes, so this effect must arise in addition to the influence of PRIME TYPE on target structure choice. Having presented results from three experiments, the following section attempts to draw some general conclusions, and to relate them to the theoretical positions outlined in the introductory sections.

2.4 Conclusions

This section begins with a discussion of the results regarding eye movements during comprehension (Section 2.4.1) and during production (2.4.2). The latter section will also return to the issue of priming of eye movements. Next, I address the question of whether these are coordinated or dissociated between interlocutors in dialogue (2.4.3). The final section (2.4.4) considers the limitations of this series of experiments, as well as suggesting possible follow-up studies.

2.4.1 Eye movements during comprehension

In comprehension, eye movements tend to reflect listeners’ interpretations of the linguistic and contextual information available up to that point (Tanenhaus et al., 1995; Altmann & Kamide, 1999). Thus, in a typical visual-world paradigm study, the timing of fixations to a particular object during sentence comprehension is assumed to depend on the processing of this information.
It has been shown that the interpretation of a sentence is influenced by various types of non-linguistic information, including the shape and form of the objects (Huettig & Altmann, 2007), object affordances (Chambers & San Juan, 2008), the availability of a contrast set of potential referents (Allopenna et al., 1998; Barr, 2008b), stereotypical expectations (Kamide et al., 2003a; Knoeferle & Crocker, 2006), the mental representation of the scene (Altmann & Kamide, 2007), and pragmatic assumptions about the speaker (Hanna & Tanenhaus, 2004). Experiment 1 reports another visual aspect which can influence gaze patterns in comprehension: the relative locations of the mentioned entities in the display, with a bias for the left-most entity. In addition, thematic roles were found to play a role (patients were fixated less than agents before speech onset), though it cannot be established here whether this was a visual effect of the scene layout (Glanemann et al., 2009), a conceptual bias towards agents (Bock, 1986b; Fisher, Hall, Rakowitz, & Gleitman, 1994), or the result of linguistic knowledge of the relative frequencies of actives and passives (Weiner & Labov, 1983).

Many of the studies cited above also showed that people can use current input to predict upcoming sentence content (cf. Altmann & Kamide, 1999). Predictive processes generate saccadic eye movements toward the entity that is likely to be mentioned next, with fixations to it beginning before the actual onset of its name. In Experiment 1, such predictions were not expected to influence fixations while participants heard the N1, since little information about the upcoming sentence was available at this point (remember that in half the trials the image seen by participants did not even match the confederate’s sentence). Indeed, fixations following the N1 were responsive: patient fixations decreased on hearing the mention of an agent. Hearing the N1 made it possible for participants to predict the N2, since only two potential referents were available on-screen. The continuous fixation of the N2 referent throughout its mention suggests that it was indeed predicted correctly, and this interpretation was confirmed by a similar finding regarding the N3 in Experiments A and B. In fact, these studies also provided more direct evidence for predictive eye movements: fixations to the recipient decreased following its mention as N2 in a DO sentence, and fixations to the next-to-be-mentioned theme increased. Clearly then, eye movements in comprehension reflect
listeners’ moment-by-moment interpretations of the input, integrating linguistic, visual, and any other forms of available information.

2.4.2 **Eye movements during production**

By contrast, eye movements during sentence production are linked to the sentence much more closely and seem almost exclusively to reflect lexical encoding. As described in more detail in Chapter 1, speakers routinely look at depicted objects before referring to them. These looks, and even their precise timing, have been linked to lexical selection and retrieval of phonological information (Meyer et al., 1998; Griffin, 2001).

The results of the experiments described here tie in well with established findings: gaze patterns during production were primarily determined by the planning of the upcoming sentence, irrespective of which noun phrase was being prepared. In addition, gazes were closely time-locked to the speech stream, decreasing just before the respective noun onset.

But despite its importance, target structure was not the only determinant of gazes in production. In Experiments A and B, the layout of the display also influenced pre-speech looks to the agent, such that agents on the right of the screen received fewer fixations. This effect was not found in Experiment 1, possibly as a consequence of incremental lexical encoding (Griffin, 2001): by necessity, the syntactic choice between actives and passives has to be made before the onset of the N1, since encoding the agent as the first-to-be-mentioned entity implies that the active structure is selected. In contrast, the choice of producing a DO or a PO dative can also occur after the agent has been encoded (see also V. S. Ferreira, 1996; Schriefers et al., 1998).

Intriguingly, the prime sentence also turned out to affect eye movements while producing a target sentence. This had of course been expected to the extent that the prime structure influenced the target choice, and that producing the chosen target sentence would then lead to formulation-related gazes in the order of mention. However, at least in Experiment 1, the effect on eye movements took the form of an interaction between prime and target type, implying that active target
sentences were differentially affected by the structure of the prime that had preceded them: coordinated active-active prime-target pairs showed more fixations to the agent than passive-active trials.

This finding seems compatible with descriptions of a facilitatory effect of syntactic priming. For example, Smith and Wheeldon (2001) found that target sentences were initiated earlier if a sentence-initial constituent had the same internal structure as the prime than if it did not. Similarly, Corley and Scheepers (2002) investigated priming of ditransitive verbs in a web-based experiment and observed shorter latencies when participants repeated the prime structure than when they produced a different structure. Although Experiment 1 found no prime effect on individual word onsets, the finding of increased fixations to the critical entity for initiating the repeated structure suggests that syntactic priming may facilitate thematic role assignment. This could in turn contribute to the preference for repeating the prime structure. In addition, the effect of priming on eye movements supports a link between linguistic and visual coordination in a dialogue setting. During trials on which participants were linguistically aligned with their interlocutor (i.e. trials on which they repeated the prime structure) their eye movements were more likely to correlate with the (hypothesised) formulation-related gazes of the ‘primer’ than when they did not prime.

On the other hand, the only previous evidence for an effect of priming on eye movements that I am aware of comes from Kuchinsky, Bock, and Irwin (2006; 2009). In a rather different experimental setup from the one used here, they investigated English and Dutch speakers in a time-telling task. They reported a partial decoupling of eye movements and speech, such that eye movements were more likely to be affected by priming over time than the speech utterance was. On the basis of their findings, Kuchinsky and colleagues claim that the link between speech and eye movements in production is relatively loose: since changes in visual attention to particular elements of the display can occur in the absence of overt changes in the way they are described, neither the sequence of fixations nor the sentence structure is predictable on the basis of the other. This assertion takes us back to the question of coordination or dissociation between eye movements and speech in dialogue. In the following section, I review how the results reported here relate to these accounts.
2.4.3 Coordination or dissociation of eye movements in dialogue

As discussed in Section 2.1.2, the three experiments discussed in this chapter were designed to test two hypotheses regarding the link between interlocutors’ eye movements: the coordination and the dissociation account.

The coordination account states that people’s eye movements in comprehension and production are highly similar and closely yoked, due to their status as moment-by-moment indicators of attention and cognitive processing. Since sentence components are generally processed in the order of their mention, irrespective of whether the sentence is being heard or spoken, eye movements should follow this order closely. In fact, the studies by Richardson and colleagues (2005; 2007) suggest that in successful communication listeners’ eye movements will echo speakers’ with only a short delay. An explanation for such a finding related to the Interactive Alignment model (Pickering & Garrod, 2004) assumes that visual attention is one of many levels of representation involved in language use. If this is the case, visual coordination could influence alignment at other levels of representation, such as syntax. Conversely, syntactic priming might also lead to greater visual coordination.

The support for the coordination account provided by Experiment 1 is mixed, at best. Certainly, participants did coordinate the syntactic structure of the target with the prime, suggesting that similar linguistic representations were accessed in comprehension and production. Importantly, linguistically coordinated trials showed an effect of visual coordination as well: producing an active target following an active prime led to a sentence-initial increase in fixations of the agent. There was also a general overall similarity between fixation orders in comprehension and in production. In both cases, participants tended to look at the two entities in the order of their mention, i.e. first at the agent and then at the patient for active sentences, and vice versa for passive sentences.

However, this basic resemblance was the only clear correspondence between comprehension and production gazes: while sentence structure affected fixations in both comprehension and production, gaze patterns in the former were also influenced by the screen positions of the agent and patient, while an effect of the
prime sentence was found on the latter. The fact that different additional factors exert an influence in the two tasks means that coordination of gazes between them can never be perfect.

Even more importantly, the timing of fixations in comprehension bears little resemblance to the timing in production. During formulation of a target sentence, the three experiments described here all found a strong preponderance of fixations to every next-to-be-mentioned entity in the 600-200 ms before its mention, regardless of the sentence position or structure. In addition, these fixations always decreased over the period of analysis. By contrast, no clear pattern was found for when fixations to the relevant entity occurred during comprehension: in Experiment 1, for instance, a large number of gazes to the N1 entity occurred only after its name onset, while fixations to the N2 referent generally reached it before it occurred in the sentence.

In short, whereas eye movements in production were closely time-locked to the onsets of the sentence components, those in comprehension did not always follow referring expressions by even a roughly similar lag. Instead, when it was possible to predict what would be mentioned next, participants looked at the predicted entity ahead of its mention. When this was not possible, or when the prediction turned out to be wrong, the relevant entities were fixated only after the referring expression had been mentioned.

On the other hand, the discrepancy between the findings described here and those of Richardson and colleagues may be due primarily to differences in design and analysis. For instance, Richardson and Dale (2005) calculated the distribution of fixation recurrence for each speaker and a listener, i.e. the amount of time that both were looking at the same object, for different time lags. They compared the amount of recurrence in these distributions with a baseline consisting of the overlap between the speaker’s fixations and those of an unsystematic listener, created by randomising the temporal information of the listener’s eye movements. However, this cross-recurrence analysis does not take into account the individual onsets of lexical expressions; in a sense, it compares fixations for two types of listener, rather than for two types of linguistic information. Thus it seems likely that Richardson and Dale’s global measure of maximum overlap in attention between speakers and
listeners may obscure the more fine-grained temporal dissociation between their respective eye movements.

Nonetheless, the clear discrepancies between determinants of and motivations for gazes in comprehension and production make it unlikely that precise coordination of visual attention between speakers and listeners plays a critical role in communicative interactions, as has been suggested (e.g., Clark, 1996; Richardson & Dale, 2005; Richardson et al., 2007). The results of Experiments 1, A, and B seem more supportive of a dissociation account, stating that eye movement patterns can differ between comprehension and production, and that a lack of visual alignment does not necessarily endanger the success of the conversation.

Interestingly, despite the fact that anticipatory gazes seem to be the main cause of the dissociation between gaze patterns in comprehension and production, it can also be argued that being able to predict what an interlocutor is going to say next actually increases helpful and felicitous contributions to the conversation. For instance, it may make it easier to restrict the set of potential referents (Altmann & Kamide, 1999) and respond more quickly (e.g., when performing a joint task such as preparing a cake; Hanna & Tanenhaus, 2004), to understand the speaker despite noisy input (Pickering & Garrod, 2007), and to provide quick and accurate feedback if something is going wrong.

In conclusion, Kuchinsky, Bock and Irwin’s suggestion that the link between eye movements and speech production is loose (Kuchinsky et al., 2006; see Section 2.4.2 above) would seem to hold for the relationship between eye movements during production and those during comprehension as well: while there are clearly some instances of visual coordination between the two, equally there are obvious ways in which they are not – and do not need to be – identical.

2.4.4 Limitations and Outlook

To recapitulate, Experiment 1 used a syntactic priming paradigm to investigate the factors determining gaze patterns in comprehension and production of similar sentences. Consistent with previous research, the primary influence in both tasks was the unfolding sentence structure. However, while the link between
gaze and speech was quite tight in production, eye movements in comprehension reflected the ongoing integration of listeners’ interpretations of the sentential content with the visual context and other conceptual factors. At the same time, even in production the pervasive influence of the target sentence structure on gazes was supplemented by an additional effect of the prime sentence structure.

This effect merits further exploration, as it is informative with regard to two critically debated issues: on the one hand, while the close connection of gazes in production with speech planning is well-documented, the function of these gazes is still unclear (see Griffin, 2004, for a discussion of the alternatives). The finding that these gazes can be affected by additional factors connected with linguistic processing but not linked immediately to lexical encoding, may provide further insight into what exactly they reflect and thus ultimately into their purpose. On the other hand, the fact that the additional influence stems from the prime sentence structure also relates to the debate concerning the mechanisms and functions of syntactic priming itself, as it could help to clarify the levels of processing affected by the priming manipulation.

However, the results of Experiment 1 contain a significant shortcoming with regard to addressing both these research questions. Due to an overall bias for using the active voice, passive targets were virtually never produced following an active prime. This meant that one cell in the four-way comparison remained empty. Since less frequent and/or marked structures may be more susceptible to syntactic priming than default structures (Pickering & Ferreira, 2008), a comparison of the gazes found during passive target production depending on the prime structure would have been particularly interesting.

Unfortunately, it is not obvious how to encourage participants to produce more passives overall. Target displays in Experiment 1 already consistently depicted the patients on the left side of the screen, although this meant that effects of target image orientation could not be analysed for production. One option would be to include target sentences using inanimate agents, since passive versions of sentences such as (24) tend to be preferred over active descriptions (Bock, 1986b).

(24) The church is being hit by lightning/ Lightning is hitting the church.
In addition, the results of Experiment B suggest that using non-repeated verbs might also increase the relative number of passive targets after active primes: although the priming effect was much reduced in Experiment B compared to Experiment A, spontaneous usage of the dispreferred structure (in this case DO datives) was relatively more frequent. Thus, it would probably be worth replicating Experiment 1 with non-repeated verbs and at least some inanimate agents, as well as possibly with counter-balanced target image direction.

Another potential limitation of Experiment 1 is the fact that the priming effect on eye movements was investigated by comparing the same person’s eye movements while comprehending the prime and producing the target. While this comparison is interesting with regard to effects of comprehension on production, a comprehensive description of the relationship between eye movements while speaking and listening would also benefit from two further comparisons: contrasting a real speaker’s eye movements during production with a simultaneous listener’s gaze in comprehension; as well as comparing the speaker’s eye movements with the previous listener’s gaze during their own subsequent production. Again however, it is hard to see how such a study could be set up, as it would probably mean abandoning the confederate-scripting paradigm and the control over the distribution of prime structures that this allows.

Finally, analyses of the scan paths, i.e. the true order in which the entities on the screen were fixated, might provide further information regarding the factors influencing gazes in comprehension and production, particularly before speech onset. Different conclusions would for instance result from the (hypothetical) finding that all participants initially fixate the left-most entity once before concentrating on the agent, compared to the finding that the agent and the left-most entity are differentially preferred on some pictures or by some participants. In a similar vein, depicting additional irrelevant characters on the scene would make it easier to separate effects stemming from the actual roles of the agent and patient from those due to the fact that there was nothing else on the screen to look at.

Using a very different paradigm, the subsequent experiments return to the issue of visual coordination between speakers and listeners. This time however, the direction was reversed: the following chapters address the question of whether comprehension is improved by following the speaker’s gaze, rather than how
people’s eye movements relate to whether they reproduce a previously comprehended structure. In addition, visual coordination in these studies was deliberately manipulated, rather than occurring relatively naturally as a result of priming.
Chapter 3: The use of an interlocutor's gaze in language processing

The critical element of Experiment 1 for the purpose of this thesis is the comparison of eye movement patterns in comprehension and production. Comprehension gazes reflected the continuing integration of all available information, while production gazes were much more consistently (though not absolutely) determined by upcoming speech. This tight link between eyes and speech means that eye movements in production reliably contain important information about how the utterance is going to continue; because speakers look at what they are about to mention, gaze direction can act as a visual pointer to upcoming referents, and thus aid comprehension of the utterance. It is even conceivable that listeners might be able to make use of the temporal characteristics of production-related gazes, preceding mention by about 800 ms (Griffin & Bock, 2000). Such an ability would be especially useful for anticipating sentence structure.

The experiments described in the remainder of this thesis explore these issues by supplying listeners with information about the speaker’s gaze behaviour, in addition to the spoken sentence. As a specific introduction to the ‘gaze-projection’ paradigm, this chapter summarises the background literature, and explains its motivation and practical details. It begins with two theoretical parts addressing the usefulness of gaze in dialogue from different angles. First, I review existing research on the role of real gaze in face-to-face conversations (Section 3.1). Next, I concentrate specifically on the information contained in these gazes and how it can be processed, presenting some examples of studies using gaze projection (3.2). The final, more practical Section 3.3 summarises the methodology and questions underlying Experiments 2 to 7.
Chapter 3: Gaze projection

3.1 The role of gaze in face-to-face conversation

Intuitively, it is helpful to be able to see where one’s interlocutor is looking during face-to-face conversation. This is because gaze helps coordinate turn-taking, provides feedback on whether an utterance has been understood, and can disclose which aspects of the world the speaker is talking about, as well as revealing interest or boredom. Experimental evidence for the multiple functions of gaze is presented in the following sub-sections; beginning with a discussion of how accurately it can be perceived.

3.1.1 Gaze detection and gaze following

Detecting where someone else is looking is a skill that humans practice constantly, with ease, and to a high degree of accuracy. In a classic study, Gibson and Pick (1963) trained a confederate to fixate certain areas of the room. Participants’ reports of her focus of gaze showed a level of differentiation which corresponded to being able to discern the smallest characters on a standard acuity chart. They were particularly accurate at detecting when the confederate was looking directly at them.

These findings have been replicated many times and suggest the existence of a specialised neuro-cognitive mechanism to detect the direction of another person’s gaze (Baron-Cohen, 1995). This ‘Eye Direction Detector’ probably involves the superior temporal sulcus (see also Emery, 2000; Ricciardelli, Ro, & Driver, 2002b), is extremely fast, and works with high geometric accuracy. As Baron-Cohen (1995) shows, such a mechanism can be found in many species, as it has a high adaptive value: since both people and animals tend to fixate the target of their next action (Hayhoe & Ballard, 2005), becoming aware of ‘eyes directed at me’ can signal that an action likely to affect me is about to take place. Obviously, the predictive value of gaze awareness is particularly useful if the action is aggressive, but it is also important for the manifold social interactions that characterise primate societies (Baron-Cohen, 1995; Emery, 2000).

But not only does it seem likely that specific brain mechanisms have evolved for processing gaze direction; several authors also argue that the human eye with its
white sclera and dark iris makes it particularly easy to extract this information (Langton, Watt, & Bruce, 2000; Kingstone et al., 2003; Schwaninger, Lobmaier, & Fischer, 2005). In addition, cues such as body and head orientation help in the rapid determination of gaze direction (Langton et al., 2000; Pusch & Loomis, 2001; Lobmaier, Fischer, & Schwaninger, 2006). Moreover, Lobmaier and colleagues (2006) point out that the presence of plausible fixation targets can bias where someone’s gaze is perceived to be directed, since they are more likely to be looking at an object than at the empty space next to it.

The relative importance of these different factors for computing the direction of a partner’s gaze varies depending on the situation, the lighting, and the relative positions of gazer and observer. Yet however it is obtained, someone else’s gaze provides a highly salient cue to direct attention to the object of their fixation, i.e. to see what they are looking at (Driver et al., 1999; Ricciardelli, Bricolo, Aglioti, & Chelazzi, 2002a). This tendency is particularly strong if the target is in the visual periphery (Langton et al., 2000), where the observer is less likely to be aware of it already. Even in photographs, the eyes of a depicted person and the object they are looking at attract substantially more and earlier fixations than any of the many competitor regions that a real-world scene contains (Castelhano, Wieth, & Henderson, 2007). This presents a challenge to models of gaze control based primarily on visual saliency (e.g., Itti & Koch, 2001, see Section 1.1.2 above), but it also reveals just how compelling eye gaze is in attracting attention. In fact, even young infants respond to changes in an adult’s gaze and head orientation by shifting their own gaze correspondingly (e.g., Scaife & Bruner, 1975; Butterworth, 1995). I return to the implications of such early forms of joint attention in Section 3.1.4. For now, the important findings are that humans are extremely skilled at estimating the focus of another’s gaze, and that face-to-face situations contain a strong incentive to recruit this kind of information.

### 3.1.2 Functions of the addressee’s gaze in dialogue

If gaze direction is such a ubiquitous and highly salient cue in interactions, how is it used in dialogue? Certainly, being able to see where one’s interlocutor is looking is one of the core features of face-to-face conversation (Clark & Brennan,
1991; Clark, 1996); but how exactly is it helpful? In this section I discuss the kind of information the speaker extracts from the addressee’s gaze pattern, before looking in the opposite direction, at the value of the speaker’s gaze for the addressee (Section 3.1.3).

One of the first detailed analyses of gaze and mutual gaze in an unscripted conversation between two interlocutors comes from Kendon (1967). Although Kendon found that overall, speakers fixate addressees less than vice versa, they are particularly likely to look towards addressees during passages of fluent speech and at the end of phrases, as well as at the end of a turn (cf. Levine & Sutton-Smith, 1973). In all of these cases, Kendon argued that they are looking for back-channel responses and other forms of feedback, signalling that the addressee has understood what has been said so far. In addition, he proposed that speaker gazes to the addressee at the end of a turn serve to determine what the listener is going to do next, and whether he or she is likely to take the floor. Consequently, listeners who wish to take a turn at contributing to the conversation will seek to meet the speaker’s gaze whenever they feel that such a turn-transition is possible.

Experimental support for these assumptions comes from studies by Bavelas, Coates, and Johnson (2000; 2002), who examined the effects on narrative storytelling of responses from the addressee. Responses could be generic (as in nodding, saying “mhm”, or making gaze contact) or they could reflect specific reactions to emotional content, such as wincing or sighing. The quality of the speaker’s narrative decreased significantly if addressees were distracted and thus less responsive to the current narrative (Bavelas et al., 2000). In fact, Bavelas et al. (2002) showed that listeners are sensitive to when speakers are looking at them, and specifically time their responses to occur at these times.

Thus, it seems that information about where addressees are looking, as well as whether they meet the speaker’s gaze, provides important and time-critical feedback, assisting the speaker in producing a fluent and appropriate contribution. In this sense, Bavelas et al. (2000) conclude that: “Listeners may not be equal co-narrators, but they are essential” (Bavelas et al., 2000, p. 950; cf. Clark & Krych, 2004, for a similar conclusion).
3.1.3 Functions of the speaker’s gaze in dialogue

As already mentioned, speakers are generally less inclined to look at their addressees, but they do show prolonged gazes towards them at the end of their utterances (Kendon, 1967; Levine & Sutton-Smith, 1973). At this point it is up the speaker to decide whether to meet this gaze, thereby conceding the floor. In contrast, the beginning of a turn is characterised by the new speaker looking away. Consequently, the availability of mutual gaze informs the current addressee of the speaker’s plans regarding turn-taking.

However, awareness of a speaker’s gaze is useful even in complete monologue situations. This is because when and where a speaker is looking can also signal attentiveness, competence, attraction, dominance, or emotional intensity (for a review of such social functions of gaze, see Kleinke, 1986). More directly, due to the propensity during language production of fixating objects before mentioning them, speakers’ gaze also reveals what they are talking about. For instance, Hanna and Brennan (2007) show that addressees can use speakers’ gaze direction to disambiguate referring expressions. In their experiments, pairs of naïve participants were separated by a barrier, so they could see only each other’s faces, but not the coloured objects in front of their partner. They were randomly assigned the roles of director and matcher, and the director instructed the matcher to move specific objects in their display. These were arranged so that there was either an identical competitor at another position in the display, or else that shape-colour combination was unique. If the displays mirrored each other, matchers were able to use the director’s gaze to identify temporarily ambiguous targets well before the point of linguistic disambiguation. In fact, they were quick to locate the correct target even when the order of the display was reversed, apparently re-mapping the director’s gaze. Hence, Hanna and Brennan propose that eye gaze in referential communication is initially processed automatically, but can be adapted based on additional contextual knowledge in a second processing step (see Staudte & Crocker, 2008, for evidence that people even use robot gaze for disambiguation).

A similar demonstration of the effectiveness of gaze direction as a cue in conversation was provided in a more elaborate setting by Monk and Gale (2002). Their investigation of video-mediated conversation used a novel type of display,
which made it possible to simultaneously project the object under discussion and a mirrored reflection of the speaker’s face. This allowed an accurate estimation of what they were looking at, as if they were sitting at the other side of the table, behind the object under discussion (see Gale & Monk, 2000). In pairs of participants, such a speaker described precise locations on complex pictures (e.g., an electron microscope image or a circuit diagram); the addressee had to identify the correct location. Remarkably, availability of the speaker’s gaze meant that dyads required only about half the number of turns and words to complete the task, compared to control conditions of playing back only the audio description, or where the face projection did not overlap with the object of conversation (Monk & Gale, 2002).

Presumably, being able to see the speaker’s gaze is beneficial precisely because speakers tend to look at what they are going to speak about. Thus, while gaze can be used deliberately in deictic communication (as when the parents of a fractious child alert each other to an approaching ice-cream wagon without overtly mentioning it), most of the time speakers fixate relevant objects automatically. It seems that addressees routinely follow these gazes and make use of the information they contain. The next section discusses the importance of such instances of joint attention for language acquisition and early cognitive development.

3.1.4 The role of joint attention

We have already seen that even young infants tend to look in the direction of an adult’s gaze (Scaife & Bruner, 1975; Butterworth, 1995; Brooks & Meltzoff, 2002). The referential information contained in this gaze is particularly valuable for infants acquiring language, as it facilitates the mapping of words to the objects they refer to (Baldwin, 1993; 1995; see also Yu, Ballard, & Aslin, 2005). In addition, the ability to follow an adult’s gaze and to establish a state of joint attention with them is critical for learning to understand others as distinct agents with intentions of their own (Tomasello, 1995) and for developing a theory of mind (Baron-Cohen, 1995).

Conversely, individuals who lack an awareness of joint attention or an inclination to adopt it show profoundly impaired social interactions and frequently also linguistic difficulties (Baron-Cohen, 1995; see Williams, Waiter, Perra, Perrett, & Whiten, 2005, for a discussion of the brain areas involved).
However, the term *joint attention* is not always used consistently in the literature. Thus, Butterworth (1995) defines it simply as “looking where someone else is looking” (p. 29), but according to other accounts, joint attention requires an additional understanding by one or both partners that the attentional focus is currently shared (see Eilan, 2005, for a summary of the debate from a philosophical perspective). To use an example from Baldwin (1995), it is quite possible to share the focus of attention with a horse, without assuming either that the horse is aware of this overlap, or that this will pave the way for a communicative exchange with it. But to benefit language acquisition, it seems critical that infants possess at least a basic awareness of when attentional focus is shared, and of the fact that this singles the fixated object out as a likely topic of conversation. Tomasello (1995) makes a similar point: “Joint attention is primarily a social, or social-cognitive, phenomenon: Two individuals know that they are attending to something in common” (p. 106).

A number of the papers cited in this section describe ingeniously designed studies to test infants of different ages for signs of such intersubjective awareness of attention (e.g., Baldwin, 1995). In summary, they suggest that while even infants in their first half year orient in the direction of an adult’s attention, this early gaze-following is probably based on a learnt association between someone else’s direction of regard and the likelihood of there being something interesting to see there (Butterworth, 1995). Just after their first birthday however, infants begin to perform the geometric triangulation linking the adult’s gaze with an object on this line of regard, and at 18 months, they competently use a range of contextual cues to assess the actual attentional focus of the adult. In this way, they can avoid learning incidental object-name pairings:

What is noteworthy here is that infants are not merely linking new words with the objects made salient by speakers’ actions. Instead, they seem to be making inferences concerning another’s attentional focus and referential intent based on close observation of how language is overlaid on subtle aspects of behavior. Already infants are treating language as linked with mental life. This is intersubjective understanding in clear relief, and it is being skillfully recruited in the service of language learning. (Baldwin, 1995, p. 152; see also Moore, Angelopoulos, & Bennett, 1999)

A similar line of development can be traced for learning to comprehend and produce pointing behaviour (Butterworth, 1995; see Bangerter, 2004, regarding pointing in adulthood). In fact, as a visually more pronounced cue than eye gaze,
pointing may be particularly helpful in establishing joint attention to a visible object, whether it is initiated by the adult or by the child.

To summarise, joint attention clearly plays a critical role in early cognitive, social, and linguistic development, but the facilitation of understanding that it provides continues in adult life as well: following the direction of someone’s gaze helps in identifying what they are talking about (Hanna & Brennan, 2007). It seems likely that this benefit at least partially motivates the strong predisposition to attend to objects in the line of other people’s gaze (Driver et al., 1999; Ricciardelli et al., 2002a; Castelhano et al., 2007). At the same time, it is not clear whether a mutual appreciation of the shared focus of attention is as critical for moment-by-moment comprehension as it is for acquiring new words. After all, an adult addressee generally knows the meaning of all the words they hear; being able to see what the speaker is looking at presumably just allows them to generate an accurate representation of the utterance meaning earlier, by increasing the salience of the relevant object. For the purpose of this thesis, I will therefore adopt the simplest possible definition of joint attention, as a situation where two people are simultaneously looking at the same thing (cf. Butterworth, 1995). Whether this also requires understanding that their gaze target is shared must remain open, though I return to this issue at several points in the following chapters. The following section discusses a family of paradigms for investigating the influence of joint attention on a variety of cognitive processes.

3.2 Gaze projection

Section 3.1 revealed that humans are strongly inclined to follow another person’s gaze, and highly skilled at detecting where it is directed. This section addresses a different aspect of the general question of how gaze can be used in dialogue: rather than focussing on how accurately the signals of another person’s attention are exploited (e.g., eye gaze, head movement, pointing, body orientation), it looks at how seeing what they are attending to can be integrated into the representation of a joint task, whether this be conversation, visual search, or collaborative construction. In gaze-projection paradigms, this is achieved by projecting
the focus of one or both participants’ gaze onto the workspace or into the visual world, frequently in form of a moving dot. This is likely to enhance joint attention, since one participant’s gaze will be drawn to whatever the other is fixating. To illustrate, I begin by summarising some previous studies using this kind of setup.

3.2.1 Precedents

To my knowledge, the first report of projecting interlocutors’ eye movements onto an experimental display comes from Velichkovsky (1995). To explore the influence of joint attention on a cooperative problem-solving task, he provided some participants with training at solving computerised puzzles, which were later used as stimuli. In the main experiment, these experts sat at screens linked to those of novices, and the pairs were instructed to solve the puzzles as quickly as possible. However, only the novice could use the mouse to move pieces, so the expert had to provide assistance verbally. The effects of two different forms of feedback were compared to such a voice-only condition: for voice & gaze, a red circle on the novice’s screen reflected where the expert was looking, while in the voice & mouse case the position of the expert’s computer mouse was projected to the novice’s screen. Both participants were fully aware of the setup, so the expert could use an extended fixation or jiggle the mouse to point out particular pieces. A second experiment reversed the direction of the manipulation by projecting the novice’s gaze to the expert’s screen.

In both experiments, the time to solve the puzzles was significantly reduced by the availability of the gaze cursor, though the difference between voice & gaze and voice & mouse was not significant. In fact, experts and novices both adapted readily to the novel modalities, frequently using them as pointing devices. For instance, experts’ instructions often took the form of: “Take this (marked by eye or mouse), place it there (again marked)” (Velichkovsky, 1995, p. 208). This supports the conclusion that the additional modality created a situation resembling direct physical copresence, which facilitated reference and grounding and benefitted understanding and task success (Clark & Brennan, 1991). At the same time, participants were clearly deliberately exploiting the additional channel of communication. While interesting from an applied point of view, this limits the
extent to which the usefulness of the gaze cursor can be explained by the largely automatic and unpremeditated likelihood of fixating currently important objects, as Velichkovsky claims (Velichkovsky, 1995, p. 216; for a discussion of the respective characteristics of eye vs. mouse movements, see Brennan, 2005).

In Velichkovsky’s experiments, the gaze cursor was useful in both directions: while the experts whose gaze was projected used it deliberately as a pointer, they also benefited from the feedback contained in the novice’s gaze. This contrasts with findings by Bard and colleagues (Bard et al., 2007), who used a map task to study speaker’s sensitivity to information which indicates that their addressee does not understand them. Speakers described a path between landmarks on a fictitious map, allegedly for an addressee to follow. The feedback provided was either visual (in the form of a simulated gaze cursor) or verbal (produced by the confederate addressee). Overall, speakers tended to ignore gaze cursors which suggested that the addressee was heading in the wrong direction, even when the two modalities conflicted; yet they responded to verbal utterances signalling problems by rephrasing their instructions, even when the cursor suggested that all was well. According to Bard et al., these results suggest that interlocutors share the responsibility for successful communication, due to a limited capacity on the part of the speaker to integrate information about the addressee’s current representation. This is plausible in view of the general tendency of speakers to look less at addressees than vice versa (see Sections 3.1.2 and 3.1.3, above), though a real gaze cursor (as opposed to a simulation) would provide stronger evidence.

At the same time, studies where it is unlikely that the gaze cursor is employed strategically indicate that the unintentional character of most fixations is an important factor in the effectiveness of gaze cursors. For instance, Stein and Brennan (2004) recorded the eye movements of a group of expert programmers locating bugs in software programs. Later, a second group of programmers were faster to locate the bugs themselves if they had seen one of these gaze cursors scanning the code. Similarly, visually highlighting those areas of ambiguous displays which people tend to fixate just before they reach an interpretation or solution has been shown to facilitate subsequent viewers’ understanding (Pomplun, Ritter, & Velichkovsky, 1996; Grant & Spivey, 2003), suggesting that “eye movements themselves can, in a bottom-up manner, kick-start cognitive processes
that lead to insight” (Knoblich, Öllinger, & Spivey, 2005, p. 355). If this is the case, then being attracted to relevant parts of the display by the gaze-cursor of a successful problem-solver should prove helpful in generating one’s own solution. This may in fact be one explanation for Richardson and colleagues’ finding that the amount of overlapping fixations between interlocutors in dialogue predicts the success of their conversation (Richardson & Dale, 2005; Richardson et al., 2007), despite the fact that the overlap in these studies was just estimated post-experimentally, making it hard to determine cause and effect.

In contrast, Brennan et al. report an experiment where success depended precisely on being able to avoid following the gaze cursor (Brennan, Chen, Dickinson, Neider, & Zelinsky, 2008): pairs of participants performed a collaborative visual search task, searching for a letter O in an array of Qs. Massive benefits were found when the partner’s gaze was available as an on-screen cursor; to be able to avoid that part of the display, participants must have monitored where their partner was looking, despite concentrating mostly on their own search. Strikingly, a combined voice & gaze condition actually delayed task completion compared to the cursor-only case, though this seems likely to be an effect of the specific task. One of the most interesting aspects of this study is that although participants’ eye movements must primarily have been driven by the search task rather than an intention to communicate with their partner, they clearly interpreted the moving cursor as exactly what it was: their partner’s gaze. I return to the question of what the cursor represents in Section 3.2.2, after discussing two studies that investigated the effect of a gaze cursor on language learning.

As part of a larger investigation into how language acquisition is affected by so-called ‘embodied intentions’ (i.e., body movements relating to language), Yu, Ballard, and Aslin (2005) recorded a native Mandarin Chinese speaker as he related a children’s picture-book in his own words. They played this recording back to monolingual adult Americans, who were subsequently tested on Chinese word segmentation and semantics. Those who had heard only the audio recording were no better than chance at these tasks, but the two groups who had seen a video as well had learnt a considerable amount of Chinese. This was particularly the case for participants who had seen a cursor reflecting the speaker’s eye movements, as opposed to those who just saw a video of the book. Yu and colleagues propose that
the gaze cursor is more effective than the simple video because it maintains the tight time-lock between speech and body movements, which they claim is essential for filtering meaningful word-object pairings out of a continuous speech and gaze stream.

Finally, seeing a speaker’s gaze cursor has been shown to assist in learning adjectives such as *big, thin, or tall*, the meaning of which depends on the available contrast set (Brown-Schmidt, Yu, & Tanenhaus, 2009). The rationale behind this study was that speakers producing such scalar adjectives typically look at both the referent and its contrast (e.g., a big frog and a small frog in an array of 10 other objects). Thus, the information contained in the gaze cursor should be sufficient to derive the meaning of the adjective. This hypothesis was confirmed: after hearing a Mandarin speaker describe this type of display and seeing a corresponding gaze cursor, participants were able to apply the adjective-noun pairings correctly when distinguishing between two different-sized frogs and two different-coloured balloons. There was even evidence for generalisation of the adjective meaning to untrained nouns, e.g. a small bed. However, participants were no better than chance if the training display had not contained the contrast.

### 3.2.2 Advantages and limitations of the paradigm

To summarise, a number of studies have used projected gaze, either while interlocutors were engaged in a collaborative task (e.g., Velichkovsky, 1995) or, more often, recording speakers or directors first and playing their audio and gaze footage back to listeners later. Gaze cursors based on real eye movements substantially facilitate task completion, whether this is solving a puzzle, finding a target, or learning a word. In addition, it seems clear in at least some studies (e.g., Yu et al., 2005; Brennan et al., 2008; Brown-Schmidt et al., 2009) that the cursor is used as if it were real face-to-face gaze, and not just a visually salient cue which happens to be worth following.

Nonetheless, it is important to note that I am not claiming that a gaze cursor *is* equivalent to ‘real’ gaze either in the way it is processed, or in how participants interpret it. To start with, they differ in too many obvious ways. Among other
things, perceiving where someone is looking in a face-to-face context involves computing their gaze target by triangulation, and the accuracy of the result depends on a number of factors (Gibson & Pick, 1963; Pusch & Loomis, 2001). Gaze cursor information is usually much more exact, pinpointing a very small area of the screen. In addition, it does not require looking backwards and forwards between the eyes and the fixated object, so the information is available constantly. This may or may not be an advantage, but it is certainly important to bear in mind when interpreting the time course of how the cursor is processed. Also, while speakers in face-to-face contexts frequently avoid the potential distraction of looking at their listeners, participants faced with a brightly moving cursor on an otherwise static display may find it hard to avert their gaze (though the results by Bard et al., 2007, and Brennan et al., 2008, suggest that it does not unavoidably attract fixations). Finally, gaze cursor information is inherently ambiguous. The fact that I see a cursor indicating that my partner is looking at a bottle of wine on the kitchen counter could mean many different things: that he is trying to work out what it is; that he thinks it is interesting although he is actually talking about a different object; that he is preparing to describe it to me; that he is using it as a reference point for something else he wishes to describe (e.g., the chocolate cake next to it); or indeed that his eyetracker has slipped and he is actually looking at the window above. If he knows that I am able to see where he is looking, his gaze could also serve a more direct communicative function, pointing the bottle out to me (Velichkovsky, 1995).

In fact, it is even possible that gaze cursors and real eye gaze are processed in completely different ways. On the one hand, there is evidence that gaze perception on the basis of eye direction benefits from specialised processing mechanisms which are available early both in phylo- and ontogenesis (Emery, 2000). This is argued to make eye gaze direction an even more salient cue than socially important symbols such as arrows (Ricciardelli et al., 2002a). On the other hand, a similar argument of specialised processing could be made for the bright and quickly-moving gaze cursor too, which according to theories of attention should capture fixations easily (Folk, Remington, & Wright, 1994; Theeuwes, Kramer, Hahn, & Irwin, 1998). Whether one or the other or maybe even both kinds of gaze stimulus benefit from such rapid bottom-up processing is beyond the limits of this thesis, but the different possibilities are important to consider when generalising between them.
Despite these caveats, gaze projection paradigms can clearly provide interesting information on how the time-course of visuo-attentional processing relates to comprehension and dialogue. In part, this is precisely because they eliminate the need to extrapolate a partner’s focus of attention from the perceived direction of their gaze. They also reduce experimental difficulties stemming from differences between speakers: in this kind of paradigm, all participants can consistently receive the same information in real-time.

In addition, gaze projection creates a situation which resembles joint attention between interlocutors: the visually salient cue signalling the object of the speaker’s attention means that the listener will tend to adopt the same perspective onto the scene – this, in turn, is likely to assist in understanding the description (Grant & Spivey, 2003; Richardson et al., 2007). One plausible explanation for this facilitation is that the jointly fixated objects become part of common ground through physical copresence (Clark & Brennan, 1991), although the two interlocutors may not even be aware of it. Among other things, this would reduce reference ambiguity, thus aiding comprehension (see Bard et al., 2007, for a more detailed discussion). In support of this suggestion, studies of video-mediated communication generally find that it is more useful to see a shared workspace than to be able to see the partner’s face (Whittaker, 2003; Clark & Krych, 2004).

In short, gaze-projection seems well-suited to exploring the interplay of attention between partners in a collaborative task, in particular if this task is language-based. The final section of this chapter outlines a series of studies conducted using this paradigm, which form the topic for the rest of the thesis.

### 3.3 Overview of the gaze-projection experiments

The experiments reported in the following chapters used gaze projection to examine how the complementary information contained in visual and speech cues is integrated during the comprehension of unscripted object descriptions. They also investigated the situations in which one modality can replace the other, focussing in particular on the importance of temporal and semantic consistency between the two forms of information.
Because the intention was to investigate the usefulness of gaze cursors based on speakers’ natural production-related eye movements, it was important to record them during unscripted speech. This ensured that the eye movements were determined only by the characteristics of the display and by the speech task itself. The latter involved describing a series of objects in such a way that subsequent listeners would be able to identify them. However, sparse displays containing only a few potential objects seemed likely to overestimate the value of the gaze cursor in such a task, as there would not be many potential targets. For this reason, it was decided to use photographic depictions of real-world scenes, which contained many objects irrelevant to the task at hand.

Speakers and listeners were recorded separately, which meant that each speaker’s speech and gaze cursor could be played back to several listeners. This made it possible to compare listeners’ performance in different speech/cursor conditions based on the same speaker, thus allowing a generalisation beyond individual descriptions and gaze traces. Speakers knew that subsequent participants would have the task of finding the objects based on their descriptions, but they were unaware that their gaze would be played back, ensuring that they did not produce deliberate communicative eye movements. However, this setup also meant that while they produced their utterances for the benefit of subsequent experimental participants, they received no feedback as to how their descriptions were understood.

Experiment 2 (Chapter 4) served to establish the general paradigm and to explore how listeners profit from the projection of a speaker’s gaze in such a task. Listeners heard descriptions by six different speakers, either with or without the additional gaze cursor. As in subsequent studies, the effect of the cursor was investigated primarily with regard to two dependent variables: the onset time of the first fixation to the target object – a measure which reflects visuo-attentional effects of the cursor – and the latency of the listener’s mouse click. The latter measure is more intentional and was assumed to require at least a basic understanding of the utterance, making it possible to see how the two forms of information are integrated. Chapter 4 also contains detailed descriptions of the stimuli, experimental method, and analysis, since these were re-used in the subsequent experiments.
Experiment 3 (Chapter 5) was designed to assess the extent to which listeners actually interpret the gaze cursor as the speaker’s gaze, and whether it is important that they understand the connection between the two. To this end, listeners were provided with differing instructions about what the cursor represented, under the assumption that they should rely on the gaze information more if they believed it to be accurate than when it was questionable.

The idea behind Experiment 4 was to offset the timing of gaze cursor and speech utterance against each other, thereby manipulating the strength of the association between them. A cursor which slightly precedes the speaker’s actual gaze might prove additionally helpful in predicting upcoming speech content, whereas one which is largely inconsistent with speech should impede comprehension. Experiment 4 tested the effect of fast-forwarding or delaying the gaze cursor against speech by one second in each direction. However, the within-listeners design revealed a number of methodological difficulties, which are discussed in Chapter 6. These led to several modifications of the paradigm, including using only one speaker and a different eyetracker (see Experiment 5, Chapter 7). These innovations allowed a renewed examination of the timing manipulation: Experiment 6 looked at highly inconsistent speech and cursor combinations, while Experiment 7 presents a more fine-grained time-course of the effects of offsetting visual and linguistic information. These two experiments are reported together in Chapter 8. On the one hand, they tested listeners’ sensitivity to the temporal discrepancy in language production between time to fixate and time to mention; on the other hand, they revealed the points in time at which supplementary information about message content is most helpful for comprehension. Finally, the results of the gaze-projection experiments are summarised and discussed in the General Discussion (Chapter 9).
Chapter 4: Testing the Paradigm – Experiment 2

This chapter describes the first experiment based on the gaze projection paradigm, including a detailed overview of the methods and stimuli used for subsequent experiments.

4.1 Research question

The purpose of Experiment 2, described in this chapter, was to ascertain the extent to which listeners profit from the projection of a speaker’s gaze onto the scene he is describing\textsuperscript{17}, and to establish which aspects of scene processing and language comprehension this affects. In this sense, it formed a manipulation check for the entire paradigm: it seemed clear that listeners would look at the gaze cursor, given both the visual salience of a moving stimulus on a static image and the fact that the cursor reflects the speaker’s production-related eye movements. This makes it inherently informative about the upcoming speech content (e.g., Bock et al., 2003), and therefore relevant to the listener’s task of clicking on the objects the speaker is describing.

However, it was less clear whether listeners would also make use of information from the gaze cursor for their comprehension task, in addition to just looking at it. Identifying and clicking on the object that the speaker is describing clearly entails listening to what he is saying. In this experiment, listeners in all conditions heard the speech utterance played back in the same way and so received identical auditory information, irrespective of whether they saw the gaze cursor or not. Hence it was possible that there would be no difference between gaze cursor conditions in the time to click on objects. Alternatively, if the cursor is useful for

\textsuperscript{17} In all gaze-projection experiments, speakers will be referred to with the personal pronoun ‘he’; listeners with ‘she’.
understanding the utterance – similar to the way a speaker’s gaze is helpful in face-to-face interaction (e.g., Hanna & Brennan, 2007) – then listeners who were shown the gaze cursor should be faster to click on the objects than those who were not.

In addition, this experiment explored the idea that seeing the speaker’s gaze might also influence other levels of processing the speech utterance. For instance, longer fixation times on the relevant objects and/or greater certainty about which object the speaker was referring to might mean that these objects were more accurately encoded. For this reason, the main experiment was followed by two memory tasks, which are described in more detail below.

### 4.2 Design and predictions

Listeners were asked to mouse-click on an ordered series of objects displayed in a complex photographic scene, according to instructions given by a pre-recorded speaker. Each listener saw and heard three different scenes described by the same speaker. Whether listeners saw a projection of the speaker’s gaze or not was varied within speakers, but between listeners, because it was unclear whether seeing a gaze cursor in some trials would affect eye movements in others. This yoked design meant that every speaker was assigned a pair of listeners who heard his description: one of them saw the gaze cursor, the other didn’t. Listeners with a gaze cursor were told that the moving dot represented where the speaker had been looking while describing. Listeners’ eye movements and mouse clicks were recorded.

Following the presentation of all three scenes, two additional offline measures of processing and memory for the scenes were obtained: first, all participants were asked to recall and name out loud the objects they had clicked on in each scene. They were then given a paper version of a slightly modified scene and asked to spot eight differences compared to the original display; four differences related to objects that had been clicked in the original task, while the other four concerned previously unmentioned objects.

The following predictions were made concerning the effect of the gaze cursor on different dependent variables:
a) Due to the attention-grabbing nature of the bright green moving dot, listeners who were shown the gaze cursor should fixate the relevant objects earlier than those who did not see the cursor.

b) If listeners can make use of the information contained in speakers’ eye movements, the fact that it is informative about upcoming speech content should lead to faster click latencies with projected speaker gaze than without.

c) If comprehension is facilitated by projecting the speaker’s gaze, listeners who saw the cursor will have paid more attention to the relevant objects and may therefore show improved memory for them in the recall task than those who did not. In the Spot-the-Difference task, they would be expected to discover more differences concerning clicked objects than listeners who saw no gaze cursor. On the other hand, the latter would presumably have spent more time scanning the entire scene, so they might be better at noticing changes to previously unmentioned objects.

d) If visual coordination between speaker and listener leads to greater linguistic alignment as well, then listeners who saw the gaze cursor should be more likely to repeat the speaker’s referring expressions in the recall task than those who didn’t. Alternatively, if the additional visual cue reduces the importance of following the speech utterance closely, it might also be encoded less well, and hence lead them to repeat fewer expressions.

e) Finally, it seems likely that the effects of seeing the cursor would change over time, as participants learned to use the green dot which they would not initially have been accustomed to processing. This could result in a speed-up in click latencies and an improvement on the recall test over the course of the three scenes. However, first fixation onsets were expected to be less influenced by previous experience in this way than click latencies, on the assumption that first fixations to an object resulted mainly from attention being drawn to its location relatively automatically.
4.3 Methods and materials

This experiment comprised three different forms of stimuli. First, I describe the scenes that were presented on-screen to both speakers and listeners (Section 4.3.1). In addition to seeing these images, listeners heard the speaker’s utterance, and half of them saw a projection of the speaker’s gaze. I therefore continue by detailing the procedure used to record speakers’ speech and gaze, as well as the way in which this data was processed before playing it back to listeners (4.3.2). The section concludes with a description of the resulting speech and gaze stimuli (4.3.3) and a final summary of the procedure for recording the eye movements and clicks from listeners (4.3.4).

4.3.1 Stimulus and item selection

Figure 4-1 shows a grey-scale example of one of the three photographs used as stimuli in this study (the other two showed a kitchen and a bedroom, respectively, and all three are reproduced in colour in Appendix B). The images measured approximately 900 × 670 pixels and were displayed in colour on a grey background of similar luminance.

Figure 4-1: Example of experimental stimulus (presented in colour in the experiment, cf. Appendix B).
For each speaker, small, brightly-coloured numbers were superimposed on the centre or functional area of twelve objects in the photograph to indicate the order in which they were to be described; these twelve objects will be referred to as “items” in the following. The order was randomly generated and differed for each speaker. The order in which the three rooms were presented was counterbalanced across speakers.

The twelve critical items in each room (e.g. the vase of flowers in Figure 4-1) were selected on the basis of a pilot study, for which the photographs had originally been created. To ensure that the items were not too obvious, for each potential item there was another similar-sized object in view which did not need to be described. All resulting items consisted of everyday objects that were relatively easy to identify and describe, as well as visually separate from any other objects. A full list of items with further details on their selection is provided in Appendix C.

4.3.2 Creating the gaze and speech stimuli

4.3.2.1 Recording procedure

Six native British English undergraduates came to the Joint Eyetracking Lab described in Chapter 2, Section 2.2.2.4, to be recorded in the role of speaker (mean age 20.4; 2 males; 4 exclusions due to poor calibration or missed items). They were seated at arm’s length in front of the 22’ screen, and fitted with one of the EyeLink 2 trackers.

Speakers were instructed that they would be shown three photos of household rooms. Twelve objects in each room would be numbered, and their task was to describe these objects in order, so that “someone who listens to your recording in the future will be able to identify the objects and mouse-click on them. This person will be looking at the same picture, but will not see the little numbers”. They were asked to begin each instruction with the number of the object they were about to describe, e.g. “Number 1 is...”.

131
In initial recordings, the relatively active nature of the task and the long trials\(^\text{18}\) during which no drift correction could be performed (12 objects per room, roughly two minutes of description) led to frequent loss of pupil trace. This made it necessary to use corneal reflection as an additional tracking mode, which reduced the proportion of blank samples to less than 1%. However, the disadvantage of using corneal reflection on EyeLink 2 trackers is the reduced sampling rate (250 Hz instead of 500 Hz) and the fact that not all participants can be tracked in this mode, which led to the relatively high proportion of excluded speakers.

An additional measure taken to improve recording accuracy consisted of a trial-final fixation target in the centre of the screen. This appeared for seven seconds when speakers pressed a button to indicate the end of their description, making it possible to determine how much drift had occurred. Speakers were also asked to sit as still as possible and encouraged to wear an airline-style inflatable neck rest to reduce head movements. Each trial was preceded by a full 13-point calibration procedure, until average error at validation < .5° and maximum error < 1°. The eye with the smallest error was selected for monocular tracking.

Prior to the three experimental trials, speakers were given a practice trial to familiarise them with the type of scene and task. It also allowed them to ask for clarification before the actual recording or to be corrected if there was any problem (e.g., leaning forward, not mentioning object names, not describing at all). The practice trial consisted of a photograph of a classroom, with six easily identifiable objects marked with superimposed numbers.

Following recalibration, speakers then described the three experimental displays at their own speed. No speaker required the maximum display time of five minutes. After describing the objects in all three scenes, speakers piloted the surprise recall and Spot-the-Difference tasks which subsequent listeners would be asked to perform (see Section 4.3.4.2). Finally, they were asked for some basic

\[^{18}\text{Due to the unusual design of these studies, it is important to clarify terminology: trial will be used to refer to the entire presentation and description of one room (i.e. the experiment consisted of three trials). Each room contained 12 objects to be described, resulting in a total of 36 items for the entire experiment (for more details on item definition see Section 4.3.2.3).}\]
demographic information and debriefed. In total, speakers were in the lab for about 55 minutes, for which they received £5.

4.3.2.2 Creating the gaze cursor and interest areas

For each speaker, samples were extracted from the recorded eye movement data file (.edf) to be used as a basis for the gaze cursor. The resulting movement file consisted of a list of time points in milliseconds from recording onset, and x- and y-coordinates reflecting where the speaker was looking at that time. If such a list is set as a custom movement pattern, EyeLink draws a gaze cursor to that position on the screen. The timing of this presentation is accurate to the screen refresh rate (17 ms).

During blinks, i.e. when the eye was occluded, the cursor was removed from the screen. In addition, any vertical movement recorded just before and just after blinks was not shown, under the assumption that this was an artefact of the movement of the eyelid and might be confusing for listeners. In all other respects, the cursor movement files corresponded exactly to the original sample reports. The presentation software ran the same cursor files for all listeners, except that for those in the no-cursor condition the gaze cursor was not actually displayed on-screen. This ensured that the timing of presentation and recording was otherwise identical for the two groups. The gaze cursor itself was a bright green filled circle, 13 pixels in diameter. This was small enough not to cover any object it happened to be focussing on, but large enough to be easily noticeable.

For the analysis, a region of interest (ROI) was defined around each experimental item. Any fixation registered in this region within the time constraints of the analysis was treated as a fixation to that object. The ROIs were created manually, incorporating about 20 pixels around the contours of the objects. They are visible as yellow lines around the items in Appendix B.

4.3.2.3 Determining items from the speech recordings

As well as the speakers’ eye movements, their speech stream was also recorded through a clip-microphone to a .wav-file, beginning at display onset. This audio file was played back to listeners without further manipulation. However, for analysis purposes, each file was segmented into 12 items (see Footnote 18, above). This was done using Praat software (Boersma & Weenink, 2008) to mark the onset of
the word “number”, with which speakers had been asked to begin each new item description (they almost always did; otherwise the actual onset of the number name was used). In this way, any speech and gazes occurring between the words “number 1 is...” and “number 2 is...” were defined as pertaining to item 1. The onset time of each occurrence of “number” will be referred to as item onset in the following.

Because item onset could occur well before it was clear which object the speaker was referring to, the onset of the first referring noun was also marked, as a rough indicator of when the speech utterance pinpointed the target object relatively unambiguously. However, this name onset is a rather vague measure, as the referring noun sometimes immediately followed “number x is the”, whereas at other times it was preceded by a large number of modifiers (e.g., “number 3 is just below and to the right, it’s a small brown...”). In addition, quantifiers such as “a pair of” and dummy noun phrases (“the black thingy”) made the definition of the relevant referring expression somewhat subjective (though classification was kept consistent across trials and speakers with the help of lists of descriptions). Particularly in cases with several pre-nominal modifiers, it is highly likely that listeners could predict the upcoming noun well before its name onset.

4.3.3 Speaker results

Before I describe how the speaker files were presented to listeners, this section provides a very basic overview of how speakers actually did refer to the objects, and how their gaze behaviour related to their utterances. This is important with a view to listeners’ interpretations, but it is also interesting in its own right: with a few exceptions (e.g., Brown-Schmidt et al., 2005; Diderichsen, 2005; Brown-Schmidt & Tanenhaus, 2008), most studies to date on eye movements during production have employed tightly constrained sentence structures or referring expressions, depending on the aim of the research (e.g., Bock et al., 2003; Meyer et al., 2004; Gleitman et al., 2007). In addition, the visual displays used for psycholinguistic eyetracking studies usually consist of disconnected arrays of objects, or impoverished clip-art ersatz scenes (e.g., Meyer et al., 1998; Knoeferle et
al., 2005). Such stimuli can be useful in answering well-defined questions on the
time-course of language processing, but fixation behaviour when speaking about
such displays can not necessarily be generalised to more natural situations (for a
discussion of the types of stimuli used in psycholinguistics, see Henderson &
Ferreira, 2004, and Chapter 1, Section 1.1.3, above).

In contrast, the studies reported here used photographic scenes and included
many unmentioned objects in each display, with the aim of keeping speakers’ gaze
behaviour as natural as possible. In fact, speakers often included references to
neighbouring objects, e.g. “number 7 is the cafetière next to the vase, which has a
little milk jug standing just in front of it”. Consequently, while the gaze cursor
would probably focus most on the object that was being mentioned or that was
going to be described next, speakers’ gaze could also be attracted to proximal
objects, to previous or future items (all twelve superimposed numbers were visible
throughout the speaker trial), or to other interesting regions (e.g., the picture on the
wall in the bedroom).

In addition to such potential influences of the stimulus, the speakers’ task
also differed from the more conventional monologue naming of simple pictures, in
that they were asked to describe photographs for the benefit of future listeners.
Speakers’ utterances were not manipulated post-recording, even if they were
disfluent or potentially misleading. Such conditions are very frequent in every-day
speech, and it may be that seeing speakers’ gaze would help clarify the meaning for
the listeners (Hanna & Brennan, 2007).

All these factors could plausibly affect the time-course of production-related
fixations, generally assumed to land on an object about 800-1000ms before it is
named (Griffin & Bock, 2000). The methodological differences between the study
described here and previous investigations of eye movements during speech
production make it worth providing an overview of speakers’ utterances (Section
4.3.3.1) and their fixation behaviour (4.3.3.2).

4.3.3.1 Speakers’ item descriptions

Generally, speakers’ descriptions were quite detailed and elaborate,
presumably in an effort to be helpful to their imagined future listeners (cf. Schober,
1993). An example of a complete trial transcript is reproduced in Appendix D, but
the following description is fairly typical (item onset and name onset are marked with asterisks):

"... *Number 3 is above the door; I think it might be the *fire alarm. But it basically looks like a white box just above the doorframe... um, made of plastic, I would assume. ... *Number 4 is at the base of the door..." (Speaker 2, bedroom)

In the absence of feedback from their listeners, speakers may frequently have been over-helpful: several listeners later complained that they usually located the relevant object long before the speaker finished describing it. In addition, disfluencies, repetitions, and self-corrections were quite frequent, and occasionally objects were actually misidentified (e.g., the laptop in the bedroom was often described as a TV).

The freedom given to speakers in how to describe the images, as well as individual differences in their speech speed, fluency, and level of detail led to large differences in trial duration between the six speakers. This is reflected in Table 4-1, which summarises the mean duration for each speaker of items (i.e. one object description) and trials (one room; 12 items). It also includes mean name-onset latency (the time from item onset to unambiguous reference), which showed similar variability. The order of presentation of the three rooms and of the items within each room did not show any consistent effect on latencies and durations.

Table 4-1: Descriptive statistics per speaker: mean item duration, name-onset latency, and trial duration, in seconds.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Mean item duration</th>
<th>SD</th>
<th>Range</th>
<th>Mean name-onset latency</th>
<th>Mean trial duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>14.5</td>
<td>4.5</td>
<td>18.6</td>
<td>8.0</td>
<td>164.0</td>
</tr>
<tr>
<td>2</td>
<td>16.0</td>
<td>6.3</td>
<td>28.5</td>
<td>5.7</td>
<td>181.1</td>
</tr>
<tr>
<td>3</td>
<td>9.0</td>
<td>3.8</td>
<td>16.2</td>
<td>2.5</td>
<td>102.6</td>
</tr>
<tr>
<td>4</td>
<td>9.9</td>
<td>3.6</td>
<td>17.4</td>
<td>2.1</td>
<td>113.0</td>
</tr>
<tr>
<td>5</td>
<td>8.3</td>
<td>4.3</td>
<td>22.0</td>
<td>2.8</td>
<td>94.0</td>
</tr>
<tr>
<td>6</td>
<td>5.0</td>
<td>1.2</td>
<td>4.6</td>
<td>1.7</td>
<td>59.0</td>
</tr>
</tbody>
</table>

| Total   | 10.4               | 5.6 | 32.8  | 3.8                     | 119.0               |

In summary, an item usually lasted between 5 and 15 seconds, and total trial duration varied correspondingly between about one and three minutes. Speakers 1 and 2 usually took significantly longer than the other speakers for their descriptions, which were very detailed (difference in item duration between Speakers 1 and 4:}
\( t(64) = 4.55, p < .001 \). Nonetheless, they tended to refer to the object unambiguously only about halfway through the description. Speaker 6, in contrast, was very concise and significantly quicker than anyone else (mean item duration compared to Speaker 5: \( t(64) = 4.19, p < .001 \)). The other three speakers fell between these extremes (all \( p > .1 \)). For Speakers 3-6, name onset was proportionally earlier: the object was usually referred to unambiguously within the first third of the item description.

### 4.3.3.2 Speakers’ fixation patterns

Several questions are of interest when considering the speakers’ gaze behaviour. On the assumption that speakers are highly likely to look at an object if they spend ten seconds or more describing it, the extent to which any fixations were registered in the appropriate ROI during the description of a particular item is a useful indication of recording accuracy. Despite the measures described above, this was far from ideal: across speakers, no fixation was registered to the relevant ROI in 46 of the total of 214 items (21.5 %). In most cases, this was due to fixations being recorded somewhat above the actual object. This general trend suggests a fault in the recordings – presumably some form of drift – rather than a characteristic of some speakers not to fixate objects directly. The number of items per speaker which received such a relevant fixation, shown in Table 4-2, is therefore taken as an indication of recording quality. According to this measure, Speakers 1 and 6 clearly produced the worst recordings.\(^{19}\)

<table>
<thead>
<tr>
<th>Speaker</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>47.1 %</td>
<td>100 %</td>
<td>77.8 %</td>
<td>100 %</td>
<td>86.1 %</td>
<td>58.3 %</td>
<td>78.5 %</td>
</tr>
</tbody>
</table>

\(^{19}\) Unfortunately, the extent of these missed fixations was only discovered at a later point in time, otherwise these two speakers would have been replaced. However, the gaze cursor still pointed in the right direction, even if it was not accurately focussed on the object.
So assuming that speakers generally fixate the objects they are describing, when do these fixations occur? Answering this question requires looking at the gazes preceding *name onset*, rather than those before *item onset*. For the latter, speakers had little choice but to look at the number superimposed on the object to know which item was next; so the first fixation to the object is not necessarily related to producing its name and description. Figure 4-2 shows that despite the complexity of the scenes and the huge variation in timing and descriptiveness between speakers, the eye-voice span is fairly typical of language production, if not very pronounced: fixations to the current object increase gradually and peak roughly 800ms before name onset.

![Figure 4-2: Regressive time graph of fixation proportions before name onset (name onset = 0 ms). For the faster speakers, item onset would be roughly at the onset of the graph.](image)

Another aspect worth exploring is the occurrence of preview-fixations to the following item during the description of the current one. These occurred about half the time, for 49.3% of items across speakers. Preview fixations have important implications for listeners' use of the gaze cursor: they imply that for half of all items, listeners who could see the cursor received a clue to the next item before there was any possibility of predicting its identity on the basis of the speech utterance. However, the usefulness of this clue would depend on how frequently speakers also looked at other, not-to-be-mentioned objects. In this context it is interesting to note that speakers did not show comparably increased fixations of the previous item (15.1% of items). In fact, most preview-fixations occurred towards the end of the
current description: 91 % began less than five seconds before the onset of the next item (mean start time = 942 ms before next onset, SD = 853). This suggests that speakers generally finished processing the object they were describing before shifting their attention to the next one.

Actually, an interesting finding in this respect is that for half the item descriptions speakers didn’t look at the next little number before mentioning it (i.e. starting a new item). It seems likely that in these cases they were trying to keep their speech fluent by mentally counting up to the next number and then using the time between saying “number x is” and the start of the actual description to locate the next object. Alternatively, they may have noticed it earlier and remembered it, or have seen it in the parafovea.

With these characteristics of the speaker’s speech and gaze stimuli in mind, it is now time to return to the listeners. The following sections describe the actual experimental procedure and the results of Experiment 2.

4.3.4 Listener methods

4.3.4.1 Participants

The actual experiment was run with 12 listeners (2 males), two for each speaker: one could see the projected gaze cursor, the other could not. Mean age was 21.8; two participants were replaced due to experimental error or poor calibration. Since the pilot Spot-the-Difference study (see Appendix C) had found substantial differences in object naming for different age groups and varieties of English, listeners, like speakers, were all native British English undergraduates.

4.3.4.2 Procedure

In most respects, the procedure for listeners was identical to the one described for speakers. In the following, I report only those aspects which differed. Listeners were told that they would hear a description of each room and that their task was to mouse-click on the objects “as soon as you have identified which object the speaker is referring to”. Listeners who saw the gaze cursor were also told: “Simultaneously, a little green dot will display eye movements that were recorded
from the speaker while they were looking at and describing the scene.” The order of the trials was the same as the order in which the speaker had originally described them.

During the experimental trials, listeners’ displays did not include numbers superimposed on the objects. However, to avoid pre-exposing them to the gaze cursor before measuring its effect in the actual experiment, they were given the same practice task as speakers: the classroom picture with six numbered objects to describe, not to click on. This had the advantage of providing them with a realistic idea of how the speaker had come up with the descriptions they heard during the actual experiment.

For the main experiment, listeners were eyetracked in pupil-only tracking mode (500 Hz) while mouse-clicking on the objects. Calibration was repeated before each trial until the average error was < .5° and maximum error < 1°. Following the end of the last description of each trial, a central fixation point was automatically projected on-screen for seven seconds. The audio playback was synchronised to begin at the onset of the display. It continued uninterrupted until the end of the last description of the trial. This meant that for listeners who saw the gaze cursor, the speaker’s audio file was replayed in sync with his or her eye movements. All listeners saw the position of their own mouse indicated by a red triangle. Timing and position of clicks were recorded into the .edf-file.

After clicking on the objects, listeners were given a surprise recall task for the three scenes, in the same order as their original presentation. They were asked to name out loud any “clicked” objects they could remember, in any order. Their speech was recorded to a .wav-file. To remind them of the presentation order, the name of each room appeared in the centre of the screen for 500 ms. Then the screen changed to blank grey for a maximum of three minutes, or until the participant pressed a button to indicate they could think of no more objects. Participants were eyetracked during this task, looking at the blank grey screen. A 9-point calibration routine preceded each recall trial until validation was labelled “good”.

Having completed the recall task for all three displays, the eye tracker was removed. Participants then received a colour A4-printout of each scene in a plastic cover, again in the order in which they were originally presented. They were told
that the new pictures differed from the original scenes in eight details (see Appendix C), which they should identify as quickly as possible by circling and numbering them on the transparency. They were timed during this task, again with a three-minute limit per scene. Finally, listeners were asked for demographic information and debriefed. In total, the experiment took about 50 minutes, for which they received £5.

4.3.4.3 Data processing and Analysis

The list of positions for listeners’ mouse clicks was manually compared to the correct regions of interest described in section 4.3.2.2 (see also Appendix B). In cases where listeners first clicked on the wrong object and then corrected themselves, only the correct click was used in subsequent analyses. With the exception of these rare corrections (< 1% of cases), click accuracy was 100%.

Fixations were automatically categorised with respect to the region of interest they fell in. However, the fact that all items were visible throughout the entire trial made it necessary to limit which fixations were counted as relevant to an item. For instance, a chance fixation to Item 12 right at the start of the trial, or while searching for Item 3, could have a strong but spurious effect on first fixation onset time. This would be particularly likely to affect listeners without a gaze cursor, who spent more time searching around the image for the next potential object. For this reason, only fixations to a specific ROI occurring during the description of the corresponding item (i.e. either spanning or beginning after the word “number”) were counted as relevant fixations and included in the analysis.

Initial analyses of the raw data also showed that it was important to correct fixations for drift. They suggested that it took even listeners who could see the gaze cursor an average of three seconds or more to reach a relevant item after the speaker began to say “number x is…” This is implausibly long, considering that some prediction would be possible by that time on the basis of speech alone, and that listeners with a gaze cursor might even have seen the speaker fixating the object before uttering “number”, i.e. before the start of the trial. Further investigations revealed that these latencies were at least partly an artefact of the large amount of drift occurring over the course of the trial: since listeners fixated the surroundings of an object quite a lot during its description, at least one of these gazes would usually
end up in the relevant ROI. However, drift over the course of the trial would mean that the first gaze to be registered was not necessarily the first true gaze to that object. In some cases, this might lead to earlier-than-true first fixations, though later-than-true would be more likely, considering that speakers tended to fixate both on the object itself and also in its vicinity during its description.

For this reason, a trial-final central fixation target was presented for seven seconds following the twelve descriptions for each room. For technical reasons, this was not an EyeLink drift correction target, so it was not possible to adjust fixations automatically. Instead, all fixations of every trial were manually shifted by the amount required to accurately align the final gaze with the central target, which participants were fixating at this point. Despite the trade-off that fixations to the first items in each room (for which the trial-initial calibration may have been more valid) became less accurate, this procedure led to a considerable reduction of items for which no relevant fixation was registered at all. This shifting of fixations with reference to the end of the trial was therefore adopted for all trials and all listeners.20

4.4 Listener results

The primary dependent measures of interest are based on listeners’ eye movements and mouse clicks, which are discussed in Sections 4.4.1 and 4.4.2, respectively. These are followed by a description of listeners’ performance in the memory tasks (4.4.3). The chapter concludes with a summary and brief discussion of the results (4.5).

4.4.1 First fixation onsets

Recording accuracy for listeners was much better than for speakers, presumably due to the more passive task: no relevant fixation was recorded for only 1.85 % of items across listeners (unsurprisingly, the worst performance was for the

20 Reassuringly, most analyses based on the raw data show the same effects as those reported using shifted fixations; only their size is affected.
listeners assigned to Speaker 1: 6.94 %). The number of missing fixations did not differ between cursor conditions ($p = .81$). For all analyses relating to fixation onsets, outliers and extreme values were excluded based on a cut-off of $+/- 2$ SD outside the mean per listener. This resulted in a total of 18.98 % missing data points.

The onset time of the first relevant fixation to each item was compared in a repeated-measures design between the listener with gaze cursor and the corresponding listener without, relative to the onset of the item description by the speaker. Despite the fact that the fixations were actually recorded from two different listeners, this experimental matched-pairs design on the basis of the speaker removed the variance resulting from the large differences in timing and descriptiveness between speakers (cf. B. H. Cohen, 1996, p. 433)\footnote{Unless otherwise mentioned, all results for Experiments 2, 3, and 4 are based on such matched-pairs analyses. However, supplementary analyses using traditional independent ANOVA confirmed the principal results reported here.}. According to the predictions outlined above, the relevant object was expected to be fixated earlier by listeners who could see the speaker’s gaze than by listeners who only heard the description. As Figure 4-3 shows, this was indeed the case.

![Figure 4-3](image)

**Figure 4-3:** Mean onset of the first relevant fixation, depending on the presence of the gaze cursor (“dot”). The three bars in each condition reflect the order of the three trials.

The 2-way repeated-measures analysis of variance with the factors CURSOR (presence or absence) and TRIAL (1\textsuperscript{st}, 2\textsuperscript{nd}, or 3\textsuperscript{rd} room seen) revealed a main effect of
CURSOR ($F1(1, 5) = 51.79; F2(1, 25) = 85.50$, both $p < .001$). Listeners with a gaze cursor fixated items much earlier ($M = 1093$ ms from item onset, $SD = 1539$) than those without ($M = 3642$ ms, $SD = 2226$). Interestingly, there was no effect of TRIAL, nor was there an interaction. This may suggest that listeners did not need much time to adapt to the gaze cursor, but made use of it consistently from the first trial onwards (though there could still have been an effect on the first items). The same implication can be drawn from the greater variance in first fixation onsets for those participants who saw no cursor, as opposed to those who did. The cursor generally attracted the listener’s gaze to whatever it was pointing to, while without a cursor, fixation locations varied much more widely.

How likely were listeners in the two groups to fixate the correct object before the speaker had unambiguously referred to it? Figure 4-4 compares the number of gazes beginning before and after name onset (cf. Section 4.3.2.3) for the two cursor conditions.

![Figure 4-4: Proportion of first gazes per cursor condition to the relevant object, beginning before and after name onset.](image)

Clearly, listeners who saw the gaze cursor were much more likely to first look at the object before hearing it described unambiguously than afterwards (on average 1719 ms before name onset, $SD = 1408$). The opposite was true for listeners without a cursor: while they sometimes managed to locate the object before its mention, more often they fixated it only afterwards ($M = 678$ ms after name onset,
The difference in frequencies is significant ($\chi^2 (1, N = 366) = 107.30; p < .001$).

### 4.4.2 Click latencies

Click latency, the time from item onset to click on a described object, was used as a measure for listeners’ understanding of the speaker’s utterance. A difference in mean click latencies across cursor conditions would suggest that the gaze cursor is integrated into the processing of the speaker’s utterance, beyond a simple low-level attraction of visual attention. Another interesting finding would be an interaction, such that latencies become shorter for later scenes or later items in the presence of the gaze cursor. This would indicate that listeners can make use of the cursor, but first need to learn how it relates to the speaker’s description. Figure 4-5 presents mean click latencies depending on the cursor condition for the three trials.

![Click latency graph](image)

**Figure 4-5: Click latency depending on the presence of the gaze cursor, for the three trials.**

A repeated-measures ANOVA with the factors CURSOR and TRIAL revealed a main effect of CURSOR on click latencies, $F(1, 5) = 6.35, p = .05$; $F(1, 35) = 90.95, p < .001$. Listeners with a gaze cursor clicked on the correct object almost two seconds earlier ($M = 5098$ ms, $SD = 3238$) than listeners without a gaze cursor.
The effect of Trial was not significant, nor was the Cursor*Trial interaction.

However, considering that mean name onset occurred about four seconds from item onset, these long click latencies suggest that while the visual cue may have reinforced listeners with cursors to feel confident in their judgement earlier, listeners in both conditions usually waited to click until they received linguistic confirmation of the object. A comparison of the number of correct clicks before name onset confirms this: overall, clicks before name onset occurred very rarely, in only 34 cases (7.9%). However, they were significantly more frequent for listeners with a cursor (6.9%) than for listeners without (0.9%); χ²(1, N = 432) = 21.58; p < .001. Interestingly, with one exception all clicks before name onset occurred for Speakers 1 and 2, whose descriptions were longer and more detailed than other speakers’.

### 4.4.3 Memory tasks

Accuracy in the two offline memory tasks, Recall and Spot-the-Difference, was determined by counting the number of correctly named or circled objects, and comparing these figures between cursor conditions. For both tasks, large individual differences were found for the number of items recalled and differences spotted, as Table 4-3 shows: some participants recalled not a single item from a room, while others remembered 11 out of 12. However, no correlation was found between individuals’ Recall and Spot-the-Difference rates (Kendall’s tau_b (12) = .17, p = .48), nor between the memory tasks and the time taken to complete them (Recall: tau_b (12) = .32, p = .16; Spot-the-Difference: tau_b (11) = .02, p = .94).  

---

22 Kendall’s tau is a measure of correlation used for small samples with many tied ranks (Field, 2005). The same results were found using the more familiar Spearman’s rho.
Table 4-3: Number of correct responses in the memory tasks, across the three trials.

<table>
<thead>
<tr>
<th></th>
<th>Mean correct</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall max. 36</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cursor</td>
<td>20.17</td>
<td>3.06</td>
<td>18</td>
<td>25</td>
</tr>
<tr>
<td>no cursor</td>
<td>22.83</td>
<td>5.64</td>
<td>16</td>
<td>29</td>
</tr>
<tr>
<td>total</td>
<td>21.5 (60 %)</td>
<td>4.54</td>
<td>16 (44 %)</td>
<td>29 (81 %)</td>
</tr>
<tr>
<td>Spot-the-Difference max. 24</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cursor</td>
<td>12.67</td>
<td>1.75</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td>no cursor</td>
<td>15.5</td>
<td>1.57</td>
<td>14</td>
<td>18</td>
</tr>
<tr>
<td>total</td>
<td>14.08 (59 %)</td>
<td>2.15</td>
<td>10 (42 %)</td>
<td>18 (75 %)</td>
</tr>
</tbody>
</table>

According to the predictions, items should have been recalled better by listeners who had seen the gaze cursor, could therefore fixate the relevant objects earlier, and so had more time to process them. If anything, however, listeners who had not seen the cursor recalled slightly more objects than those who had, though the correlation was not significant ($\tau_b (12) = -.22, p = .41$). There was also a marginal trend for more objects to be recalled from more recent trials (Trial 1: $M = 5.83$, Trial 2: $M = 7.33$, Trial 3: $M = 8.33$; $\rho = .30$, $N = 36$, $p = .08$).

In the Spot-the-Difference task, differences were predicted to be easier to spot for listeners with a gaze cursor, particularly if the difference concerned an object they had clicked on. In contrast, differences regarding objects that had not been described might be more apparent to listeners who had not seen the cursor. Generally, more changes were discovered to clicked items ($M = 9.33$ out of 12, $SD = 1.72$) than to new items ($M = 4.92$, $SD = 1.88$; Wilcoxon Signed Ranks $Z = 3.07$, $p < .01$). In addition, listeners without a gaze cursor discovered more differences overall ($M = 7.83$) than those with a cursor ($M = 6.42$; correlation between cursor and spotted differences: $\tau_b (12) = -.62, p = .02$). However, there was no indication of an advantage for new objects without a gaze cursor ($\tau_b (12) = -.26, p = .33$).

Finally, one of the most interesting questions was whether listeners who saw the gaze cursor would repeat more of the speaker’s referring expressions in the recall task. Addressing this issue required determining all the items which were correctly recalled for both groups of listener, as well as for speakers (listeners with cursor: 122 items correct out of a total of 216 (56.5 %); without cursor: 137 correct
(63.4 %); speakers: 152 correct (70.4 %)). Figure 4-6 shows the extent to which each of these groups repeated the referring expressions from the original descriptions.

![Figure 4-6: Re-use of speakers' original referring expression for listeners with and without cursor, and for speakers. Items that were not recalled at all were excluded.](image)

In general, there was a strong tendency to repeat speakers’ expressions, despite the fact that they were uttered/heard approximately 20 minutes earlier and in a very different task. When correctly recalling an item, speakers showed a 74.3 % likelihood of repeating the original expression, while listeners did so in 60.6 % of cases. This repetition was not just due to the fact that everyone used the same names anyway: in the original descriptions, only eight of the 36 items (22.2 %) were described using the same expression by all six speakers. However, there was no difference in repetition rate between listeners with a gaze cursor (59.8 %) and those without (61.3 %; $\chi^2(1, N = 259) = .06, p = .8$), which would have been predicted if visual coordination led to referential alignment.

### 4.5 Summary and Conclusions

Experiment 2 served as a preliminary investigation into the effects of seeing a speaker’s projected gaze while hearing what he is saying. How does this affect the eye movements of listeners, their language comprehension, and even their subsequent production? To explore these questions, listeners saw three photographs
of scenes and were instructed to click on the objects they heard a speaker describe. During this task, the presence of a gaze cursor representing the focus of the speaker’s gaze was varied between listeners. Subsequently, they received a surprise recall test for the objects that had been described. Finally, they were asked to compare a new version of the scene to the one they had originally seen and to spot any differences between the two.

Seeing the speaker’s gaze led to much earlier first fixations to the relevant objects than just listening to the speaker’s speech, and to less variance in first fixation onset time. Clearly, if the gaze cursor is present, listeners follow it closely with their own eyes and spend less time looking around. More interestingly, the presence of the gaze cursor also helped participants to identify and click on the objects earlier, suggesting that listeners are willing and able to integrate the visual cue with the auditory information for this language-based task. At the same time, the fact that almost all clicks were made only after linguistic disambiguation reflects how strongly listeners relied on the speech utterance; it seems they waited for the speech to confirm what they really already knew before committing to an object. This is supported by the difference between first fixations and click latencies in the size of the effect: the availability of the gaze cursor reduced the former by a massive 2549 ms, compared to a smaller (though still sizeable) effect of 1833 ms on click latencies.

No clear result emerged from the memory tasks. Unexpectedly, listeners who had not seen the cursor outperformed those who had, at least in the Spot-the-Difference task. This could be due to the fact that they spent more time looking around the displays and therefore encoded the objects better, but it may also be caused by the combination of a small sample size and large individual differences in performance. Certainly, these factors reduced the suitability of the recall task for investigating lexical repetition based on visual coordination, since only items that were correctly recalled could be compared. Tighter experimental control of speaker’s expressions, and allowing the listener to see the display again during recall, might be a more appropriate way of testing these issues.
Another analysis which was not conducted here would be to examine the fixation locations of listeners in the recall task. Several current theories and strands of research allow predictions relating to gaze behaviour in this task. For instance, studies using the blank screen paradigm find that people tend to fixate object locations even if the object in question is no longer present on the screen (e.g., Richardson & Spivey, 2000; Spivey & Geng, 2001; Altmann, 2004; Knoeferle & Crocker, 2007). Indeed, it has recently been suggested that looking at the correct scene location might actually help participants to recall an object that had been there originally (F. Ferreira et al., 2008; Kent & Lambert, 2008). From this perspective, listeners in front of the blank screen would be expected to move their gaze back to where they originally saw and clicked on the object they are trying to recall. In fact, if looking back to the original location really benefits recall, listeners who were drawn there by the gaze cursor in the clicking task and consequently spent a long time processing the item might also recall more objects than those who didn’t see the gaze cursor.

From a different perspective, the Interactive Alignment Model (Pickering & Garrod, 2004) and other dialogue-oriented research (Richardson & Dale, 2005; Richardson et al., 2007) would also suggest that successful recall would correlate with looking back to where the object used to be. However, here the argument is based not so much on the fact that the object was previously seen there by the listener himself, but because the speaker originally looked at it there. According to this line of thought, such visual coordination between speaker and listener should be beneficial to language comprehension and dialogue success. Clearly then, an analysis of blank screen fixation locations during recall would be interesting for several reasons, though it does not form part of this thesis, due to time constraints.

Without a doubt, Experiment 2 also suffered from several methodological problems: inaccurate speaker recordings, drift correction of the listener fixations, non-normal distributions, and massive variance including extreme values, to name just some. Many of these practical issues will be addressed in detail in Chapter 6. First though, Experiment 3 will explore a more theoretically relevant question: is the

23 The main reason this was not done was that a preliminary exploration of the data showed that many participants hardly looked at the screen at all: fixations were clustered at the very edges of the screen or missing completely.
gaze cursor actually interpreted as gaze, or is it just a salient visual cue which happens to be in the right place at the right time?
Chapter 5: Varying instructions about the gaze cursor – Experiment 3

5.1 Research question and predictions

This chapter extends the introduction of the gaze projection paradigm by exploring the importance of listeners’ interpretations of the gaze cursor. Experiment 2 showed that a projection of the speaker’s gaze leads listeners to fixate the relevant objects earlier, and also enables them to click on the correct object more quickly than if no gaze cursor is presented. However, the reason for this is not clear.

On the one hand, many previous studies have found that humans make extensive use of non-linguistic cues – such as a speaker’s gaze – to help infer their interlocutors’ meanings (Baldwin, 1993; Yu et al., 2005), and especially for disambiguating referents (e.g., Clark et al., 1983; Hanna & Brennan, 2007). Several recent papers have also shown quite convincingly that a super-imposed gaze cursor can be processed in a similar way to real gaze in a face-to-face situation (e.g., Yu et al., 2005; Brennan et al., 2008).

On the other hand, there will be many tasks where the connection between the gaze cursor and real gaze is less obvious: processing a bright green dot moving on the screen is after all a very different task to looking at another person’s face, detecting the direction of their eye gaze, and deducing what they are looking at. Therefore, while there is no doubt that the gaze cursor in this task is helpful to listeners, this need not necessarily be because they interpret it as the speaker’s gaze (see Chapter 3, Section 3.2.2, above). Instead, the gaze cursor’s visually salient, attention-grabbing properties could be attracting listeners’ attention to wherever it is on the screen; because it is linked to what the speaker is saying, this would be useful of itself. For example, if the speaker were speaking about the eggs in the kitchen photo, the gaze cursor would be moving around the bottom right-hand corner of the image. If the listener followed it there, she would probably stumble...
over the eggs much earlier than if she had begun to look for them near the fridge or by the sink. In this way, the onset time of her first fixation and her click latency could both be reduced, without any need for her to have considered the relationship between the cursor and where the speaker was looking.

Experiment 3 was therefore designed to explore the extent to which listeners actually interpret the gaze cursor as the speaker’s gaze. Is it important that they are aware of this connection? This question was investigated by providing listeners with different instructions about what the green dot represented. If listeners are aware of the link between speakers’ language-based attention and eye movements, and if they are able to generalise from the cursor to where a speaker is looking, they should be faster to fixate and click if they believe the cursor to be an accurate representation of these eye movements, than if they have reason to doubt the information it contains. Since the speech and gaze cursor are identical in both conditions, such an effect of doubting the validity of the cue would be a strong rejection of the possibility that the gaze cursor is just helpful because of its visual properties.

5.2 Methods

Experiment 3 used the speech and gaze stimuli that were recorded for Experiment 2, and in almost every respect the Method was identical to the ‘with cursor’ condition in that study.

5.2.1 Design

In Experiment 3, all listeners saw the gaze cursor. The two experimental conditions differed only in what participants were told the dot represented: either the speakers’ eye movements while making the description (‘speaking’ condition), or while remembering the objects later (‘remembering’). Again, each speaker was assigned two listeners, one in each condition, so the instructions were varied between listeners but within speakers. Results were again analysed as a matched-pairs design (see Section 4.4.1).


5.2.2 Materials

The critical manipulation in this experiment was in the two instructions. On the one hand, for the instructions to cast doubt on the informativeness of the cursor, the reason for mistrusting and potentially ignoring it had to be fairly evident to participants. On the other hand, because the cursor really was linked to the speaker’s speech, there would always be a fairly obvious connection between the two. As a result, a blatantly wrong description (e.g., “the cursor is based on eye movements when describing a different scene”) seemed too noticeably false. For this reason, a cover story was set up. As part of the general introduction, all participants were first told:

... Speakers were recorded in a previous experiment. They had two tasks: first, they saw the scene and described the twelve numbered objects for future listeners. The numbers were then removed. Speakers then looked at the objects again, but this time without speaking. They had to try to remember how they had described the objects. ...

From this point on, the actual instructions were varied between participants – either:

... In the main experiment, a little green dot on the screen will display eye movements that were recorded while the speaker was performing the first task: looking at the picture and describing the set of objects.

or:

... In the main experiment, a little green dot on the screen will display eye movements that were recorded while the speaker was performing the second task: looking at the picture without speaking, trying to remember how they had previously described the set of objects.

Before each trial, participants also received an on-screen reminder of what was allegedly the basis for their gaze cursor. In the correct, speaking condition, they were told “The moving green dot on the screen shows where the speaker was looking during the first task: describing.” In the ambivalent, remembering condition, the instruction read: “The moving green dot on the screen shows where the speaker was looking during the second task: remembering.”

At the end of the experiment, a paper-and-pencil questionnaire served as a manipulation check. It is reprinted in Appendix E and contains rating questions about how helpful the cursor was felt to be, and a direct query about which condition participants believed they had been assigned to.
5.2.3 Participants

Twelve listeners (4 males) participated in Experiment 3. Pairs of two listeners were assigned to each of the six Experiment 2 speakers. All were right-handed, spoke British English as a native language, and came from the University of Edinburgh student community. Mean age was 21.4 years; one participant had to be replaced because he failed to comply with the experimental procedure. None of the participants had taken part in Experiment 2.

5.2.4 Procedure

Experiment 3 was conducted in the same soundproof chamber as Experiments 1 and 2. Before being fitted with the eyetracker, participants were presented with a full-length paper version of the instructions described above (Section 5.2.2). Following an opportunity to ask questions, the head-mounted EyeLink 2 tracker was then set up and calibrated. The eye with least error at validation was selected for monocular recording.

To make both instructions salient and understandable, all participants received two different practice trials, corresponding to the two tasks the speakers had allegedly performed. First, they saw the photo of a classroom that had been used as a practice item in Experiment 2, with small red numbers superimposed on six objects. For the speaking task they were asked to “describe the six numbered objects so that someone who can’t see the numbers would be able to identify the objects and mouse-click on them” (speaking task). Next, the image re-appeared without the little numbers, and participants were told to “look at the objects in turn, and try to remember how you described them” (remembering task).

The practice trials were followed by the main experiment, which consisted of the same three trials and descriptions as in Experiment 2, presented in the order in which the speaker had originally described them. Each trial was preceded by an instruction screen, asking participants to mouse-click on the objects “as soon as you have identified which object the speaker is referring to”, and reminding them whether their gaze cursor displayed the speaker’s eye movements during ‘speaking’ or during ‘remembering’. In reality, the two listeners assigned to one speaker saw
exactly the same gaze cursor. Like the audio playback, it was synchronised to begin at the onset of the display and continued uninterrupted throughout the trial. Participants were eyetracked in pupil-only tracking mode while mouse-clicking on the objects. Again, the position of the mouse was indicated by a red triangle, and timing and clicks were recorded into the .edf-file. Calibration was repeated before each trial to minimise recording error. Following the last description of each trial, participants were asked to fixate a central target for seven seconds.

There were no memory tasks in Experiment 3. Instead, the eyetracker was removed after the three trials, and participants completed the post-experimental questionnaire (see Appendix E). Among other things, this contained a question alerting them to the experimental manipulation, which was also explained in more detail once they had finished. The total experiment lasted about 45 minutes, for which participants received £5.

5.2.5 Data processing and analysis

All participants’ mouse clicks were manually compared to the ROIs described for Experiment 2 (see Section 4.3.2.2, above). Only correct clicks were used in subsequent analyses. Self-corrections were rare (< 1 % of items), and apart from these, click accuracy was 100%. In addition, all participants’ fixations were manually drift-corrected with reference to the trial-final fixation point, as described for Experiment 2; they were then automatically categorised with respect to the ROI they fell in.

Since participants in both conditions of Experiment 2 saw the same gaze cursor, fewer problems with differences in variance between conditions were expected than in Experiment 2. This made it possible to extend the definition of first relevant fixation to include any fixation to the correct object occurring between the onset of the previous item and throughout the following one, thereby allowing the inclusion of gazes to the item prior to the onset of its description. Such gazes seem likely considering the substantial number of preview fixations made by speakers; in such situations, listeners would be likely to follow the cursor to the upcoming item well in advance of its mention. This new definition of relevant fixations seemed
sensible because listeners who did look at the object before item onset might well make several other fixations before refixating it, thus inflating the mean first fixation time. However, the downside of this wider definition is that it will also include more spurious fixations, which are not yet connected with the speaker’s description of the object and corresponding gaze behaviour. According to this definition, no relevant fixation was recorded in 24 cases (5.6%). The two instruction conditions did not affect the number of missing fixations ($p = .67$).

### 5.3 Results

Three measures were of interest in Experiment 3, any or all of which might be affected by the different instructions regarding the gaze cursor: onset times of the first relevant fixation for each item (5.3.1), click latencies to identify the objects (5.3.2), and participants’ responses to the questionnaire (5.3.3). All three measures were compared in a repeated-measures design between the two participants who had listened to the same speaker (one with the ‘speaking’, the other with the ‘remembering’ instruction).

#### 5.3.1 First fixation onsets

For analyses relating to fixation onsets, outliers and extreme values were excluded based on a cut-off of +/- 2 SD outside the mean per participant (6.1% of items), resulting in a total of 11.3% missing data points, including where no relevant fixation was recorded. The onset of the first relevant fixation to each item was compared between listeners with ‘speaking’ and ‘remembering’ instructions, relative to the onset of the item description by their speaker. If participants with a ‘remembering’ instruction were able to disregard their potentially less reliable gaze cursor, the prediction is that they would take longer to make their first fixation to the objects. However, as Figure 5-1 suggests, this was not the case.
The 2-way analysis of variance with the factors INSTRUCTION (‘speaking’ or ‘remembering’) and TRIAL (the order of the display presentation) found no reliable main effect of INSTRUCTION ($F_1(1, 5) = 1.77, p = .24; F_2(1, 33) = 4.93, p = .03$). If anything, first fixations were slightly earlier for participants who were told that the gaze cursor was recorded during ‘remembering’ ($M = 501$ ms from item onset, $SD = 2302$) than during ‘speaking’ ($M = 857$ ms, $SD = 2624$). There was no effect of TRIAL and no interaction (all $p > .3$).

It is interesting to note that, using the new definition of relevant gaze (see Section 5.2.5, above), participants’ first fixations appear somewhat earlier here than in Experiment 2. Overall, the average onset of the first relevant fixation occurred 677 ms from item onset ($SD = 2469$), as opposed to 1093 ms ($SD = 1539$) in the gaze cursor condition of Experiment 2. An analysis of Experiment 3 using the same relevant gaze definition as in Experiment 2 produced a much more similar mean first fixation onset of 1148 ms across conditions ($SD = 1511$), but did not affect the pattern of results (though the by-items effect of INSTRUCTION was no longer significant, $p = .62$).
5.3.2 Click latencies

As in Experiment 2, click latency to the correct item served as a measure for listeners' understanding of the speaker's utterance. While Experiment 2 showed that participants generally waited to click on an object until it had been confirmed linguistically, that study also discovered a substantial advantage for listeners with a gaze cursor. A similar difference between conditions might be found in the current study if participants in the ‘remembering’ condition were less inclined to trust the gaze cursor than those in the ‘speaking’ condition. Figure 5-2 presents mean click latencies depending on the instruction.

![Figure 5-2: Mean click latency depending on the gaze cursor instruction and trial number.](image)

Clearly, the instruction did not affect click latency in the least: the mean time to click on the correct objects was 5312 ms from item onset ($SD = 3573$) in the ‘speaking’ condition, compared to 5253 ms ($SD = 3482$) for ‘remembering’. The ANOVA confirmed the absence of any effect of INSTRUCTION (both $p > .6$). There was also no effect of TRIAL and no interaction (all $p > .5$).
5.3.3 Questionnaire results

In view of the null results for both online measures, the post-experimental questionnaire data is particularly important for this experiment. The questionnaire was designed to capture participants’ subjective intuitions about the gaze cursor. As a result, it is informative about the extent to which participants drew the intended conclusion from the alleged manipulation, and in this way may help to explain the null effects results of the fixation and click results. As can be seen in Appendix E, the questionnaire contained both qualitative and quantitative items. The results are summarised in the following, mainly with regard to any light they can throw on participants’ processing of the gaze cursor.

In Question 1, participants described the study they had just taken part in. Their descriptions were subsequently coded as to whether they included a reference to the gaze cursor and what it stood for. Despite frequently quite detailed descriptions, a third of all participants (3 ‘speaking’ [= SP], 1 ‘remembering’ [= RM]) did not mention the cursor at all, suggesting that it was not the most memorable aspect of the experiment. Of those who did mention it, the three remaining participants in the ‘speaking’ condition all referred to the dot correctly, e.g. “there was a green dot on the screen to indicate where the ‘describer’ had been looking” (SP6). However, at least one participant in the ‘remembering’ condition also stated clearly that “a dot represented the person’s eye movements as they described the objects” (RM3). In fact, only two of the ‘remembering’ participants unambiguously described the gaze cursor as ‘remembering’ (e.g., “My role was to identify the objects participants had previously described using their verbal descriptions, taking into account their eye movements when remembering their descriptions” (RM4)). Unfortunately, this suggests that the ‘remembering’ instruction was not particularly convincing.

Questions 2 and 3 concerned the usefulness and clarity of the speaker’s verbal descriptions. They were intended mainly as foils to detract attention away from the importance of the gaze cursor, but the high average ratings in both conditions (SP: $M = 6.2$, RM: $M = 6.0$ on a 7-point scale, difference not significant) showed that participants experienced no difficulties understanding their speakers.
Similarly, Questions 4 and 5 concerned participants’ perceptions of the gaze cursor. In response to Question 4 – rating its usefulness – five out of the six ‘speaking’ participants rated it as “extremely useful” (7 points), though one found it only “somewhat useful” (4 points). In contrast, all ratings in the ‘remembering’ condition were high but usually not maximal (M = 5.8). A Wilcoxon Signed-Rank test between conditions was not significant (z = -1.3; p = .36.), though this is clearly due to the outlier in the ‘speaking’ condition. The effect size was moderate (r = -.38), suggesting that there were probably not enough participants for this analysis. No difference at all was found between instructions for Question 5: How tightly would you say the dot was linked to what the speaker was saying? (both M = 6.0). Obviously, participants in both conditions noticed that speech and gaze cursor were tightly related.

Questions 6-8 explored participants’ strategies concerning the gaze cursor. Generally, participants in both conditions tended not to ignore it (SP: M = 2.8, RM: M = 3.2), tried to follow it most of the time (SP: M = 5.7, RM: M = 5.3), and – apart from two outliers in the ‘speaking’ task who rated it with 5 points – usually didn’t find the cursor distracting (both M = 2.7). None of the differences between instructions reached significance.

In Question 9, participants were asked to guess the aim of the experiment. The answers are hard to compare as they address different levels of analysis, from quite abstract speculation (e.g., “Link between eye movement and mental analysis of the images on screen? Perhaps whether it helps when describing something to somebody to be able to view the picture ‘through their eyes’” (SP3); “how your eyes recognise things before your brain does” (SP4)), to fairly precise predictions (e.g., ‘if ‘looking’ (with the green dot) or describing the objects identifies them quicker” (RM5); “whether I trusted the dot entirely or ignored it and tried to locate the objects with the description or whether I used a mixture of the two” (SP5)).

Any further comments (Question 10) related mainly to experimental procedure. The most interesting responses are probably the ones made to Questions 11 and 12. Here, participants were told that there had been two conditions, ‘speaking’ and ‘remembering’, but that they had not necessarily been assigned to the one they were told about originally. They were then invited to guess which condition they had been in, and to provide a brief explanation for their guess. Neither condition
seems to have been obviously evident to participants: in total, seven participants across instruction conditions believed that they had seen a ‘speaking’ cursor (SP: \( N = 4 \), RM: \( N = 3 \)), while five thought they had seen a ‘remembering’ cursor (SP: \( N = 2 \), RM: \( N = 3 \)). Another way of phrasing this result is that 2/6 did not believe the instruction they were given in the ‘speaking’ condition vs. 3/6 in ‘remembering’. Evidently, both instructions could be either believed or disbelieved, without a clear preference; and neither instruction was particularly obvious.

This is supported by a rough categorisation of the reasons participants gave for their guesses; the same reasons were sometimes used to argue for ‘remembering’, at other times for ‘speaking’. For example, lack of accuracy of the cursor relative to the object could be seen as an indication of ‘remembering’ (e.g., “The dot did not always match exactly with what the speaker was describing - it often moved around the scene, not directly to the next numbered object” (RM3; guessed remembering)). However, to other participants the same lack of accuracy suggested that the cursor was recorded during ‘speaking’:

“When describing I tend to look at around the object to describe it and its location relative to other things or points of reference, e.g. ‘next to’...or ‘left or right of’... When thinking about an object I know that it is already there and don't have to describe it or its location.” (RM5; guessed speaking)

Similarly, too much (rather than a lack of) accuracy could also suggest both recording conditions. Some other arguments mentioned for one or both conditions were task-specific scanning patterns, the fact that the cursor preceded speech, and memory capacity limits.

To summarise, the ‘guess’ question proves that the instruction manipulation in Experiment 3 was not successful: a large proportion of participants were not convinced by the account they were given regarding the cursor, even if it was actually correct. Those who did believe what they were told did so for different reasons, and so presumably also drew different conclusions regarding the usefulness of the cursor. While it is likely that the number of doubters was enhanced by the direct question and its implication that some participants had been ‘tricked’, it is also clear that a more direct manipulation would be required to investigate effects of participants’ confidence in the predictive validity of the gaze cursor.
5.4 Conclusions

Experiment 3 found no effect of what participants were told the gaze cursor represented, either on first fixations or on click latencies. In both instruction conditions, the presence of the cursor had a similar facilitatory effect to the one shown in Experiment 2. The obvious conclusion is that participants use the cursor as a convenient visual cue without considering its meaning; and specifically without appreciating its connection to speaker gaze.

However, this conclusion must be qualified by the fact that the instruction manipulation itself was not successful and would therefore have been unable to uncover potential effects of listeners’ beliefs about the cursor. While setting up the experiment, we initially discussed several alternative instructions (e.g., the cursor could represent gaze recorded from someone randomly looking at the picture; or not mentioning the cursor at all). They were rejected because none of them seemed plausible. However, the instruction which was selected in the end actually required an extra cognitive step: participants needed to realise that gaze in remembering might be unhelpful, as the gazer would not necessarily have remembered all the objects correctly. Only if this link was obvious to them could it have affected their confidence in the gaze cursor, which in turn would be necessary to influence how it was processed. Therefore, while participants did receive some practice at the task of remembering, retrospectively it was optimistic to assume that they would reflect on the meaning of the green dot to this extent.

One alternative explanation for the lack of difference between conditions on the online measures should also be considered: it is also possible that the instructions were different enough, but that both had the same effect. This could be the case if eye movements made while scanning a picture in order to describe it (‘speaking’) and while retrieving information from it (‘remembering’) are actually quite similar, as has been proposed recently (e.g., Kent & Lambert, 2008). However, for such an account to explain the lack of effects in this study would require listeners to be aware of such a similarity. This seems a lot less likely than that they just followed the gaze cursor, irrespective of what they were told it represented.

Nonetheless, there are some indications that participants were capable of reflecting on the meaning of the gaze cursor, even if they didn’t necessarily do so...
during the experiment. None of the rating scales in the questionnaire showed significant differences between instructions (note that ratings were made before participants were supplied with a reason to doubt what they had been told in the instruction). However, the trends were always consistent with the predictions: subjectively at least, participants were slightly less confident in the ‘remembering’ condition. It is possible that with many more participants such differences would be reliable; in this experiment, comparisons were made between only six pairs of listeners to six very different speakers. However, reliable differences in confidence ratings would seem to be a pre-requisite for finding any effects on the online measures of processing.

Importantly therefore, interpretations linking the processing of the cursor to face-to-face gaze must remain tenuous at least. It seems more likely that the cursor is processed primarily as a prominent visual cue which inherently attracts attention. However, it is precisely these attention-grabbing attributes that also ensure that the listener’s attention is drawn to what the speaker is focussing on, generating joint attention between the two, at least in the sense of “looking where someone else is looking” (Butterworth, 1995, p. 29). Consequently, presenting an on-screen gaze cursor may be a tool to dramatically reduce the usual time lag between the speaker’s and listener’s attention (though it can’t completely eliminate it, since both still need time to program the fixations).

This idea forms the basis of Experiments 6 and 7, which explore the effect of varying the timing between gaze cursor and speech. On the one hand, listeners show a strong inclination to follow the gaze cursor to the objects of the speaker’s attention. A gaze cursor that is slightly ahead of the speaker’s actual gaze should therefore prove helpful in predicting upcoming speech content. This is particularly interesting in light of recent suggestions that comprehenders actually use the production system for predicting upcoming speech (Pickering & Garrod, 2007). On the other hand, a gaze cursor which was massively out of sync with speech would provide a much stronger manipulation of its overall usefulness. If listeners are able to disengage attention from the cursor, surely this should happen when the cursor is delayed and therefore no longer helpful in interpreting the speech utterance.
Chapter 6: Methodological issues – Experiment 4

Experiment 4, described in the following, was conducted immediately after Experiment 3, the subject of the previous chapter. However, its analysis revealed a number of methodological issues and problems with the procedure, some of which have already been alluded to. I am therefore including it in this thesis not so much as an empirical study in its own right, but for the purpose of discussing some design aspects of the gaze projection experiments.

For this reason, while the motivation for the experiment and the method used will be presented in full in Sections 6.1 and 6.2, respectively, the results will be summarised only briefly (Section 6.3). Instead, they will be followed by an extended discussion of problems with the previous design (6.4) and suggestions for modifying the paradigm (6.5). I will return to theoretical questions in Chapter 7.

6.1 Research questions and predictions

Two findings from the previous studies provided the motivation for Experiment 4. For one, Experiment 3 found no evidence that listeners’ assumptions about the gaze cursor had any effect on the way they processed it. However, it also seemed that the critical instruction was ineffective. A much stronger manipulation would be to vary the extent to which the gaze cursor really is informative about upcoming speech. The idea behind Experiment 4 was therefore to examine the effect of altering the lag between the gaze cursor and the speech utterance. In addition, neither of the previous experiments found any change over time in the way listeners used the gaze cursor. It was therefore deemed safe to vary the timing lag within listeners, such that each participant took part in one fast, one delayed, and one unmanipulated cursor condition.
One of the key features of the gaze projection paradigm is that it provides listeners with spatially and temporally accurate information about the speaker’s gaze location – more accurate than information they could extract themselves by looking at the speaker’s face. An initial motivation for adopting this paradigm was the idea that by varying whether the gaze cursor was played back in parallel or offset in time against the speech stream, it would be possible to investigate whether listeners understood the precise timing of production-related gazes, and whether they were able to use this information to predict upcoming sentence content or structure.

However, it now seems doubtful that the gaze projection paradigm is particularly suited to answering such a question. Mainly, this would require that the cursor really is interpreted as speaker gaze – an assumption which was not particularly well supported by Experiments 2 and 3. If the cursor is nothing more than a salient visual cue which automatically attracts attention, then changing its timing would not be very informative about listeners’ ability to interpret the timing information contained in speakers’ gazes.

In addition, the way the gazes were recorded created a confound: because the speaker’s first fixation usually went to the little number, but the listener saw this as a fixation to the object, the timing of this first speaker fixation was not necessarily associated with the production of the object’s name. The timing link was further weakened by the fact that speakers also looked at future objects while they were still describing the current ones (though these gazes did of course contain useful information about speech content).

For these reasons, the purpose of varying the time lag between speech and cursor in Experiment 4 was not primarily to investigate listeners’ intuitions about speakers’ gaze timing. Instead, varying the timing was employed as a means of making the cursor more or less informative about speech content. Manipulating the amount of overlap between the two forms of information in this way changes the relative importance of speech versus gaze cursor for the task of identifying and clicking on the mentioned objects. Generally speaking, if the cursor precedes speech content, listeners should rely more on this visual information for identifying the objects. In contrast, the more the gaze cursor is delayed behind the natural fixation, the less helpful it will be, until it actually becomes redundant behind the speech.
These predictions also contain a test of the extent to which the cursor automatically attracts attention: if listeners are forced to focus on the bright moving dot their fixation patterns should follow the gaze cursor exactly, irrespective of how useful it is. On the other hand, if they are able to disengage attention from it, the extent to which they follow it should be based on its usefulness.

6.2 Methods

6.2.1 Design

The six speaker recordings from Experiment 2 were re-used for Experiment 4, after modifying the gaze cursor files. For every speaker description, one file was created with the cursor shifted one second ahead of the original gaze, and another with the cursor delayed one second behind it. Together with the ‘natural’ baseline, this resulted in three speech-cursor timing conditions, which were varied within participants. In this way, each person heard and saw three descriptions by the same speaker, but in one the cursor was shifted ahead, in one it was delayed, and in one it reflected the natural lag. The order of rooms and cursor conditions was counterbalanced.

6.2.2 Materials

Shifted gaze cursors were based on the original movement files described in Chapter 4, Section 4.3.2.2. To generate a gaze cursor shifted ahead by one second, all x/y-coordinate values for time points before 1000 ms were removed from the start of the file and re-inserted at the end. In this way, the cursor first appeared at the position it would originally have been at one second further into the trial. The only time when this was not the case was the very last second of each description, when the fixations shown had originally occurred right at the beginning of the trial. However, because the speaker fixated the centre of the screen both at the beginning and at the end of the trial, this was hardly noticeable. Also, by this time the
description of the last object had already finished, so listeners would usually have made their last click and be waiting for the final fixation point.

‘Delayed’ gaze cursors were created by shifting the x/y-coordinates later in time, so that the values which used to be at Time Point 0 now occurred at Time Point 1000. For all earlier time points, coordinates outside the range of the screen were inserted, which meant that no cursor at all was shown during the first second of the trial (generally before speech onset). The final second of the movement file, where the cursor was always at the centre of the screen, was removed.

### 6.2.3 Participants

Eighteen native British students from Edinburgh University took part in Experiment 4, i.e. three listeners for each of the six speakers. Their mean age was 22.4; five were male. None had taken part in any previous gaze projection experiment. In total, the experiment lasted about 45 minutes, for which participants were paid £5.

### 6.2.4 Procedure

Participants were eyetracked at 500 Hz using EyeLink 2, again in the JEL. Following the setup procedure described for Experiments 2 and 3, all participants received the same instructions with regard to the gaze cursor:

Your task will be to mouse-click on the twelve objects as soon as you are sure which object the speaker is referring to. At the same time, a little green dot on the screen will show where the speaker was looking.

For the practice trial, a female speaker was recorded describing six objects in the classroom image that had been used in the previous studies. Her description (but no gaze cursor) was played back to participants, who were asked to click on the objects. This made the practice trial more useful than in the previous studies for clarifying the procedure of the main experiment, without giving participants a preview of the gaze cursor.
Apart from these changes to the instructions and the practice trial, the procedure was the same as in Experiment 3. Following the three trials, participants were asked to fill in a questionnaire, which in most respects was identical to the one used previously (see Appendix E). However, one question explicitly described the manipulation of the gaze cursor and asked participants to guess which room they had seen in which condition.

In addition, the questionnaire included a recognition task intended as a pilot study for an experiment manipulating referring expressions. Participants received a list of all the names used for the items by any of the six speakers and were asked to circle the term they remembered hearing. To date, the results of the pilot study have not actually been used, so it will not be mentioned further; but I have included the questionnaire in Appendix F for reference.

### 6.3 Preliminary results

The analysis of Experiment 4 was initially conducted much as for the previous experiments: data from participants who had heard the same speaker was treated as repeated measurements. Fixations and clicks were categorised with respect to the regions of interest, after manual drift correction of fixations based on the final fixation point. As in Experiment 2, and in contrast to Experiment 3, a relevant gaze was defined as any gaze to an item which spanned the onset of its description or occurred while it was being described. According to this definition, 9.8% of items received no relevant fixation. Cursor condition did not affect the probability of missed fixations ($p = .32$).

Experiments 2 and 3 established that if a gaze cursor is present, participants tend to look at it. For this reason, the predictions regarding first fixation onsets in this study served mainly as a manipulation check: since the three cursor conditions differed by one second intervals, first fixation onsets were expected to show a similar stepwise decrease from the ‘delayed’ cursor over the ‘natural’ to the ‘ahead’ cursor condition. However, as Table 6-1 shows, this was not the case. In fact, the mean time till first fixation was remarkably similar in the three conditions, at about
two seconds from item onset. In addition, the large standard deviations indicate a lot of variance within conditions.

Table 6-1: First fixation onsets and click latencies for the three cursor conditions, in ms from item onset.

<table>
<thead>
<tr>
<th>Cursor condition</th>
<th>delayed</th>
<th>natural</th>
<th>ahead</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>First fixation onset</td>
<td>M</td>
<td>2225</td>
<td>1998</td>
<td>2091</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>2506</td>
<td>3028</td>
<td>3149</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>193</td>
<td>189</td>
<td>200</td>
</tr>
<tr>
<td>Click latency</td>
<td>M</td>
<td>5285</td>
<td>5145</td>
<td>5206</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>3508</td>
<td>3319</td>
<td>3336</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>215</td>
<td>215</td>
<td>214</td>
</tr>
</tbody>
</table>

As might be expected in such a case, a 2-way repeated-measures analysis of variance with the factors CURSOR (‘ahead’, ‘natural’, or ‘delayed’) and TRIAL (1-3) revealed no significant effect of either factor (all $F1/F2 < 1$). The interaction was closer to significance ($F1(4, 20) = 1.70, p = .19$; $F2(2.7, 83.6) = 2.58, p = .07$), but the pattern of results was not easily interpretable. The corresponding analysis of click latencies showed a similar lack of main effects (all $p > .3$). The interaction was significant by items ($F1(2.1, 10.4) = 1.03, p = .4$; $F2(3.0, 103.5) = 7.66, p < .001$; both Greenhouse-Geisser corrected due to lack of sphericity), but without an obvious pattern.

6.4 Problems

It seems clear that something about the design of this experiment was problematic. In particular, the fact that the first fixation onsets in no way reflected the substantial changes in the timing of the cursor seems improbable. An intensive problem-solving process showed that the cursor timing and audio synchronisation had run as intended, that all raw fixation times were measured from trial onset (i.e. the appearance of the photo), that the item onsets had been correctly imported for all six speakers, and that the SPSS-script had identified the correct first relevant gazes. This suggests that the absence of any clear effect and the huge variances were
due not to a flaw in experimental procedure, but to the design itself. In the following sections I will point out a number of sources of variance and other factors that may have overpowered any effects of the cursor condition.

### 6.4.1 Speakers’ differences in descriptiveness

As described in Chapter 4, Section 4.3.3.1, the length and detail of speaker utterances varied widely. Yet detailed descriptions (e.g., Speakers 1 and 2) differ from more concise object naming in several ways: each trial was longer, as well as the time between item onset and unambiguous referring expression; and more neighbouring objects were fixated in an effort to produce helpful descriptions (e.g., “the kettle next to the stove and in front of the books”). Also, these speakers generally provided a higher level of detail, so it was often possible to identify the item before its unambiguous mention. Consequently, averaging over speakers added a lot of noise – even the earliest first fixation times for one speaker might be relatively long for another. Figure 6-1 compares the onset times of first relevant gazes between speakers. It confirms that averaging was problematic, particularly for Speaker 1.

![Figure 6-1: Median first fixation onsets for each cursor condition and speaker, in ms from item onset. Medians were used as measures of central tendency which don’t require excluding extreme values.](image-url)
6.4.2 Differences in the accuracy of speakers’ eye movement records

Another way in which speakers differed was in the accuracy of their eye movement recordings, and therefore of the gaze cursors based on these. While listener eye movements were manually drift-corrected before analysis, this was not possible for speakers, because the raw x/y-coordinates were required to create the gaze cursor. Despite only including speakers whose gaze trace seemed reasonably accurate, the active nature of their description task meant that the resulting cursors were certainly not perfect. This may have been particularly problematic when they were moved in time relative to speech: it is possible that a spatially accurate cursor is still useful if it is shifted in time, but that one which is projected to the upper right of the object as well as being lagged (i.e. both spatially and temporally inaccurate) could make it hard to establish a connection with the correct object. It therefore seems likely that some of the discrepancies in Figure 6-1 were caused by differences in the quality of the speaker recordings and resultingly, how listeners treated the cursors.

6.4.3 Low listener-speaker ratio

Both of the problems discussed so far were exacerbated by using six different speakers, who differed on these and other parameters. Although a fairly respectable $N = 18$ listeners were tested, this still resulted in only three listeners per speaker. Also, the matched-pairs $F1$-analyses were in effect run with an $N$ of 6 speakers, which was probably not sufficient to detect small effects. On the other hand, the expected pattern of results regarding first fixation onsets would have been a step of roughly 1000 ms between each of the three cursor conditions. This should have been visible even with just one listener – but it wasn’t, as Figure 6-1 shows.

\[ \text{A more traditional by-subjects analysis between listeners rather than between speakers (} N = 18 \text{) showed no effect either - not surprisingly, considering the lack of an obvious difference between the means per condition and in view of the variance due to the differences between speakers’ descriptions.} \]
Originally, the reason for using several speakers was to ensure generalisability, instead of just picking a ‘good’ speaker. I believe that this is a valid concern. However, it’s not one that is generally addressed in spoken comprehension experiments, which tend to use a single speaker (e.g., Swinney, 1979). Conventionally, such a speaker is also fully scripted, allowing control over what he/she says. In dialogue paradigms using a confederate, the same person normally plays this role for all naïve participants (e.g., Branigan et al., 2000). In these cases, the confederate is trained and scripted to perform as similarly as possible on each occasion. Lockridge and Brennan (2002) question the assumptions that even a well-trained confederate can treat all participants identically and that this would be similar to how a naïve person would interact with them. Their experiment involved two naïve partners telling each other stories about line drawings which were visible either only to the speaker or to both of them. With regard to the spontaneity and idiosyncrasy of the resulting speech utterances, their paradigm resembles the one reported here. However, their main motivation for using naïve pairs of participants was so that listeners would provide realistic and informative feedback, which is not possible in our setup.

This makes it particularly desirable to test the generalisability of our results across different speakers. Initially though, it is probably more important to use a controlled design to ascertain which factors actually affect the processing of the gaze cursor. Since the speakers cannot be scripted without affecting the naturalness of their production-related eye movements (which are reproduced in the gaze cursor), one way to increase control in the design would be to use only one speaker. Once the relevant factors have been established, a subsequent question could be whether they affect the use of the cursor differently for different people.

### 6.4.4 Order of trial presentation

The main difference between Experiment 4 and the previous studies was the variation of the cursor condition within, rather than between participants. The order in which a particular trial-cursor combination occurred was counterbalanced across participants, but full counterbalancing of all trial orders, cursor conditions, and rooms would have required 36 participants per speaker. In effect, it seems that the
order in which the different cursor conditions were experienced affected how participants made use of the information contained in the cursor. For instance, it could be that if they saw and used a ‘natural’ cursor in the first trial, but the second trial presented a ‘delayed’ one, then by the third trial they wouldn’t know what to make of the ‘ahead’ cursor. It’s hard to predict exactly which order of conditions would affect participants in which way, but Figure 6-2 clearly suggests that not everyone drew the same conclusions. Collapsing across speakers, it displays first fixation onsets for the first, second, and third trial, depending on the order in which the conditions were experienced. Clearly, first fixation onsets changed over the course of the experiment, but this change seems to have no clear direction.25

![Figure 6-2: Median first fixation onsets depending on the order in which each cursor condition was experienced, in ms from item onset. (D = ‘delayed’, N = ‘natural’, A = ‘ahead’). As in Figure 6-1, medians were used as a rough measure of central tendency and error bars are not shown for reasons of clarity.](image)

### 6.4.5 **Was the lag manipulation too short?**

In most psycholinguistic and visual studies, shifting the time course between conditions by one second would be a comparatively large manipulation. This lag

---

25 On the other hand, an argument against a strong influence of trial order is the fact that even the very first trial (where participants didn’t yet have any experience with the cursor) did not show the expected pattern.
was selected based on the consensus in the language production literature that speakers look at objects roughly one second before referring to them (cf. Griffin & Bock, 2000), suggesting that this might be a theoretically interesting lag to investigate. Surprisingly however, the cursor manipulations proved quite hard to detect, even if you knew about them. In retrospect, this is presumably due to the length of speakers’ descriptions: in a period of ten seconds with frequent refixations of the object they were describing, one second more or less did not make much difference.

In fact, none of the participants noticed the changes spontaneously. Even when they were pushed to guess which condition was which in the questionnaire, they were not good at it: only 2/18 participants guessed all conditions correctly (out of six possible guesses); 10/18 (56 %) guessed one out of three. Of course, this may be because these were retrospective judgements after the main experiment and they just didn’t remember, but certainly it demonstrates that the shifts were not very obvious. It does seem possible that – particularly in view of the long speaker descriptions – the one-second shifts between conditions were just too short to create a noticeable discrepancy between speech and cursor information.

6.4.6 Inaccuracies of listener eye movements

Finally, the number of missing relevant fixations after drift correction (~10 %, see Section 6.3) suggests that listener recordings also suffered from drift. Though this does not seem to have been a major problem, it would have led to more missing values for some participants than others, as well as increasing the variance if manual drift correction led to occasional misidentifications of the first relevant fixation.
6.5 Conclusions and potential paradigm modifications

Experiment 4 found no difference between cursor conditions on any measures. As explained in the previous section, this may well have been due at least in part to the following factors:

- differences between speakers, both in the length and detail of descriptions and in the quality of the gaze cursor;
- an ambitious design which made full counterbalancing impractical;
- varying timing within listeners, the order of which seems to have affected how the different conditions were processed;
- a relatively small manipulation of the speech-cursor lag;
- effects of drift on listener eye movements across long trials.

This list suggests a number of ways to modify the paradigm and hopefully increase its power to detect the effects we are interested in. These include:

- providing speakers with clearer instructions regarding descriptiveness;
- reducing trial length by creating more photos with fewer items;
- recording speakers using a different eyetracker, or attempting to manually drift-correct the movement files before creating the gaze cursor;
- or alternatively using just one representative speaker;
- varying conditions only between, not within listeners;
- testing larger cursor lags;
- recording listeners’ eye movements with a different tracker.

Some of these suggestions were implemented in the studies described in the remainder of this thesis. For instance, all subsequent experiments varied conditions between listeners. They were conducted using EyeLink 1000, an eyetracker which is much less affected by drift than EyeLink 2 (J. Shen, SR Research, personal communication, September 24, 2007). However, it does also restrict participants’ head movements, which is one reason we did not record any new speakers. Instead, a single speaker was selected from the previous six, as will be described in Chapter
7. Finally, Chapter 8 will return to the question of lag differences between speech and gaze cursor and test a whole scale of these.

Despite all the problems, Experiment 4 produced one encouraging and important finding, which is worth exploring further. In contrast to what was suggested by Experiments 2 and 3, this data demonstrates in several ways that participants do seem to regulate how they treat the gaze cursor. While the effect of TRIAL was far from significant, there was some indication that experiencing the three cursor conditions in different orders resulted in differential responses from participants. In addition, and more obviously, if participants had no choice but to look at the green moving dot on their screen, the pattern of first fixation onsets should have reflected a clear and unambiguous series of one-second steps between conditions. The very fact that it seems to have been influenced by so many different aspects of the experimental design is evidence against strictly automatic processing of the cursor, and implies at least some top-down control over how it is used.

To summarise, Experiment 4 was originally devised to investigate how shifting the lag between the speech utterance and the gaze cursor would affect listeners’ fixations, comprehension, and potentially even their recall of referring expressions. However, any potential results were swamped by a huge amount of variance. A discussion of possible reasons for this outcome changed the motivation for including Experiment 4 in this thesis: rather than answering a question in its own right, it provided the backdrop for some general considerations regarding the paradigm. Following this departure from the main thrust of the argument, the remaining chapters return to questions of theoretical interest: what exactly makes the gaze cursor helpful for listeners, and how important is it that it is tightly linked to the speech utterance?
Chapter 7: Replication and Extension – Experiment 5

Based on the experiences with Experiment 4, the gaze projection paradigm was modified in several important ways, which will be reported in this chapter. Consequently, the main practical purpose of Experiment 5 was to replicate previous findings with the revised version of the paradigm. To this end, it re-examined the general question of how the presence of the gaze cursor influences the processing of the speech utterance. At the same time, it also expanded on Experiment 2 by including an additional condition, where the cursor was presented without any speech at all. In this sense, Experiment 5 provides a complete comparison of the individual and combined effects of using the gaze cursor and/or hearing the description.

7.1 Research question and predictions

The theoretical interest behind Experiment 5 is how the speech utterance and gaze cursor individually affect fixations and performance on the comprehension task, as well as whether and how they interact. Three conditions were compared for this purpose: a speech-only version, a cursor-only version, and a combined speech & cursor condition. The latter corresponded to the ‘natural’ baseline in the previous experiments.

Despite the implication from Experiment 4 that the gaze cursor can be ignored, it very obviously does attract a lot of attention. It was therefore predicted that, as in Experiment 2, first fixations to relevant objects would be much earlier with a cursor than without, but that there would not be much difference between the two conditions with cursors. The interesting question was whether this pattern would carry through to the click responses: would participants’ click latencies in the combined speech & cursor condition be affected primarily by the description they
heard, as suggested by the fact that most clicks occur only after unambiguous
naming? If language is critical to the click task, participants in the cursor-only
condition should show severely impaired performance relative to the two ‘speech’
conditions. On the other hand, if the cursor is so unambiguous that it can to some
extent replace the information provided by speech, then cursor-only click latencies
should be more similar to the combined speech & cursor condition.

7.2 Methods

7.2.1 Design

In Experiment 5, three presentation conditions were varied between
participants: speech-only, cursor-only, speech & cursor. All speech utterances and gaze
cursors were based on recordings from the same speaker, but the order in which
participants saw the three rooms was counterbalanced across listeners and
conditions.

7.2.2 Materials: Selection of speaker recordings

The gaze cursor used in Experiment 5 was based on one speaker’s eye
movements and the corresponding speech file. Because of the accuracy problems
with some of the gaze cursors used previously, it was essential to select a speaker
whose recordings were as accurate as possible, while using reasonably objective
criteria to decide which speaker was ‘good’. This was achieved by comparing the
audio and movement files for all six speakers from Experiment 1 on a number of
parameters, and selecting the ‘best’ overall.

With regard to gaze recording accuracy, at least one relevant fixation was
recorded for each of the 36 items described by Speakers 2 and 4. This would
guarantee that a gaze cursor based on either of these speakers would focus on each
item while it was being described. However, Speaker 2 produced the longest and
most detailed descriptions of all speakers (mean item duration: 16 s). In contrast,
Speaker 4’s descriptions seemed more typical for the task: his mean description duration for each item \( (M = 9.9 \text{ s}) \) was closest to the grand mean of all speakers \( (M = 10.4 \text{ s}) \), and did not differ significantly from the mean durations of Speakers 3 and 5. His total trial duration also showed a comparatively small standard deviation \( (SD = 3.6) \), and mean onset of the unambiguous referring expression was early \( (M = 2.1 \text{ s}) \), but again did not differ significantly from Speakers 3, 5, and 6.

For these reasons, Speaker 4’s item descriptions and the corresponding gaze cursor were selected as stimuli for Experiment 5. This speaker was a 21-year-old male student from Kent. The gaze cursor was based on the recording of his left eye. An example of his speech style can be found in Appendix D, which contains a complete transcript of his description of the kitchen scene.

### 7.2.3 Participants

Eighteen undergraduate students or recent graduates of Edinburgh University took part in Experiment 5 (2 datasets had to be replaced due to calibration issues). They were native speakers of British English, eight were female, and their mean age was 21.1. All participants confirmed that they had normal or corrected-to-normal vision on both eyes. None had participated in any of the previous studies.

### 7.2.4 Apparatus

The experiment was conducted in a soundproof chamber in the Psycholinguistics and Visual Cognition (PVC) lab in the Edinburgh Psychology department, on a computer running OS Windows XP®. Stimulus presentation and response recording were controlled by Experiment Builder (SR Research, Canada). Participants were seated on an adjustable chair, 90 cm away from a 21’ computer screen (resolution 1024 × 768 pixel, refresh rate 140 Hz). Having signed the consent form, they were fitted with a tower-mounted EyeLink 1000 eyetracker (SR Research). This tracker samples the eye position every millisecond with a resolution of .01° visual angle. A 9-point calibration procedure ensured that average error was < .5° and maximum error < 1°. Full calibration was repeated before each trial. Due
to the nature of the tower-mounted eyetracker, tracking was always monocular to the right eye, though viewing was binocular.

### 7.2.5 Procedure

Instructions were presented on-screen, and participants pressed a key to confirm that they had read and understood them. The experiment began with a practice trial, which was the same as the one in Experiment 4 (Section 6.2.4). Up to the end of the practice trial, instructions were identical for all conditions. This meant that although participants in the cursor-only condition wouldn’t actually hear any speech during the main experiment, they experienced the same speech-only practice as the other participants. This ensured that they had some idea of the level of description the speaker had originally used.

Following the practice trial, the instructions differed depending on the condition the participant was assigned to, indicated by the bullet points and capital letters in the following. All participants received the same non-bulleted text:

In the following image, the speaker described 12 objects.

- You will hear the speaker’s descriptions and you will also see a moving green dot on the screen, showing where they were looking while they spoke. (SPEECH & CURSOR)

- You won’t actually hear the speaker’s descriptions, but you will see a moving green dot on the screen, showing where they were looking while they spoke. (CURSOR-ONLY)

- You will hear the speaker’s descriptions, but you won’t see where they were looking while they spoke. (SPEECH-ONLY)

Please mouse-click on the 12 objects as soon as you think you know which object the speaker is referring to.

At the end of the trial, please look at the fixation dot in the centre until the screen changes.

These instructions were displayed for at least 25 seconds before the start of the main experiment, and participants were given the opportunity to ask for clarification. If necessary, cursor-only participants were told that they would have to guess what the speaker was talking about. The instructions reappeared on-screen.
before each of the three trials. Each trial ended with the presentation of a final central fixation target, which was overlaid on the image for three seconds. This allowed us to verify the insensitivity of the EyeLink 1000 tracker to drift over the course of the three trials, each of which lasted on average 113 s.

The post-experimental questionnaire also differed depending on the experimental condition. For instance, cursor-only participants were asked to rate the usefulness of the green dot, speech-only participants judged the usefulness of the description, and those in the combined condition rated both. All differences are indicated in the questionnaire in Appendix E. In total, the experiment lasted about 35 minutes, for which participants received £4.

7.3 Results

The analysis of Experiment 5 proceeded much as described for the previous experiments. To begin with, a comparison of first fixation onset times revealed how consistently listeners followed the cursor in the different conditions, and whether the previous findings generalised to the new version of the paradigm (Section 7.3.1). More interestingly, the comprehension-based measure of click-latencies (7.3.2) allowed me to investigate the relative importance of speech utterance and visual information. Participants’ responses to the questionnaire will be briefly discussed in Section 7.3.3, which is followed by an overview of the main results (7.4.1), a comparison between the original and the revised version of the paradigm (7.4.2), and a final discussion (7.5).

7.3.1 First fixation onsets

In contrast to the previous experiments, a comparison of gazes to the trial-final fixation point confirmed that the EyeLink 1000 recordings made it unnecessary to manually drift-correct fixations. Consequently, the fixation record was analysed with no further adjustment. As before, fixations were automatically categorised with respect to the ROI they fell in (see Chapter 4, Section 4.3.2.2 and Appendix B).
The boxplots in Figure 7-1 visualise the distribution of first relevant gaze onsets for each condition, according to the wide definition of relevant gazes adopted in Experiment 3 (see Chapter 5, Section 5.2.5). This definition incorporates any gaze to the current object occurring during the description of the previous, current, and subsequent items and ensured that early anticipatory gazes would be included in the analysis.

However, it is obvious from Figure 7-1 that – as in Experiment 2 – first relevant gaze onsets by participants who saw no cursor at all showed much greater variance in onset times than those in the two conditions with a cursor (speech-only: SD = 5605; cursor-only: SD = 2433, speech & cursor: SD = 2404; Levene’s F(2, 629) = 61.50, p < .001).

This is presumably due to the fact that these participants look around the image more, trying to guess the next item. Thus, in all conditions Figure 7-1 also contains something like anticipatory gazes before item onset, i.e. before the speaker has even begun to talk about that item, let alone mentioned it.

![Figure 7-1: Onset of the first relevant fixation according to the wide definition, in ms from item onset, depending on the type of stimulus. Each box contains 50 % of the data in that condition, with the midline representing the median. Outliers (more than 1.5 times the interquartile range outside the central box) and extreme values (more than 3 times the IQR) are represented as ○ and *, respectively.](image-url)

26 Levene’s F tests the hypothesis that variances are equal between groups (Field, 2005); a significant result therefore shows that the tested variances differ substantially.
For this reason, the ‘narrower’ definition of relevant gazes used in Experiment 2 was applied for this study: Only gazes to a specific ROI occurring during the description of the corresponding item (i.e. either spanning or beginning after the word “number”) were counted as relevant for the analysis. While this measure risks excluding some instances of true anticipation based on the gaze cursor, this would work against the hypothesis that participants in the cursor conditions will fixate relevant items earlier.

Using the narrow definition, no relevant gaze was recorded for 4.6 % of items across the entire experiment, but the likelihood did not differ significantly for the three conditions ($p = .43$). No recording included more than four missed items out of 36. As before, outliers and extreme values (+/- 2 SD from the mean per condition) were removed from the first relevant fixation data. The remaining 584 items constituted 90.1 % of the original dataset.

As Figure 7-2 suggests, a clear dissociation was found in first fixation onset latency between the conditions with and without a gaze cursor.

![Figure 7-2: Mean onset times of the first relevant gaze (narrow definition), in ms from item onset.](image)

Since only one speaker was used in the revised version of the paradigm, all analyses were run as conventional 2-way mixed measures ANOVAs, with the factors STIMULUS (between-subjects: speech-only, cursor-only, and speech & cursor) and TRIAL (within-subjects: 1-3). The analysis of first fixation onsets confirmed the main effect of STIMULUS which is apparent from Figure 7-2: $(F1(2, 15) = 354.45,$
$F_2(1.3, 36.1) = 158.76$ (Greenhouse-Geisser corrected due to lack of sphericity; both $p < .001$). Planned comparisons showed that participants who saw the gaze cursor fixated the relevant object much earlier ($\text{cursor-only}: M = 266$ ms from item onset, $SD = 850$; $\text{speech & cursor}: M = 307$ ms, $SD = 699$) than participants who didn’t ($M = 2965$ ms, $SD = 1595$; $p < .001$). For participants who saw the cursor, the onset of the first relevant fixation was not affected by whether they also heard speech ($p = .75$). There was no effect of TRIAL, and no significant interaction (all $p > .14$).

### 7.3.2 Click accuracy and latencies

The x/y-coordinates of participants’ mouse clicks occurring from the onset of each item were manually compared to the correct regions of interest. Significantly less correct clicks were made in the $\text{cursor-only}$ condition (185 out of 216; 85.7 %) than for $\text{speech-only}$ and $\text{speech & cursor}$ (both: 215/216; > 99 %): $\chi^2(2) = 57.47$, $p < .001$. At the same time, $\text{cursor-only}$ participants actually made slightly more clicks in total, and the numbers varied much more widely with regard to the number of clicks ($M = 37.2$ per participant, Range: 32-49) than for participants in the $\text{speech}$ conditions ($M = 35.9$, Range: 35-36), suggesting that participants were less sure of the correct referents.

This poses a problem for the analysis of click latencies. If $\text{cursor-only}$ participants adopted a strategy of “I’ll click on any object the green dot lands on”, without considering whether the speaker had actually been talking about it, both the location and the latency of their clicks would be based on a completely different premise than the two groups of $\text{speech}$ participants’. All correct clicks, i.e. any click to an object that was being described at the time, were therefore sub-classified further. The results are presented in Table 7-1, which also contains the number of wrong clicks and complete misses by condition.
Table 7-1: Subcategories of clicks by condition. The labels are explained in the text.

<table>
<thead>
<tr>
<th></th>
<th>single correct click</th>
<th>several clicks</th>
<th>wrong click</th>
<th>no click</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>speech-only</td>
<td>212</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>216</td>
</tr>
<tr>
<td></td>
<td>98.1 %</td>
<td>1.4 %</td>
<td>0.5 %</td>
<td>0 %</td>
<td>100 %</td>
</tr>
<tr>
<td>cursor-only</td>
<td>153</td>
<td>32</td>
<td>18</td>
<td>13</td>
<td>216</td>
</tr>
<tr>
<td></td>
<td>70.8 %</td>
<td>14.8 %</td>
<td>8.3 %</td>
<td>6 %</td>
<td>100 %</td>
</tr>
<tr>
<td>speech &amp; cursor</td>
<td>215</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>216</td>
</tr>
<tr>
<td></td>
<td>99.5 %</td>
<td>0 %</td>
<td>0.5 %</td>
<td>0 %</td>
<td>100 %</td>
</tr>
<tr>
<td>Total</td>
<td>580</td>
<td>35</td>
<td>20</td>
<td>13</td>
<td>648</td>
</tr>
<tr>
<td></td>
<td>89.5 %</td>
<td>5.4 %</td>
<td>3.1 %</td>
<td>2 %</td>
<td>100 %</td>
</tr>
</tbody>
</table>

A single correct click was usually preceded directly by a correct click to the previous item, though this label could also be assigned if the previous item was clearly misidentified. This would be the case if click \( n-2 \) had been a correct identification of item \( n-2 \), but click \( n-1 \) did not go to item \( n-1 \) (for instance, the large plant in the bedroom was sometimes clicked instead of the small cactus). Several clicks refers to correct clicks which occurred amongst clicks to other objects during the description of the same item, suggesting that participants were guessing. Wrong click denotes a clear misidentification, as described above. Finally, for no click a correct click or clear misidentification of item \( n-1 \) was directly followed by one to item \( n+1 \).

Clearly, cursor-only participants not only made significantly less single correct clicks than participants who could hear the speech utterance \( (\chi^2(1) = 120.28, p < .001) \), they also used a guessing strategy more frequently (several clicks: \( \chi^2(1) = 56.19, p < .001 \)), identified the wrong object more often (wrong click: \( \chi^2(1) = 29.82, p < .001 \)), and completely missed more objects (no click: \( \chi^2(1) = 26.53, p < .001 \)). On the other hand, while performance was clearly impaired for these participants, they still unambiguously identified the correct object more than 70 % of the time – not really so bad, considering they couldn’t hear the description at all. In addition, the comparatively low frequency of ‘several clicks’ (~15 %) indicates that cursor-only participants generally tried to be fairly certain of the object before clicking.
Nonetheless, to avoid including clicks which may have been the result of a guessing strategy, inferential analyses were run only on the latencies of single correct clicks. This criterion reduced the number of included clicks in the cursor-only condition to 70.8 % (vs. 98.1 % and 99.5 % for the speech-only and speech & cursor conditions, respectively). Figure 7-3 provides mean latencies for these unambiguously correct clicks, for each of the three trials and stimulus conditions.

![Figure 7-3: Mean latencies for unambiguously correct clicks depending on gaze cursor condition and trial number, in ms from item onset.](image)

A 2-way analysis of variance confirmed the effect of STIMULUS: $F(2, 15) = 22.26; F(1.6, 46.3) = 30.29$ (Greenhouse-Geisser corrected due to lack of sphericity), both $p < .001$. Bonferroni-corrected posthoc tests showed that participants who saw a gaze cursor not only fixated but also clicked on the correct object earlier (cursor-only: $M = 2712$ ms, $SD = 1712$; speech & cursor: $M = 3180$ ms, $SD = 1668$) than participants who didn’t ($M = 4898$ ms, $SD = 2121$; both $p = .001$). Again, whether they heard speech as well as seeing the cursor was irrelevant ($p = .40$).

In this task, there was also a main effect of TRIAL: $F(2, 30) = 5.21, p = .01; F(2, 58) = 8.59, p = .001$. Polynomial contrasts revealed that this effect was linear, in that it reflected a monotonic reduction in click time across trials ($F(1, 15) = 9.83,$

---

27 Additional analyses on latencies of all correct clicks (i.e. including those in the several clicks category) showed exactly the same pattern.
participants were slowest on the first trial ($M = 3972$ ms) and became increasingly faster at responding ($M = 3631$ ms and $M = 3457$ ms for trials 2 and 3, respectively). There was no interaction of STIMULUS and TRIAL by participants ($p = .20$), though it was marginal by items ($F(3.01, 87.4) = 2.31, p = .08$). Figure 7-3 shows that over the course of the experiment participants who saw a gaze cursor became faster at correctly identifying the objects, whereas speech-only participants hardly improved at all. At the same time, it also suggests that the improvement may have occurred earlier for participants in the speech & cursor condition (by Trial 2) than for those who saw experienced the cursor-only presentation (speed-up not until Trial 3). It seems probable that looking for a learning effect across three 12-item trials is too coarse a measure to capture the developing adaptation to the presence of the cursor; this would most likely require an item-by-item analysis.

### 7.3.3 Questionnaire results

In Experiment 3, the questionnaire data was important because it showed the extent to which participants had believed and taken into account the instruction manipulation. In contrast, the three conditions in Experiment 5 were blatantly different from each other, so the questionnaire was not required as a manipulation check. Instead, it provided supplementary information on the relative importance of the two forms of stimulus, and on participants’ intuitions about how they might interact. The questionnaire items with a rating scale are particularly suited to this purpose and will be discussed in the following. First, descriptive statistics for these items are presented in Table 7-2.
Table 7-2: Descriptive statistics for rating questions (all on a 7-point scale).

<table>
<thead>
<tr>
<th></th>
<th>usefulness of speech</th>
<th>clarity of speech</th>
<th>usefulness of cursor</th>
<th>clarity of cursor</th>
<th>How helpful would it have been to hear/see the speaker?</th>
<th>How tightly was the cursor linked to speech?</th>
</tr>
</thead>
<tbody>
<tr>
<td>speech-only (N = 6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>5.50</td>
<td>4.83</td>
<td>-</td>
<td>-</td>
<td>5.17</td>
<td>6.00</td>
</tr>
<tr>
<td>SD</td>
<td>1.05</td>
<td>1.33</td>
<td>-</td>
<td>-</td>
<td>2.56</td>
<td>0.89</td>
</tr>
<tr>
<td>Range</td>
<td>4-7</td>
<td>3-7</td>
<td>-</td>
<td>-</td>
<td>1-7</td>
<td>5-7</td>
</tr>
<tr>
<td>cursor-only (N = 6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>-</td>
<td>-</td>
<td>5.67</td>
<td>5.67</td>
<td>6.83</td>
<td>5.33</td>
</tr>
<tr>
<td>SD</td>
<td>-</td>
<td>-</td>
<td>1.03</td>
<td>1.21</td>
<td>0.41</td>
<td>0.82</td>
</tr>
<tr>
<td>Range</td>
<td>-</td>
<td>-</td>
<td>4-7</td>
<td>4-7</td>
<td>6-7</td>
<td>4-6</td>
</tr>
<tr>
<td>speech &amp; cursor (N = 6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>4.67</td>
<td>4.50</td>
<td>6.00</td>
<td>-</td>
<td>-</td>
<td>6.17</td>
</tr>
<tr>
<td>SD</td>
<td>1.86</td>
<td>1.38</td>
<td>0.89</td>
<td>-</td>
<td>-</td>
<td>0.75</td>
</tr>
<tr>
<td>Range</td>
<td>2-7</td>
<td>2-6</td>
<td>5-7</td>
<td>-</td>
<td>-</td>
<td>5-7</td>
</tr>
<tr>
<td>Mean across conditions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>5.08</td>
<td>4.67</td>
<td>5.83</td>
<td>5.67</td>
<td>6.00</td>
<td>5.83</td>
</tr>
<tr>
<td>SD</td>
<td>1.51</td>
<td>1.3</td>
<td>0.94</td>
<td>1.21</td>
<td>1.95</td>
<td>0.86</td>
</tr>
</tbody>
</table>

After an initial description of the study, the first set of questions addressed the usefulness and clarity of the speech (for speech-only participants) and of the gaze cursor (for cursor-only participants). Speech & cursor participants rated both kinds of stimuli, so their responses were used for comparison.

On average, speech-only participants found the descriptions more useful than speech & cursor participants by almost one scale point, though the difference did not reach significance ($t(10) = .96, p = .36$). If anything, this might be a hint that speech & cursor participants had somewhat less need for the speech information. The two groups did not differ in their judgement of the clarity of the speech utterance ($p = .68$) – not surprisingly, considering that they had heard exactly the same description. Similarly, no difference was found between cursor-only and speech & cursor ratings of the usefulness of the cursor ($p = .56$).

---

Note that participants in this condition didn’t actually hear any speech, although they were told that the gaze cursor reflected the speaker’s eye movements while he was describing.
The next question asked each of the groups of participants who had received only one form of stimulus was whether it would have been helpful to be see or hear the other one as well. Across conditions, participants indicated that this would have been helpful ($M = 6.0$), but for cursor-only participants this rating was almost at maximum ($M = 6.83$). This contrasts with the speech-only participants, among whom the variance is significantly wider ($\text{Levene's } F(1, 10) = 15.02, p < .01$), though the two group means did not differ significantly ($t(5.3) = -1.57, p = .17$, equal variances not assumed): some speech-only participants agreed that it would have been ‘extremely helpful’ to see where the speaker was looking, but others felt this would have been ‘not at all helpful’. It seems that while all cursor-only participants were certain that their task would have been easier if they had been able to hear the speaker, the usefulness of a gaze cursor was less intuitively obvious to the speech-only group.

Only the final question in Table 7-2 was answered by participants in all conditions, though in fact it was not precisely the same question. Participants in the two cursor conditions had to rate how tightly they felt the dot had been linked to what the speaker was saying, whereas participants in the speech-only condition were asked more generally for their intuitions about the strength of the link between eye movements and speech. In effect, their assumptions about this association did not differ in strength from speech & cursor participants’ consistently high ratings ($p = .73$). Cursor-only participants’ rating of the tightness of the link was marginally lower than the rating by the speech & cursor participants ($t(10) = 1.84, p = .096$).

To summarise, though the comparisons between groups did not reach significance – probably due in part to very low power – there seems to have been a tendency for cursor-only participants to find their rather unusual task harder than those who heard the speech stimulus as well. Although they agreed that the gaze cursor was useful, they seem to have trusted it somewhat less than the speech & cursor participants, and generally felt that some important information was lacking. The latter, by contrast, were convinced of the utility of the gaze cursor, to the extent that they may have relied somewhat less on the speech utterance than the speech-only participants. This final group showed yet another pattern: despite assuming that there would be a tight link between eye movements and speech, the potential worth of being provided with the extra information does not seem to have been evident to all of them.
7.4 Conclusions

Experiment 5 served two separate purposes. Its theoretical goal was to identify the individual and combined effects of hearing the description and seeing a projection of the speaker’s gaze. From a practical point of view, it also served to evaluate the effects of the changes made to the paradigm, and whether the effects found in previous experiments would generalise to studies using the revised version. The results of these two separate purposes will be summarised and analysed in Sections 7.4.1 and 7.4.2, respectively.

7.4.1 Contributions of speech and gaze cursor

Considering the results from several quite different dependent measures, it seems that if a gaze cursor is played back as it occurred with a description, the main benefit of hearing the speech utterance is in avoiding errors. With regard to the latencies both of first fixating the relevant object and of identifying and clicking on it, participants in the cursor-only condition were at least as fast as those who also heard speech; whereas participants with speech but no cursor were much slower. However, when accuracy was taken into account, the pattern reversed: cursor-only participants were considerably more likely to misidentify objects or not to find them at all, while both speech conditions performed at ceiling. This may in part be due to a speed-accuracy trade-off, but the fact that only unambiguously correct fixations were included in the analysis suggests that this is not the whole story: there is no reason to assume that cursor-only participants would correctly identify more objects if they were forced in some way to delay their clicks.

At the risk of over-interpreting the questionnaire data, which found no significant differences of means, the numeric trends suggest that while cursor-only participants were generally able to extract the information they required from the cursor, they may also have been somewhat less certain of their conclusions. By contrast, speech-only participants correctly seem to have been more confident in their judgements, but were not particularly aware of the fact that being able to see the speaker’s gaze might have improved their performance.
7.4.2 Comparison with previous experiments

Clearly, the most basic result of Experiment 2 was replicated despite the changes in experimental procedure: seeing an indication of the speaker’s gaze helps listeners to locate the mentioned objects sooner, as well as providing them with confirmation of their selection. But in view of the two key changes to the paradigm (only one speaker and a different eyetracker), it is also interesting to compare the exact time scales of the latencies. Experiments 2 and 5 both used the same narrow definition of relevant gazes, as well as employing almost identical conditions, so they are well-suited to such a comparison.

Mean first fixation onset in Experiment 5, beginning 307 ms from item onset for the speech & cursor condition and 2965 ms for speech-only, is much earlier than in Experiment 1 (M = 1093 ms and M = 3642 ms, respectively). Although the means for Speaker 4 in Experiment 2 are somewhat more similar (speech & cursor: M = 614 ms; speech-only: M = 3365 ms), these are based only on one participant per condition. This makes it hard to determine how much of the discrepancy in means is due to the different speakers in Experiment 2, vs. the different eyetracker. However, click latencies were measured independently of the tracker and show a similar divergence (speech & cursor: M = 5098 ms (Exp. 2) vs. M = 3180 ms (Exp. 5); speech-only: M = 6931 ms vs. M = 4898 ms), suggesting that the differences are due primarily to the aggregation over several speakers in Experiment 2. Interestingly, the difference between the two conditions is remarkably similar in the two experiments and for both measures (first fixations: M = 2549 ms (Exp. 2) vs. M = 2658 ms (Exp. 5); click latencies: M = 1833 ms vs. M = 1718 ms).

Overall, it would seem that while the absolute timing of these measures may be affected by the combination of an individual speaker’s utterance and gaze record, the relative effect of presenting the gaze cue remains fairly constant. This makes it possible to generalise across speakers and experiments. It is also an important finding with regard to manipulating the time lag between speech and cursor, which will be the topic of the following chapter.
7.5 Summary and Discussion

The problems encountered in Experiment 4, described in Chapter 6, led to some modifications of the paradigm. Consequently, a practical goal of Experiment 5 was to replicate previous findings before returning to the question of lag differences between gaze cursor and speech utterance, which was the original idea behind Experiment 4. In addition, Experiment 5 also addressed an important issue in its own right: by including a condition where participants did not hear the speaker’s utterance at all, this study explored the individual contributions of each of the two kinds of stimulus. It found that the cursor alone was completely sufficient to explain the speed advantage for first fixations and also click times, which had been found in all previous speech & cursor conditions. Hearing the speech utterance, by contrast, substantially increased the accuracy of participants’ responses, and to some extent also their confidence in these.

However, these findings assume that speech and cursor are in sync, or rather that the former follows the latter in the same way that an utterance would normally follow eye movements. In this sense, participants in speech & cursor conditions simultaneously receive very similar information from two different stimuli. Whether they actually use both or give priority to one over the other will depend on a number of factors, such as the ease of extracting the relevant information, its (perceived) accuracy, and whether one form is available earlier than the other. This last factor will be explored in Chapter 8, which employs the improved paradigm to re-examine the lag manipulation attempted in Experiment 4.
Chapter 8: Varying lag times - Experiments 6 & 7

This chapter comprises two experiments, both of which explored the effect of varying the overlap between speech utterance and gaze cursor. These two stimuli contain very similar information, and Experiment 5 confirmed that either is sufficient on its own to adequately perform the task of clicking on the mentioned objects. However, that study also provided an explanation for the fact that performance is best when both stimuli are available: while listeners identify the relevant objects earlier when their decisions are based on the cursor alone, they are more accurate when they follow the speech utterance.

The additional facilitation found when gaze cursor and speech utterance are presented simultaneously indicates that both forms of information are integrated whenever possible. This leads to the question of what happens if the two stimuli are not consistent with each other. Are listeners able to allocate attention flexibly to whichever form of information is more useful? If so, can an unhelpful stimulus be ignored completely, or does its presence decrease performance relative to conditions with just a unique stimulus?

Experiment 6 addressed both these questions. On the one hand, it compared listeners’ performance in a condition where speech and gaze cursor co-occurred naturally with conditions where the gaze cursor was dramatically shifted in time, thus making it relatively more or less informative than the speech utterance. On the other hand, it included a condition with a completely uninformative gaze cursor, which listeners would ideally have to ignore to perform their task.

Next, Experiment 7 explored the effect of different lags in more detail. Its aim was to determine when and how the preference for one stimulus over the other evolves, and whether this results in an impairment of task performance. The chapter concludes with combined analyses across both experiments to compare fixation onsets and click latencies along a scale of different lag times. Among other things, these analyses should finally reveal the extent to which the gaze cursor is fixated.
and processed automatically, and whether listeners can and do sometimes take into account that it represents speaker gaze. Lastly, the questionnaire-based data on participants’ subjective impressions of the two stimuli was also pooled and compared across both experiments.

8.1 Experiment 6: Large lag variations and a misleading cursor

Experiment 6 investigated the effect of seeing a gaze cursor that was inconsistent with the speech utterance. This could be either an inconsistency in time, such that the gaze cursor was not in sync with the speech (but still informative), or it could be total, in form of a complete lack of connection between the two stimuli.

8.1.1 Research question and predictions

Experiment 6 extended the earlier studies in several ways. For one, a gaze cursor that really is out of sync with speech is a much stronger manipulation of its trustworthiness than the instruction changes used in Experiment 3, because it doesn’t rely on listeners interpreting the dot as speaker gaze. Finding an interaction of the respective usefulness of the speech- and cursor-information therefore seemed more likely with this design.

Additionally, the lag of five seconds employed in this experiment is much longer than the one-second shifts attempted in Experiment 4, presumably raising the probability of finding an effect of the inconsistency. Finally, all the successful previous experiments involved cursor-speech pairings that reflected the way speech and gaze co-occurred naturally in the speaker’s production. These experiments established that the gaze cursor contributed to the speed of the click decision, and the speech utterance to its accuracy. Whether a relative reduction of their respective usefulness leads to a corresponding decline in speed or accuracy remains to be seen.

Generally, shifting the cursor ahead of its original position by five seconds means that it is even further ahead of the speech content than in normal speech. The relevant items should therefore be fixated at least five seconds earlier than in the
baseline condition. Click latencies should also be dramatically reduced, unless the inconsistency between the two stimuli causes participants to prioritise accuracy and wait for the speech utterance. In contrast, a severely delayed cursor will lose its speed advantage over the speech utterance. Therefore, first fixation onsets in this condition were predicted to be much delayed, and click latencies longer than in the baseline. Finally, if the cursor has no connection at all with where the speaker was looking and what he was saying, fixating it would improve neither speed nor accuracy of the decision. If participants are able to ignore the green dot completely, their fixation and click latencies should therefore be comparable to the speech-only condition in Experiment 4. Any further increase in the two latencies would imply that the unhelpful cursor was distracting attention away from the task.

**8.1.2 Methods**

**8.1.2.1 Design**

Each participant in Experiment 6 was assigned to one of four cursor conditions, relative to speech: in the condition hereafter referred to as natural, the cursor represented the speaker’s gaze as it had accompanied his speech in the original recording. The three other conditions consisted of a gaze cursor shifted five seconds ahead of where it had originally been (ahead); one shifted five seconds behind this point (delayed); or a dot which was not connected to the speech in any way (random). All cursor conditions were based on the same original speaker, and varied between participants. The order in which the three rooms were seen was counterbalanced.

Following the main experiment, participants were presented with an additional recall task, similar to the one following Experiment 2: they were asked to recall as many items as possible from each room. During this task, half the participants saw the natural gaze cursor and the other half the random dot; but apart from this the screen was blank. As the recall task was only tangential to the main question, it is described in Appendix G.
8.1.2.2 Materials

As in Experiment 5, all descriptions and movement files were taken from the original recordings of Speaker 4. The movement files were used to create the shifted gaze cursors, using the algorithm described in Section 6.2.2, above. A shift of five seconds was decided upon because this is approximately half an average item description. This lag therefore seemed extreme enough to be disturbing, without being entirely disconnected from the speech utterance: fixations would be directed at the correct object for roughly half the item duration, but would arrive there much later than in the non-shifted condition.

In creating the random cursor, on the one hand it was important that it was not linked to the speech utterance in any way. On the other hand, participants were told that the dot represented speaker gaze, so it could not be truly random. For instance, fixation durations and saccades had to be reasonably similar to a real viewing pattern, but not directed at relevant objects. For this purpose, a movement file from the same speaker was used, but while he was describing one of the other rooms. In addition, the x- and y-coordinates were reversed, and the whole resulting file was played backwards, resulting in a fairly realistic impression of fixation behaviour.

8.1.2.3 Participants

Twenty-nine members of the University of Edinburgh student community took part in the study. Five were excluded from the analyses, either because they turned out to be much older than the rest (N = 2), or because calibration was not sufficiently accurate (N = 3). The remaining 16 females and 8 males were all native British English speakers, with normal or corrected-to-normal eyesight and aged between 18 and 29 (M = 21.4). None had taken part in any previous gaze-projection experiment.

8.1.2.4 Procedure

Movements of the participants’ right eyes were recorded using the EyeLink 1000 tracker in the PVC lab, as described in Chapter 7. Full 9-point calibration was carried out at the start of the experiment and before each trial, ensuring an average
error < .5° and maximum error < 1°. The instructions with regard to the gaze cursor were identical for all participants:

In this experiment, three photographs of different rooms will be displayed on the screen, one room at a time. You will hear a speaker describing twelve objects in each room. Your task will be to mouse-click on the twelve objects as soon as you are sure which object the speaker is referring to.

At the same time, a little green dot will move around on the screen. This dot is based on the speaker's eye movements. You may look at this, but you don't have to – it's up to you.

The practice trial and main experiment were the same as in Experiment 5. They were followed by three trials of surprise recall, as described in Appendix G. Following the three recall trials, participants responded to the questionnaire, which in most respects was identical to the one used previously (changes are indicated in Appendix E). Again, one question directly addressed the cursor manipulation: participants were informed of the different conditions and asked to guess which one they had experienced. Subsequently, they were thoroughly debriefed and told whether their guess had been correct. They left the lab after about 45 minutes, for which they received £5.

**8.1.3 Results and Discussion**

Because of the drift-resistance of the EyeLink 1000, no post-experimental manual drift correction was required. Fixations and clicks were directly classified according to the regions of interest they fell in.

**8.1.3.1 First fixation onsets**

The wide definition of relevant gazes was used, i.e. including any fixation to the correct object which occurred during the description of the previous, current, or subsequent item. This definition was used because of the five-second differences between cursors: obviously, if participants look at the objects well before item onset in the ahead condition, but not in the natural and delayed cursor conditions, this is a real result which could be obscured by including only fixations beginning around
item onset. There were only 18 cases without such a relevant gaze (2.1% of all items for all participants). Outliers and extreme values (+/- 2 SD from the condition mean) were excluded from the fixation onset time analyses. This increased the total missing data to 8.1%. Figure 8-1 presents mean onset times of the first relevant fixation to the item, depending on the timing relationship between the speech utterance and the gaze cursor.

![Figure 8-1: Mean onset of the first relevant fixation, depending on gaze cursor condition and trial number, in ms from item onset.](image)

A 2-way mixed measures ANOVA with the factors CURSOR (four conditions, between subjects) and TRIAL confirmed the main effect of CURSOR: $F1(3, 20) = 77.05; F2(2.4, 68.5) = 39.67$, Greenhouse-Geisser corrected due to lack of sphericity; both $p < .001$. There was no effect of TRIAL and no interaction (all $F < 1$).

Planned comparisons showed that first fixations in the 5s ahead condition ($M = 3132$ ms before item onset, $SD = 2901$) were significantly earlier than in the

---

29 An alternative strategy would have been to allow fixations within the last five seconds before item onset in the 5s ahead condition only, though this was not considered at the time. However, the narrow definition of relevant gazes was used for the comparison with Experiment 6, and the values for the natural, delayed, and random cursor conditions appear in Table 8-1. In addition, the analyses of fixation-to-click times in Section 8.3.2 suggest that spurious early fixations did not influence the results dramatically.
natural condition \( (M = 62\) ms after item onset, \( SD = 1477; p < .001\)). Evidently, participants in both cases tended to follow the gaze cursor to the item before the speaker had even begun talking about it. This was significantly earlier \( (p < .001)\) than first fixations by participants in the 5s delay and random conditions, whose mean onset times were \( M = 1543\) ms \( (SD = 4523)\) and \( M = 1852\) ms \( (SD = 4202)\) after item onset, respectively. Post-hoc Bonferroni-corrected \( t\)-tests showed that these two conditions did not differ significantly from each other \( (p > .9)\).

Clearly, first fixations were strongly affected by the speech-cursor lag, such that earlier gaze cursors (relative to speech) led to earlier fixations, as predicted. Two aspects of these results are worth noting: the predictions outlined in Section 8.1.1 stated that if the salient gaze cursor grabbed participants’ visual attention automatically, first fixation means in the different lag conditions should be separated by (at least) five seconds. At first glance, this seems not to be the case (mean difference between 5s ahead and natural: 3194 ms; between natural and 5s delayed: 1481 ms). Tests of whether these differences are significantly less than five seconds and a discussion of what this would imply will be postponed until the combined analyses with Experiment 7 in Section 8.3.1.

Another interesting finding is the similarity between random and 5s delayed. The distinction between these two conditions is that in the former, the green dot is completely unconnected to the speech content, whereas in the latter it is delayed, but still correctly points to the items. The fact that first fixation onset times did not differ significantly between them suggests that five seconds constituted such a strong delay that the cursor became more or less redundant, leading participants to rely almost entirely on the speech utterance to identify the items – just like those in the random condition. On the other hand, this would also mean that the random participants were not substantially distracted by the unhelpful cursor; otherwise their first fixations to the items should have been even more delayed.

The obvious way to test this speculation is to compare both conditions to the speech-only condition of Experiment 5. This resembled the random condition in that the speech utterance was the only form of information available, but was more like 5s delayed in that there was no additional distraction. For this comparison, the narrow definition of relevant gaze was used, including only those fixations to the item which began just before or during its description. Table 8-1 provides an
overview of mean first fixation onset times using this definition, for all the conditions included in the analysis.

Table 8-1: Descriptive statistics for comparing the (relatively) disconnected conditions of Experiments 5 and 6. All means in ms from item onset, using the narrow definition of relevant gaze. Note that all participants heard the same speaker's descriptions.

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Disconnected cursors</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>natural, Exp. 5</td>
<td>natural, Exp. 6</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>307</td>
<td>409</td>
<td>2965</td>
</tr>
<tr>
<td>N</td>
<td>194</td>
<td>207</td>
<td>203</td>
</tr>
<tr>
<td>SD</td>
<td>699</td>
<td>889</td>
<td>1595</td>
</tr>
<tr>
<td></td>
<td>speech-only, Exp. 5</td>
<td>5s delay, Exp. 6</td>
<td>1437</td>
</tr>
<tr>
<td>Mean</td>
<td>2965</td>
<td>3344</td>
<td>3385</td>
</tr>
<tr>
<td>N</td>
<td>194</td>
<td>207</td>
<td>199</td>
</tr>
<tr>
<td>SD</td>
<td>699</td>
<td>889</td>
<td>1917</td>
</tr>
<tr>
<td></td>
<td>random, Exp. 6</td>
<td></td>
<td>193</td>
</tr>
<tr>
<td>Mean</td>
<td>3344</td>
<td></td>
<td>3385</td>
</tr>
<tr>
<td>N</td>
<td>194</td>
<td></td>
<td>199</td>
</tr>
<tr>
<td>SD</td>
<td>699</td>
<td></td>
<td>1917</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>2073</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>996</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1975</td>
</tr>
</tbody>
</table>

Collapsing over trial order, these first fixation onsets were compared across the five conditions using ANOVA: $F1(4, 25) = 141.92$, $F2(2.1, 74.5) = 130.67$, both $p < .001$ (Greenhouse-Geisser corrected due to lack of sphericity). Games-Howell-corrected post-hoc tests (adjusted for non-homogenous variance between conditions) confirmed that there was no difference between the baseline natural conditions in the two experiments ($p = .94$). More interestingly, the three other conditions did not differ significantly either (all $p > .27$), despite the slightly lower mean for the speech-only condition. Thus there is no evidence that the random cursor caused substantial disturbance, nor that the 5s delay gaze cursor in any way improved speech-based processing.

8.1.3.2 Click latencies

Arguably, click latencies are the more interesting dependent measure, as they should reflect which kinds of inconsistency between speech and gaze cursor lead participants to ignore the cursor information in favour of the more accurate speech utterance. All analyses were based on correct clicks, which were made for 99.8% of items. Among these, cases where participants had initially clicked the wrong object and then corrected themselves were very rare (< 1%).

The mean click latencies in the different conditions presented in Figure 8-2 show a rather different pattern from the first fixation results: participants identified the items fastest when they saw a natural pairing of speech and gaze cursor ($M = 3639$ ms from item onset, $SD = 1709$). Latencies in the two lag conditions
resembled each other (5s ahead: \(M = 4559\) ms, \(SD = 2192\); 5s delayed: \(M = 4725\) ms, \(SD = 1569\)); while participants who saw the random cursor were almost a second slower to click (\(M = 5704\) ms, \(SD = 2569\)).

![Figure 8-2: Mean latency to correct click depending on gaze cursor condition and trial number, in ms from item onset.](image)

Two-way analyses of variance with the factors CURSOR (4 levels) and TRIAL (3 levels) showed a main effect of CURSOR: \(F1(3, 20) = 7.35, p < .01; F2(1.9, 67.02) = 44.28, p < .001\) (Greenhouse-Geisser corrected due to lack of sphericity). However, only the difference between natural and random was fully significant in Bonferroni-corrected post-hoc tests (\(p = .001\)). TRIAL had no significant effect by participants (\(p = .22\)), but was marginal by items (\(F2(1.5, 52.5) = 3.09, p = .07,\) Greenhouse-Geisser corrected), with slower clicks in the first trial than in trials 2 and 3. The interaction of CURSOR and TRIAL was not significant (both \(F < 1\)).

To summarise, it had been expected that participants in the 5s ahead condition would be faster to click than those seeing the natural gaze cursor, unless the discrepancy compared to the speech utterance was sufficiently disturbing for them to sacrifice the speed advantage provided by the cursor in favour of accurate decisions based on the descriptions alone. This seems to have been the case: while it may not be entirely clear that participants were actually slower to identify the
objects, they were certainly no faster than participants in the natural condition. It seems then that the asynchrony of five seconds was large enough for participants to lose confidence in the usefulness of speaker gaze. Despite the fact that they looked at the cursor, as evidenced by the predicted pattern for first fixations, their click decision at this lag seems to be based entirely on the utterance they heard. At the same time, the significant difference in click latencies between natural and random shows that the unconnected dot also led to a substantial lack of confidence for the click decision.

8.1.3.3 Summary of recall task results

Full details of the analysis and results of the recall task can be found in Appendix G. Similar to Experiment 2, the type of gaze cursor presented to participants did not affect the number of items they recalled correctly. This was the case both for the type of cursor during encoding in the main experiment and for the type of cursor during recall. There was no effect of whether the cursors in the two tasks matched each other.

8.1.4 Conclusions

To summarise the main results of Experiment 6, it seems that first fixations to the relevant object depended largely on the timing lag between the gaze cursor and speech: the earlier the cursor, the earlier the first fixation. In addition, the lack of a difference between a completely unhelpful dot and a 5s delayed cursor indicates that as the cursor became less informative relative to speech, participants cut down on following it in favour of fixations driven by what they were hearing.

Click latencies, in contrast, appeared to be less influenced by the exact timing of speech and cursor than by whether the two provided consistent information. If this was the case, participants did make use of the gaze cursor − at least as confirmation of their speech-based prediction of the next item − and this led to earlier clicks. However, if the connection between speech and cursor was weak, they ignored the cursor almost completely. The implications of these results will be discussed in greater depth at the end of the chapter. First however, Experiment 7
serves to explore in more detail how varying the lag between cursor and speech affects the way the two stimuli are used.

8.2 Experiment 7: Small lag variations

Experiment 6 established that a gaze cursor that was out of sync with speech by five seconds was experienced as a dubious source of information and profoundly affected both fixation onsets and click latencies. But a lag of five seconds constitutes a massive manipulation, which was specifically selected to be distracting. For this reason, it can’t inform us about whether the naturally occurring timing relationship between speech and cursor is important, or whether participants can use both forms of information as long as they are clearly related. Experiment 7 therefore examined a range of intermediate lags between the five-second extremes.

8.2.1 Research question and Predictions

Five different lags between speech and gaze cursor were tested in the following study: the cursor could be 2s delayed, 1s delayed, natural, 1s ahead, or 2s ahead of speech. Of these, the one-second lags were of particular interest for several reasons: gaze typically precedes speech by roughly one second, so the 1s delayed condition is an approximation of speech and gaze information appearing roughly in parallel (though this is qualified by the speaker’s initial focussing of the little number). In contrast, 1s ahead should slightly boost the predictive information naturally inherent in speaker gaze, without creating a noticeable inconsistency. These conditions therefore used a more powerful design to re-examine the manipulation that had shown no effect in Experiment 4. Finally, a combined analysis of this data compared to two-second shifts and the five-second conditions from Experiment 6 should reveal how the use of speech and gaze cursor information separately and in conjunction develops over time.

Based on the findings of Experiment 6, first fixation onsets were predicted to be strongly dependent on the extent to which the gaze cursor preceded the speech utterance – the further ahead the gaze cursor, the earlier the first fixation, and vice
versa. Click latencies, by contrast, were expected to be larger for greater inconsistencies between the two stimuli. This should be more or less irrespective of whether the gaze cursor was shifted forwards or back in time, relative to speech.

8.2.2 Methods

Five speech-cursor lags were tested between participants, all of whom heard the same descriptions. The order in which participants heard and saw the three rooms was counterbalanced across participants and cursor conditions. As in the previous two studies, the descriptions originated from Speaker 4, and his eye movement record was used to construct the shifted gaze cursors.

For reasons of efficiency, the data from the natural lag condition was used as a baseline for both Experiments 5 and 7, which were tested during the same two-week period. A randomised list controlled the allocation of participants to the natural (or speech & cursor) condition, as well as to 1s delayed and 1s ahead in this experiment, and the speech-only and cursor-only conditions in Experiment 5. The two 2s ahead and delayed conditions were tested immediately following the other five, but all within a space of four weeks.

Hence, all 30 participants included in Experiment 7 came from the same Edinburgh University student population. They were mostly undergraduates, between 18 and 24 years old ($M = 21.7$), and consisted of 16 females and 14 males. All were native speakers of British English, and none had taken part in any of the previous gaze-projection experiments. One participant had to be replaced because of eye problems. Participants’ fixation behaviour was recorded from their right eye, using the EyeLink 1000 tracker in the PVC lab.

Since this study used the same experimental script as Experiment 5, the procedure was exactly as described there for the natural condition. There was no

---

30 Observant readers may note that the reported means for the speech & cursor and natural conditions are not identical in the results of Experiments 5 and 7. This is because, although the two experiments used the same data, relevant gazes were defined differently in the two studies. Details can be found in the respective results sections.
recall task, so participants were usually able to complete the practice and three experimental trials and fill in the questionnaire in less than 30 minutes. As indicated in Appendix E, the questionnaire included a question asking them to guess the condition they had been assigned to. They received £4 for their participation.

### 8.2.3 Results

This section includes only the analyses of first fixation onsets and click latencies for the five lag conditions of Experiment 7. It is followed in Section 8.3 by combined analyses with Experiment 6, including the questionnaire data, before a full discussion in Section 8.4.

#### 8.2.3.1 First fixation onsets

Figure 8-3 shows the distribution of onset times of the first relevant gaze for each lag condition. It reveals that strongly delayed gaze cursors resulted in much wider variance \((\text{Levene’s } F(4, 985) = 36.48, p < .001)\), resembling the speech-only conditions in Experiments 2 and 5.

![Figure 8-3: Onsets of first relevant gazes (wide definition), depending on the gaze cursor lag, in ms from item onset. Each box contains 50 % of the data in that condition, with the midline representing the median. Outliers (more than 1.5 times the IQR outside the central box) and extreme values (more than 3 times the IQR) are represented as ○ and *, respectively.](image)
Nonetheless, the wide definition of relevant gazes was used, meaning that any gaze to the relevant region of interest which began during the previous, current, or subsequent item description was included in the analyses. No such relevant gaze was recorded in 2% of cases; these cases had to be dropped. In addition, outliers and extreme values (+/- 2 SD from the mean per condition) were excluded, resulting in a total of 8.3% missing cases. Because of the many conditions involved, Table 8-2 provides an overview of mean onset times and click latencies, followed by a graphical presentation of first fixation onsets (including trial order) in Figure 8-4.

Table 8-2: Descriptive statistics for the five speech-cursor lags: Onset times of first relevant gazes and click latencies (described in 7.3.2), both in ms from item onset.

<table>
<thead>
<tr>
<th></th>
<th>2s delay</th>
<th>1s delay</th>
<th>natural</th>
<th>1s ahead</th>
<th>2s ahead</th>
</tr>
</thead>
<tbody>
<tr>
<td>First fixation onset</td>
<td>M</td>
<td>558</td>
<td>471</td>
<td>-174</td>
<td>-981</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>3417</td>
<td>1833</td>
<td>1288</td>
<td>1348</td>
</tr>
<tr>
<td>Click latency</td>
<td>M</td>
<td>4068</td>
<td>3615</td>
<td>3181</td>
<td>3541</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>1758</td>
<td>1782</td>
<td>1668</td>
<td>1873</td>
</tr>
</tbody>
</table>

Figure 8-4: Onset times of first relevant gazes (wide definition) depending on the gaze cursor lag and trial number, in ms from item onset.

31 This was necessary to be able to compare delayed with ahead conditions, as the latter usually led to first fixations before item onset. However, it should be borne in mind that many of the early fixation onsets in the delayed conditions will have been due to chance, and that the mean onset time in this condition is therefore artificially reduced.
The usual analyses of variance confirmed that the cursor LAG significantly affected first fixation onsets \((F1(4, 25) = 53.91; F2(1.6, 45.5) = 26.96,\) Greenhouse-Geisser corrected; both \(p < .001\)). There was no effect of TRIAL and no interaction. Planned comparisons (repeated) showed that, apart from a non-significant decrease between \(2s\) delay and \(1s\) delay \((p = .52)\), each further speeding of the cursor relative to speech significantly decreased the time till first fixation (all \(p < .01\)).

### 8.2.3.2 Click latencies

As before, participants’ clicks to the objects were manually compared to the regions of interest. In cases where participants had self-corrected an initial wrong click, only the time of the subsequent correct click was entered into the latency analysis. The resulting accuracy was above 99%.

Mean latencies for the five gaze cursor conditions are presented graphically in Figure 8-5; Table 8-2 (above) includes the numerical values. These reveal that participants were fastest to click the correct item in the natural and \(2s\) ahead conditions \((M = 3181\) ms from item onset and \(M = 3037\) ms, respectively\); somewhat slower in the \(1s\) ahead \((M= 3615\) ms) and delayed conditions \((M= 3541\) ms), and took longest to click following the \(2s\) delayed condition \((M = 4068\) ms from item onset).
Analyses of variance with the factors LAG (5 levels, between participants) and TRIAL (3 levels, within participants) found a main effect of LAG on click latencies ($F_1(4, 25) = 3.53, p = .02; F_2(3.1, 110.1) = 30.88, p < .001$, Greenhouse-Geisser corrected). However, planned comparisons (simple contrasts of each condition against the natural baseline) found that clicks were significantly slower than the natural condition only if the cursor was delayed by two seconds ($p < .01$), although there was a trend in this direction for 1s delayed ($F_1: p = .16, F_2: p < .001$). Clicks in the two ahead conditions did not differ consistently from natural.

A main effect of TRIAL was marginal by participants ($F_1(2, 50) = 2.49, p = .09$) and significant by items ($F_2(2, 70) = 7.83, p = .001$). This was due to somewhat later clicks in the first trial ($M = 3673$ ms from item onset) than in trials 2 ($M = 3407$ ms) and 3 ($M = 3388$). The interaction of LAG and TRIAL was also significant by items ($F_2(4.6, 162.1) = 2.89, p = .02$, Greenhouse-Geisser corrected), but not by participants ($p = .58$). Looking at the graph, it again seems possible that a consistent speed-up would have been found across individual items, although it was not fully significant across trials.

### 8.2.4 Conclusions

As predicted, first fixations to the relative object were quite closely tied to the timing lag between cursor and speech: earlier cursors led to earlier first fixations. The pattern of means suggests that this tendency was less pronounced for cursors delayed behind speech than those preceding it, presumably due to the fact that the first fixation was more often driven by the speech content if the cursor was delayed. In addition, the mean differences between conditions were always smaller than the one-second differences between the actual cursor presentations. Both these observations suggest that while the cursor was generally the main factor affecting gaze behaviour, the content of the speech utterance also had some influence on fixations. The exact relationship is investigated further in the following section, which combines analyses of Experiments 6 and 7.

Click latencies were also longest in the 2s delayed condition; but the predicted pattern of longer latencies for greater inconsistencies between speech and cursor was not found. While natural clicks were indeed faster than those in the delayed and
1s ahead conditions, clicks for 2s ahead were at least as fast. Tentatively, it seems more as if there is a fairly strong link between cursor lag and click latency in the delayed conditions, but that ahead cursors encourage more strategic processing. This could be why there seems to be no clear pattern across participants for these cursor conditions. This suggestion will be explored further by comparing the two ahead conditions to the substantially greater shift of 5s ahead in Experiment 6.

### 8.3 Cross-experiment analyses

Experiments 6 and 7 are complementary in many respects, with fairly extreme lag shifts tested in Experiment 6, and Experiment 7 filling out the smaller discrepancies in between. The following section therefore contains analyses in which the relevant data from both has been pooled, allowing some additional questions to be addressed. An initial combined analysis of first fixations and click latencies (Section 8.3.1) is followed by a comparison of the time from first fixation to when participants actually make the click (8.3.2), comparisons of correlations between fixating and clicking over time (8.3.3), as well as a combined investigation of the questionnaire data in Section 8.3.4.

#### 8.3.1 First fixation onsets and click latencies

Before focussing on the combined analyses in detail, Figure 8-6 summarises Experiments 6 and 7 by graphing the development of first fixation onsets and click latencies across time. It illustrates several findings that have already been mentioned. For instance, first fixation onsets are clearly dependent on the appearance of the gaze cursor relative to speech. Interestingly, however, the relationship with the cursor timing does not seem to be perfect: no difference in mean first fixation onsets between two neighbouring conditions is as large as the

---

32 Unless otherwise mentioned, the two natural conditions were combined and the random condition from Experiment 5 was excluded.
actual difference in timing between the two cursors. For example, the difference in first fixation onset times between 1s ahead and 2s ahead is only 625 ms rather than 1000 ms. Even more strikingly, the difference between 5s delayed and 2s delayed is just 925 ms, not the 3000 ms that would be expected if the moving green dot was so salient that participants automatically looked at it wherever it happened to be.

To test whether the first fixation onsets deviated from the lag differences statistically as well as numerically, the total gaze cursor lag was added to each first fixation onset for all ahead conditions (e.g., onset + 1000 ms for 1s ahead) and subtracted in all delay conditions. Analyses of variance with the factors TRIAL and LAG were run on the means of these new values, with planned contrasts testing whether each condition differed significantly from the natural mean. Note that after adjusting first fixation onsets in this way there should be no significant differences between conditions if participants were solely following the cursor. Conversely, in any case where the difference in fixation onsets between conditions is significant (and, based on Figure 8-6, significantly smaller than the lag difference), the assumption must be that participants’ fixation patterns were influenced by factors other than the cursor. This was indeed the case: the main effect of LAG ($F1(6, 41) = $)
73.55, $F_2(2.6, 66.5) = 33.79$, Greenhouse-Geisser corrected; both $p < .001$) was significant for all comparisons except 1s ahead and natural ($p = .7$); 2s ahead and natural was marginal by participants ($p = .06$). This suggests that the earlier fixations in these two conditions compared to natural were indeed driven mainly by the earlier cursor. However, the highly significant difference between lag-adjusted 5s ahead and natural confirms that a strong inconsistency between speech and gaze cursor affects gaze behaviour as well as click latencies – until now, gaze had been assumed to be driven primarily by the bright moving cursor, irrespective of speech. It seems that, similarly to what was found in the random condition of Experiment 6, the tight link between cursor and gaze was weakened for participants experiencing a discrepancy between the cursor and speech. In addition, the fact that differences between conditions tended to be even smaller for delayed than for ahead conditions (all $p < .05$) confirms the increasing importance of the speech content for driving fixations, as it gains in informativeness relative to the gaze cursor.

Figure 8-6 also shows that the effect of the gaze cursor lag on click latencies is much weaker than on first fixation onsets. Analyses of variance revealed a significant main effect of LAG: $F_1(6, 41) = 6.73; F_2(6, 210) = 48.64$, both $p < .001$. However, planned comparisons (repeated) did not show the stepwise progression found for first fixations: Clicks were significantly later than in the natural condition at both five-second lags ($p < .001$), as well as at 2s delayed ($p = .04$). All other conditions did not differ significantly from natural ($p > .2$). There was also a significant effect of TRIAL ($F_1(2, 82) = 3.59, p = .03; F_2(1.6, 56.3) = 7.62, p < .01$, Greenhouse-Geisser corrected), such that participants consistently took longer to click on the first than on the subsequent trials. The interaction was not significant (both $p > .14$).

So, while clicks tended to be earlier for close couplings of speech and gaze cursor, the precise lag did not seem to be important. Only extremely asynchronous conditions led to considerably longer click latencies, almost irrespective of whether the cursor was shifted forwards or back.
8.3.2 Fix-to-click times

The dissociation between first fixation onsets and click latencies described in the previous section suggests another way of looking at the data: how long does it take from the time a participant first looks at an object until they click on it? Participants seeing a (reasonably helpful) gaze cursor move to the right side of the living room picture would presumably follow it there with their eyes. They would seem likely to make a first guess quite quickly (e.g., that the speaker was going to mention the vase or the flowers next), and would also probably move the mouse in that direction already. However, we know that clicks usually only occurred after linguistic disambiguation, e.g., after the speaker had explicitly mentioned the “vase”, “flowers”, or “bunch of roses”. This makes the clicks a plausible measure of when listeners decided they had accumulated sufficient evidence to confirm their initial guess regarding the next object. Presumably, this evidence could be both in the form of the speech utterance and of the cursor continuing to focus on an object. The fixation-to-click measure therefore provides information on the extent to which looking at an object is helpful for recognising it as the item the speaker is referring to, but it can also further clarify how participants integrated the two stimuli.

The fix-to-click measure was created by subtracting the onset time of the first relevant fixation from the time point of the first correct click to the object. In 36 cases, fix-to-click turned out to be negative. These were removed, as it seems more likely that they were caused by recording problems than that participants clicked on the correct object without once looking at it. Together with cases in which there was no correct click at all (N = 37), this resulted in a total of 4.2 % missing data. Apart from removing negative fix-to-click values, extreme values were not excluded, as they may have been caused by different combinations of factors in the different conditions.

However, as Figure 8-7 and Table 8-3 show, this meant that the data turned out to be strongly positively skewed in most conditions, with long tails towards higher fix-to-click values.
Figure 8-7: Fix-to-click measures for all lag conditions. Each box contains 50% of the data in that condition, with the midline representing the median. Outliers (more than 1.5 times the IQR outside the central box) and extreme values (more than 3 times the IQR) are represented as ○ and *, respectively.

To avoid biasing the analyses by the large number of outliers and extreme values, the following graphs and analyses will therefore be based on aggregated medians, rather than on means. First however, Table 8-3 provides an overview of all the relevant descriptive statistics, while Figure 8-8 presents the median values graphically for each condition.

Table 8-3: Time (in ms) from first relevant fixation to correct click, for each lag.

<table>
<thead>
<tr>
<th>Lag</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>5s delay</td>
<td>198</td>
<td>4351</td>
<td>5579</td>
<td>1435</td>
</tr>
<tr>
<td>2s delay</td>
<td>201</td>
<td>4749</td>
<td>4953</td>
<td>2501</td>
</tr>
<tr>
<td>1s delay</td>
<td>212</td>
<td>4271</td>
<td>4557</td>
<td>2521.5</td>
</tr>
<tr>
<td>natural</td>
<td>418</td>
<td>4012</td>
<td>3058</td>
<td>2998</td>
</tr>
<tr>
<td>1s ahead</td>
<td>209</td>
<td>4896</td>
<td>3141</td>
<td>3973</td>
</tr>
<tr>
<td>2s ahead</td>
<td>208</td>
<td>4754</td>
<td>2497</td>
<td>4324</td>
</tr>
<tr>
<td>5s ahead</td>
<td>209</td>
<td>7898</td>
<td>3922</td>
<td>7629</td>
</tr>
<tr>
<td>Total</td>
<td>1655</td>
<td>4871</td>
<td>4131</td>
<td>3571</td>
</tr>
</tbody>
</table>
Figure 8-8: Median time from first looking at an item to correctly clicking on it, in ms.

Trial was collapsed for the analysis, and analyses of variance compared the effect of LAG (7 levels, between participants, within items) on median fix-to-click latencies. Predictably, it revealed a main effect of LAG: $F(6, 41) = 71.48; F(2.6, 89.4) = 11.12$ (Greenhouse-Geisser corrected), both $p < .001$. Bonferroni-corrected posthoc tests showed that in all ahead conditions participants fixated the relevant item significantly earlier before clicking on it than with a natural speech-to-cursor pairing (the comparison with 1s ahead was marginal by participants ($p = .053$), all other $p < .001$). Fix-to-click latencies for 1s delayed and 2s delay did not differ from natural (both $p > .9$), but 5s delay led to a clear reduction compared to natural ($p < .001$).

The pattern of results suggests that the time spent attending to an object is not particularly important for deciding whether the speaker is referring to it. The pronounced differences in fix-to-click time for the two five-second shift conditions both reflect this: participants in the 5s ahead condition often first looked at the object a long time before its unambiguous mention. But while it seems probable that from this point on they considered it a likely candidate for the next object, they waited to click on it until it was clearly mentioned in the speech. In contrast, participants in the 5s delayed condition had already heard a substantial part of the description by the time they fixated the object, and therefore were able to click on it almost as soon as they recognised it.
Another way of looking at these results is that a gaze cursor preceding speech seems helpful when an absolute reaction time measure is used (first fixation onsets and – to a lesser extent – click latencies), but not with a relative measure which takes into account the time in which participants were already looking at the object and could have clicked it (fix-to-click times). In contrast, a gaze cursor lagged behind speech performs well on the relative measure, but leads to slower clicks in absolute terms (i.e. relative to speech). Clicks remain fast on both measures if speech and gaze cursor are in sync, as in the natural condition.

### 8.3.3 Correlation analyses

For a final alternative exploration of the link between cursor lag, first fixation onset, and click latency, correlations were calculated between the two dependent variables for each lag condition. For this purpose, outliers and extreme values beyond two standard deviations of the mean per condition were excluded from both measures (first fixation onset: 8% excluded; click latency: 4.2%). Table 8-4 presents Spearman’s rank correlations for each lag condition.

<table>
<thead>
<tr>
<th></th>
<th>5s delay</th>
<th>2s delay</th>
<th>1s delay</th>
<th>natural</th>
<th>1s ahead</th>
<th>2s ahead</th>
<th>5s ahead</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>N</strong></td>
<td>193</td>
<td>187</td>
<td>186</td>
<td>384</td>
<td>189</td>
<td>188</td>
<td>186</td>
</tr>
<tr>
<td><strong>rho</strong></td>
<td>.466***</td>
<td>.195**</td>
<td>.252**</td>
<td>.226***</td>
<td>.17*</td>
<td>.132</td>
<td>.129</td>
</tr>
<tr>
<td><strong>p</strong></td>
<td>&lt; .001</td>
<td>= .007</td>
<td>= .001</td>
<td>&lt; .001</td>
<td>= .019</td>
<td>= .071</td>
<td>= .078</td>
</tr>
</tbody>
</table>

Despite the fact that there is a consistent correlation between fixating and clicking, this is much smaller and less reliable for conditions where the cursor is shifted ahead of speech. These results provide further evidence that participants’ handling of the gaze cursor in the ahead conditions was less consistent and predictable than in the natural and delayed conditions. Since first fixations in the 1s ahead and 2s ahead conditions were affected almost exclusively by the gaze cursor (see the lag-adjusted analysis in section 8.3.1), this variability must stem mainly from participants’ click latencies. A likely cause would be increased strategic processing in the ahead conditions, which in turn could be influenced by a number
of different factors (e.g., individual preference, ease of identification of the object, quality of speaker description).

8.3.4 Questionnaire data

The previous sections have presented a number of different ways of comparing fixation and click data depending on the timing relationship between speech and gaze cursor. Prior to a final summary and discussion of the results, this section explores participants’ subjective assessment of the relative importance and usefulness of the two stimuli for their task, as well as the extent to which they noticed any lag manipulation.

Again, the analysis concentrates on those questionnaire items that used a rating scale, for which descriptive statistics are presented in Table 8-5. The comparison of ratings across lag times was based on correlation analyses: for delayed conditions up to natural, negative correlations between lags and ratings were expected with regard to the usefulness of the cursor, the link between cursor and speech, and the extent to which participants tried to follow the cursor. For example, as the lag decreased toward natural, participants were expected to rate the cursor’s usefulness more highly. Positive correlations with lag size were expected for “ignored the cursor”, and “found cursor distracting”. The ratings of speech were not expected to differ across conditions, as all participants had heard the same recordings.
Table 8-5: Descriptive statistics for rating questions (all on a 7-point scale). Ranges are included as a more intuitively understandable measure of dispersion than standard deviation for small sample sizes.

<table>
<thead>
<tr>
<th></th>
<th>5s delayed</th>
<th>2s delayed</th>
<th>1s delayed</th>
<th>natural</th>
<th>1s ahead</th>
<th>2s ahead</th>
<th>5s ahead</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N = 6</td>
<td>N = 6</td>
<td>N = 6</td>
<td>N = 12</td>
<td>N = 6</td>
<td>N = 6</td>
<td>N = 6</td>
<td>N = 48</td>
</tr>
<tr>
<td>usefulness of descriptions</td>
<td>M</td>
<td>6.0</td>
<td>5.0</td>
<td>5.5</td>
<td>5.0</td>
<td>5.5</td>
<td>5.17</td>
<td>6.33</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>.89</td>
<td>1.1</td>
<td>1.38</td>
<td>1.54</td>
<td>1.87</td>
<td>1.61</td>
<td>.82</td>
</tr>
<tr>
<td></td>
<td>Range</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>clarity of descriptions</td>
<td>M</td>
<td>5.33</td>
<td>4.83</td>
<td>5.5</td>
<td>5.08</td>
<td>5.17</td>
<td>4.67</td>
<td>5.5</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>.82</td>
<td>1.47</td>
<td>.84</td>
<td>1.24</td>
<td>.98</td>
<td>1.86</td>
<td>1.05</td>
</tr>
<tr>
<td></td>
<td>Range</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>usefulness of cursor</td>
<td>M</td>
<td>4.0</td>
<td>5.83</td>
<td>6.17</td>
<td>6.0</td>
<td>6.5</td>
<td>6.17</td>
<td>5.0</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>1.41</td>
<td>1.47</td>
<td>1.17</td>
<td>.85</td>
<td>.55</td>
<td>1.17</td>
<td>2.19</td>
</tr>
<tr>
<td></td>
<td>Range</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>how tightly was the cursor linked to the speech?</td>
<td>M</td>
<td>3.83</td>
<td>5.83</td>
<td>6.0</td>
<td>6.08</td>
<td>6.0</td>
<td>6.5</td>
<td>5.17</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>.75</td>
<td>.75</td>
<td>.89</td>
<td>.67</td>
<td>.63</td>
<td>.55</td>
<td>1.47</td>
</tr>
<tr>
<td></td>
<td>Range</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>ignored cursor; concentrated on speech</td>
<td>M</td>
<td>5.0</td>
<td>2.83</td>
<td>3.33</td>
<td>3.08</td>
<td>2.33</td>
<td>2.33</td>
<td>3.0</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>.63</td>
<td>1.33</td>
<td>1.51</td>
<td>1.78</td>
<td>1.51</td>
<td>1.37</td>
<td>1.67</td>
</tr>
<tr>
<td></td>
<td>Range</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>tried to follow cursor</td>
<td>M</td>
<td>3.33</td>
<td>5.5</td>
<td>5.83</td>
<td>5.17</td>
<td>6.0</td>
<td>5.33</td>
<td>5.67</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>1.03</td>
<td>1.23</td>
<td>.41</td>
<td>1.03</td>
<td>1.1</td>
<td>1.21</td>
<td>1.86</td>
</tr>
<tr>
<td></td>
<td>Range</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>found cursor distracting from task</td>
<td>M</td>
<td>2.67</td>
<td>2.83</td>
<td>2.67</td>
<td>2.50</td>
<td>3.00</td>
<td>1.33</td>
<td>2.0</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>1.03</td>
<td>2.14</td>
<td>.82</td>
<td>1.51</td>
<td>2.1</td>
<td>.52</td>
<td>1.1</td>
</tr>
<tr>
<td></td>
<td>Range</td>
<td>2</td>
<td>5</td>
<td>2</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

In fact, this is exactly what was found: cursor lag correlated negatively (Kendall’s $\tau_b = -0.49$, $p < .01$) with the experienced strength of the link between cursor and speech. The predicted negative correlation of lag size with rated usefulness of the cursor just missed significance ($\tau_b = -0.29$, $p = .057$), as did the predicted positive correlation with participants’ explicit strategy of ignoring the dot ($\tau_b = 0.29$, $p = .057$; $N = 30$ for all analyses). A negative correlation with the strategy of trying to follow the dot was also marginal ($\tau_b = -0.27$, $p = .08$), though
there was no evidence that participants experienced the dot as distracting
($\tau_b = .08, p = .63$). Ratings of the descriptions did not correlate with cursor lag
(both $p > .27$). For natural and ahead conditions, the opposite pattern would be
expected, if participants were aware of making use of the predictive qualities of the
gaze cursor. Interestingly however, this was not the case: cursor lag didn’t correlate
significantly with any of the rating items (all $p > .17, N = 30$). This is the clearest
evidence yet that participants in ahead conditions differ in how they experience and
use the cursor, with no single strategy emerging.

Finally, were participants retrospectively able to guess which condition they
had been in? Towards the end of the questionnaire they were informed of the
experimental conditions and asked to tick a box next to the one they thought they
might have experienced. Because it seemed unlikely that participants would notice
the difference between a one- and two-second shift, any correct identification of a
delayed or ahead cursor was scored as accurate, even if the exact lag was not correct
(e.g., if 1s delayed was ticked when the participant had actually experienced a 2s
delayed cursor, this would be counted as a correct guess). Also, due to the different
options available in Experiments 6 (five-second lags and random) and 7 (one- and
two-second lags), these guesses will not be compared statistically. However, the
overview contained in Table 8-6 shows a very clear pattern: participants noticed a
delayed cursor only if it lagged behind speech considerably. They were more
accurate for ahead cursors, although this is presumably partly due to the fact that
natural and 1s and 2s ahead were the most popular guesses for all conditions.

<table>
<thead>
<tr>
<th>Actual cursor lag</th>
<th>Condition guess</th>
<th>Total</th>
<th>% correct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5s delay</td>
<td>2s delay</td>
<td>1s delay</td>
</tr>
<tr>
<td>5s delay</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2s delay</td>
<td>1</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>1s delay</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>natural</td>
<td>6</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>1s ahead</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2s ahead</td>
<td>1</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>5s ahead</td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>Total</td>
<td>4</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
In fact, this pattern may help to explain the tight link between fixation and clicking for delayed conditions but not for ahead. If participants in delayed conditions actually tended to remember the gaze cursor as being natural or even ahead, and therefore helpful and predictive of speech, this suggests that they used it extensively, continuously integrating the visual information with the utterance. For participants in the ahead conditions, the same strategy may have been somewhat less helpful, as the connection between speech and gaze cursor was actually weaker (remember that speakers tend to look at objects before they mention them), so such integration may have been harder to accomplish.

8.4 Summary and Discussion

The two experiments described in this chapter primarily investigated the effect of shifting the gaze cursor in time relative to speech, thereby making it more or less informative. Experiment 6 compared participants’ use of gaze cursors that had been subjected to comparatively large shifts forwards and backwards in time, as well how they dealt with a completely uninformative cursor.

The speech-cursor lag had a strong influence on first fixation onset times, such that earlier gaze cursors (relative to speech) led to earlier fixations. However, this was not solely due to attention being attracted by the bright green moving dot. Particularly in the delayed condition, participants’ first fixations were substantially earlier than expected based on the cursor, suggesting that they were attending more to the content of the speaker’s description. A similar result was found for the random cursor: first fixations in this condition were in fact no later than in 5s delayed. This shows that participants were not distracted by the uninformative cursor and seem to have been able to ignore it most of the time.

A different pattern emerged for click latencies, i.e. the time before participants actively identified the object being described. Although accuracy remained at ceiling for both the 5s delayed and the random conditions, participants were significantly slower to click on items if they only had a random gaze cursor to assist them. There was also a suggestion that both five-second shifted cursors
direction made it harder to locate the object correctly, though neither comparison
was fully significant.

The proposal that clicks to the objects were based almost entirely on the
speech utterance, irrespective of the time spent looking at them, was borne out by
the results of Experiment 7, which explored smaller, one- and two-second lag shifts.
This was particularly true for ahead conditions, which did not differ consistently
from the natural baseline, though in the delayed conditions participants were
somewhat slower to click.

A number of cross-experimental analyses confirmed this proposition:
participants in ahead conditions took much longer from first fixation to correct click
than those in other conditions. In addition, the correlation between first fixation and
correct click time was considerably weaker for them, suggesting that they were
inclined to look around searching for the correct object. Finally, their subjective
ratings of the gaze cursor’s usefulness on a questionnaire presented after the main
experiment also showed a much greater amount of variability than that found
among participants in delayed conditions.

It seems then that while seeing a lag-shifted gaze cursor accompanying a
speaker’s description affects the time to fixate the described items in a very
predictable way, how it is integrated with the speech utterance and hence its effect
on the time to click on the item is more complicated. Large inconsistencies between
speech and cursor, such as those created in the five-second shift conditions and with
the random dot, cause participants more or less to abandon the cursor completely
and rely just on the speech utterance, which allows them to identify the objects
accurately. How listeners responded to smaller inconsistencies depended on the
direction of the shift: a gaze cursor which was displayed one or two seconds later
than the speaker’s original gaze led to reliably later clicks to the objects, as had been
expected – despite the fact that participants seem to have had no awareness of
anything unusual.

In contrast, no clear pattern emerged for gaze cursors that were shifted
slightly ahead of the speaker’s gaze. In some cases, this predictive cursor may
actually have been helpful, as suggested by the somewhat shorter click times in the
2s ahead condition compared to natural. In other cases, it was not, creating overall
click times that were roughly comparable to the natural condition. Interestingly, the correlations between first fixations and click latencies were extremely small in these conditions, and participants’ subjective impressions of the cursor varied widely as well.

It seems likely that a number of factors contributed to the rather ambiguous results in the ahead conditions. These could include participants’ prior assumptions about the relationship between eye movements and speech, their individual willingness to risk clicking on the wrong object, the ease of identifying the objects, and the descriptiveness and sentence structure of the speaker’s actual description. However, it does seem as if providing gaze information somewhat earlier than it could realistically occur does not lead to a clear benefit for comprehension.

On the other hand, from a more applied point of view, this is probably a good thing, since relaying naturally occurring or slightly delayed gaze-speech combinations is technically much more feasible in an online setting than projecting an interlocutor’s gaze ahead of their speech. On-time or slightly delayed gaze information also seems more likely to occur in face-to-face conversation, unless the speaker is extremely disfluent: detecting the object of a speaker’s gaze will generally involve at least two fixations (first to his or her eyes and then in the direction of the gaze), so this information is likely to be subject to delay quite frequently. In contrast, while speakers require a certain amount of time to produce an object name after looking at it, listeners can process the auditory information as soon as it is articulated, and even while they are visually attending to something else.

Slightly delayed visual confirmation is therefore a frequent occurrence, and this study provides further evidence that such visual cueing can be very helpful for listeners, whether or not they are aware of the connection between the moving dot and the speaker’s gaze. The implications will be discussed further in the concluding chapter.
Chapter 9: General Discussion

This chapter serves to round off and conclude the thesis on a number of different levels. To begin with, I summarise the results and conclusions for each of the experiments described, focussing on what they show regarding eye movements in comprehension and production, and what implications this might have for dialogue. Next, Section 9.2 looks at the gaze-projection paradigm as a whole, discussing its limitations, but also suggesting interesting possibilities for further investigation. These include novel research questions concerning interactive language use, as well as potential applications. The final section (9.3) takes a step back from the specific details of the paradigm and puts forward a more general conclusion on the basis of the work reported here.

9.1 Summary of the experiments

The aim of this thesis was to examine how eye movements during comprehension and production relate to each other, and the implications this might have for interactive language use.

To this end, Experiment 1 used a syntactic priming paradigm with confederate scripting, comparing eye movements during comprehension of simple descriptive sentences to those recorded during the subsequent production of similar sentences (e.g., the nun is punching the policeman). As predicted on the basis of previous literature (Tanenhaus et al., 1995; Altmann & Kamide, 1999), eye movements in comprehension were sensitive to listeners’ developing interpretation of the input, based on all forms of information available up to that point. Thus, upon hearing the first noun phrase (the nun), participants responded by fixating the mentioned entity. However, once the identity of the first referent was established, they were able to anticipate the referent of the N2; continued fixations throughout its mention revealed that participants had predicted it correctly. In addition to these effects of incremental comprehension, visual and scene-semantic factors such as the
relative positions of the entities on the screen and their thematic roles also affected
the overall likelihood of their fixation during comprehension.

Eye movements during production were connected much more tightly with
whichever entity was going to be mentioned next, confirming that they primarily
reflect the ease of lexical selection and phonological retrieval (Meyer et al., 1998;
Griffin & Bock, 2000; Griffin, 2001). Yet fixations were not driven exclusively by
upcoming mention, but also by whether the target structure had been primed: the
agent was more likely to be fixated sentence-initially before production of an active
target if the prime had been active as well, compared to when the two sentences
differed in voice. The effect was not very strong, but it seems quite credible that
focussing on a particular thematic role during comprehension increases the
likelihood of fixating the corresponding player while preparing for production. This
could facilitate thematic role assignment, thus contributing to the preference for
repeating the prime structure.

Finding priming of eye movements supports the suggestion that successful
interlocutors show somewhat similar patterns of fixation (Richardson & Dale, 2005;
Richardson et al., 2007), although the association is not perfect. One plausible
explanation for how this resemblance develops is through alignment spreading
interactively between levels of representation (Pickering & Garrod, 2004), in this
case between the visual-conceptual and the syntactic level. So, although the specific
time-course and determinants of fixation are likely to vary between comprehension
and production as a result of differences in the function and motivation of gazes, the
general tendency to adopt the same perspective as one’s interlocutor may well be at
the heart of the ubiquity of repetition and priming in interactive language use
(Garrod & Pickering, 2004).

Experiments 2-7 examined the relationship between gaze patterns in
comprehension and production from a different angle. Assuming that visual
coordination with the speaker would assist comprehension, a gaze-projection
paradigm encouraged listeners to adopt the speaker’s visual perspective onto the
scene he was describing (Velichkovsky, 1995; Brennan et al., 2008). Experiment 2
contrasted fixation onsets and overt object-identification times between listeners
who saw a green dot reflecting where their speaker was looking and those who did
Chapter 9: Discussion

not. As in all experiments in this series, seeing a gaze cursor made it much easier to recognise which object the speaker was describing.

Experiment 3 replicated this facilitation, but found no effect of varying what listeners assumed the cursor to represent. However, the abstract manipulation based only on differences in the instructions may not have been strong enough to create a measurable effect, compared to the blatant visibility of the cursor. For this reason, the cursor itself – and thus its relation to the speech utterance – was manipulated in all following experiments.

Experiment 4 hinted that listeners were not just mechanically following the cursor, but it also revealed a number of difficulties with the original paradigm. This resulted in a number of changes to the design of subsequent experiments, including presenting the same speaker’s description and gaze trace to all participants, varying cursor conditions only between listeners, and recording eye movements using a more drift-resistant eyetracker.

Using this revised version of the paradigm, Experiment 5 compared the respective contributions to listeners’ fixation patterns and click latencies of the auditory description and the visual gaze information. Both measures were primarily affected by the gaze cursor: whether listeners heard the description as well reduced the number of identification errors but did not influence the temporal pattern of their gaze and click behaviour.

On its own, this might imply that listeners almost unavoidably follow the cursor wherever it goes; its usefulness would seem to be due to the natural association with what the speaker is about to mention. However, Experiments 6 and 7 qualified this conclusion, varying the spatial and temporal link between speech and cursor. A random cursor with no connection to what was being said did not lead to slower first fixation times than one which was delayed by five seconds but otherwise accurate, though listeners were somewhat slower to click on the correct object. This indicates that participants in the random condition were at least partially able to disengage from the distracting cursor and perform the task satisfactorily based on the spoken description. Even more convincingly, while variations of the temporal lag between speech and gaze cursor consistently affected first fixation onsets (such that ‘faster’ cursors lead to earlier fixations), the effect was always
smaller than the actual difference between each condition and the natural lag. This was particularly obvious for ‘delayed’ cursors, where the speaker’s fixations were projected later than they had actually occurred. In these cases, as the cursor became less informative relative to speech, listeners were less likely to follow it and more likely to use the description to locate the object. In contrast to fixation onsets, click latencies were hardly affected at all for small lag variations (one second in either direction). Only large shifts created a sufficient inconsistency between cursor and speech to impair object identification.

In short, although the gaze of a speaker producing an unscripted description of a messy real-world scene moves around much more than in typical visual-world-style displays, it still contains enough information about the speaker’s speech planning to provide massive benefits to the listener’s comprehension. This is the case even when the gaze cursor is delayed relative to when the speaker actually looked at the objects, as long as an apparent connection between cursor and speech remains. Thus, the exact timing between the two forms of information seems comparatively unimportant; instead, listeners appear to integrate the cursor as one form of information among others (e.g., expectations based on the depicted objects or the quality of the description), all of which contribute to their developing interpretation of the utterance.

It seems likely that providing such a cursor in a truly interactive conversation would further increase its usefulness, as it would allow a fast and direct assessment of whether the interlocutors have arrived at a shared representation (though see Bard et al., 2007). The following sections discuss such potential extensions of the gaze-projection paradigm, as well as the implications of the current findings.

### 9.2 Discussion of the gaze-projection experiments

The most obvious conclusion from the gaze-projection experiments reported in this thesis is that seeing an on-screen representation of the speaker’s gaze is exceedingly useful for identifying the object being described. In fact, previous research suggests that this effect is not specific to object descriptions, but generalises to a large number of situations that involve generating some form of shared
representation (e.g., Velichkovsky, 1995; Stein & Brennan, 2004; Yu et al., 2005; Brennan et al., 2008; Brown-Schmidt et al., 2009).

9.2.1 Limitations: A return to the ‘real gaze’ question

Yet despite the pervasiveness of the effect, it is less clear what causes the facilitation, and even whether it has the same source in the different tasks and experiments. For instance, it seems fairly clear in Brennan et al.’s (2008) joint visual search task that the cursor was considered as a signal of the partner’s actual gaze. But its value in Velichkovsky’s (1995) setup could plausibly have been due to strategic use on the part of the speaker, while in a third set of studies (perhaps Stein & Brennan, 2004, or even the experiments in the previous chapters) the effect may have resulted from the attraction of visual attention to a useful location. Yu and colleagues (2005) provide yet another account of the facilitatory effect: in their model, eye movements are microlevel behaviours which allow an online and temporally contingent estimation of what someone is about to say or do – in fact, they see eye movements as embodied representations of intention (see also Ballard et al., 1997).

All these potential explanations of how the projected gaze cursor works lead back to the question of what viewers interpret it as representing (see also Chapter 3, Section 3.2.2). Although the experiments reported in this thesis suggest that the cursor does not monopolise listeners’ attention to such an extent that they ignore other forms of information completely, many of the results do point to a largely attentional basis for the effect. At the same time, participants seemed quite comfortable with instructions stating that the green dot represented where the speaker had been looking, and post-experimentally used phrases such as “he kept looking in the top right corner” to describe the movement of dot. This suggests that the cursor was interpreted as speaker’s gaze to some extent at least, though whether this means that it was also processed in the same way as face-to-face gaze must remain open.

In fact, this question may not be as important as it appears at first glance: the variety of tasks which have been combined with projected gaze confirms that such cursors are effective at enhancing joint attention on critical aspects of the shared
environment between interactants. Actually, it is likely that even within a single interaction, the development of this kind of shared perspective can have different causes – sometimes because the bright moving dot attracts attention to a region which happens to be task-relevant, but at other times because one interlocutor deliberately checks to see what the other is looking at.

At least in this respect, processing of the gaze cursor could actually be quite similar to processing of real gaze: on the one hand, perception of eye gaze can trigger reflexive shifts of attention in the relevant direction (Driver et al., 1999; Langton et al., 2000; Ricciardelli et al., 2002a; Castelhano et al., 2007). On the other hand, even young children look at the speaker to determine where he or she is looking (Baldwin, 1995; Brooks & Meltzoff, 2002), particularly on encountering a mismatch between a heard utterance and their current locus of attention – a deliberate recruitment of visual feedback for the sake of disambiguation. Similarly, the speaker’s gaze behaviour can be incidental or deliberate too. Yet the information contained in the gaze is useful in all cases, and assists the listener in preparing a timely and accurate response.

This makes the gaze-projection paradigm a valuable tool for addressing quite diverse questions. In the following section, I begin by suggesting further improvements and interesting modifications of the experiments reported here, before considering some more far-reaching applications.

### 9.2.2 Possibilities for further studies

Clearly, there are ways in which the experiments in this thesis could be improved. In particular, the initial speakers recorded for Experiment 2 varied so widely both in their descriptions (e.g., level of detail, extent of disfluent speech) and in the gaze recordings (amount of drift) that it was very difficult to compare listeners who had heard different speakers. So, while the decision to use only one speaker from Experiment 4 onwards was essential with regard to obtaining comparable listener data, it meant that a lot of data regarding the speaker was not taken into account. Yet characteristics of a particular speaker might have implications for the usefulness and interpretability of the gaze cursor and its relation to speech. In view of the unusual task and the cluttered nature of the scenes,
it would certainly be worth comparing speech-related eye movements for production as well, exploring in a more systematic way how tightly gaze and speech really are linked in this kind of setup. Similarly, it would be interesting to see whether telling speakers that their gaze will later be shown to listeners would lead them to monitor and manipulate it more consciously, as in Velichkovsky (1995).

From the listeners’ perspective, it is conceivable that the relative importance of the gaze cursor varies depending on the speech rate and on whether the speaker uses a very detailed description. It would therefore be desirable to record speakers in a slightly more controlled setting, while maintaining the naturalness of their spontaneous production. This could be done by way of instructions, examples, or a training session before the actual recording. In retrospect, I would also like to have used more photographs with fewer items per room. This would ensure that items are not too close together and allow more frequent calibration or drift correction.

Another idea which would require creating new stimuli would be to include more deliberate contrasts between objects, such that their description would be likely to create temporary ambiguities (cf. Brown-Schmidt & Tanenhaus, 2008). For example, the kitchen picture could show a red saucepan and a red mixing bowl, or a bottle of wine and a bottle of whisky. In fact, it would probably not affect speakers’ eye movements substantially to have them learn specific terms for the items prior to the recording (e.g., a saucepan vs. a cooking pot) or indeed instruct them to use specific constructions (the pot that’s red vs. the red pot; Cleland & Pickering, 2003).

This kind of control over speaker utterances would make it possible to address a question similar to the rationale of Experiment 1, but using gaze projection: does visual coordination (in the sense of a shared perspective generated by the gaze cursor) affect the extent of lexical or syntactic priming, as interactive alignment might predict (Pickering & Garrod, 2004)? Of course, such a study would require the listener to speak as well. This could be achieved by using a confederate as in Experiment 1, with the two participants taking turns to describe highlighted objects to each other. Alternatively, as in Experiments 2-7, the speaker/confederate could be recorded first, and the listener would be asked to remember the

33 The latter contrast was actually present in the kitchen picture, but there were not enough similar items to compare performance on such ambiguous pairs.
instructions and repeat them for a new listener. This idea relates back to the memory tasks described in Experiments 2 and 3, which would also be worth returning to in a more systematic fashion.

Further insights into the use of speech and gaze information in comprehension could be gained even from simple modifications of the paradigm. The cursor itself could be manipulated in several more ways, among which I think hand-crafting a cursor which is not actually based on the speaker’s gaze is one of the most interesting: Bard et al. (2007) show that a misleading artificial cursor is likely to be ignored, but how about a cursor which is in some way better than the speaker’s gaze, for instance in not pointing to unrelated objects? Such a ‘perfect’ cursor would provide a strong test of just how much improvement to comprehension can be expected on the basis of projected gaze. An alternative strategy would be to filter real gaze traces for informative gazes and to project only these. For instance, Velichkovsky (1995) proposed that long fixations (e.g., > 500 ms) are more likely to be task-related than shorter fixations, and thus may be more suited to fulfilling a communicative function. To my knowledge, no subsequent research on eye movements in production has investigated potential differences in significance depending on fixation length, but duration would certainly be a convenient criterion for filtering a gaze record. In fact, Velichkovsky’s speculation could be easily tested by contrasting the effects of projecting only long fixations with a gaze cursor based on shorter fixations.

After finding no effect of manipulating what the cursor was said to represent in Experiment 3, the question of whether listeners’ interpretation of the cursor can affect how they make use of it was not explored further. To do this, it would seem important to find a stronger and more plausible way of manipulating this interpretation. One possibility for making the cursor less obviously related to speech would be to record speakers in a dual-task situation. For instance, while producing their descriptions they might be required to monitor for a suddenly appearing target elsewhere in the display. In addition, an important baseline condition would be not to tell listeners about the dot at all, and to see how this affects the extent to which they follow it, as well as to note their guesses about what it represented.

Particularly for this kind of manipulation it might be necessary to look more closely at the time-course of listener fixations. In fact, if I continue to use this
paradigm it may be worth considering additional dependent measures, and possibly also employing other methods of analysis (cf. the discussion for Experiment 1 in Section 2.2.3.2). The novelty of the design of the gaze-projection experiments meant that more advanced methods of analysis could not be adapted to them easily. Analysing latencies only avoided some of the pitfalls of using ANOVA for eye movement data, but the wide spread of variance remained a concern, and I am sure that a more appropriate form of analysis could be found, given time. Among other things, this should involve rethinking the two competing definitions of relevant gazes, neither of which is ideally suited to the data (see the repeated discussion in Sections 5.2.5, 7.3.1, and 8.1.3.1). More convincing results might be obtained fairly simply by making relevant gazes contingent on the speaker having fixated the object before the listener. Another interesting possibility – probably additionally to, rather than instead of inferential statistics – would be to conduct cross-recurrence analyses (Eckmann et al., 1987), promoted for the study of dialogue by Richardson and colleagues (Richardson & Dale, 2005; Richardson et al., 2007). This would allow a precise quantification of the amount of overlap between the speaker’s and listener’s perspectives, and an estimation of whether this really facilitates comprehension.

Finally, the aim in establishing gaze projection as a paradigm for investigating the interplay of visual attention between interlocutors is obviously not just to understand the effects of seeing a dot on the screen, but to use it as a tool for experimentally investigating the processing of natural language in situations which rely on establishing joint visual attention. As described in Chapter 3, these could range from issues relating to language acquisition, via establishing memory representations, to task success in collaborative interaction. One example is outlined in the following.

In creating the paradigm, I originally hoped to be able to include a gaze cursor in a truly interactive setting, where the speaker is actually co-present and able to adapt his or her speech according to the listener’s perceived requirements or feedback. One possibility would be to present pairs of interlocutors with slightly different versions of the same photographic scene. Their joint task could be to describe the pictures to each other in order to discover the differences (as in the Spot-the-Difference pilot study reported in Appendix C). One of the two
participants would be eyetracked, and the gaze trace projected to the other’s screen. An initial series of experiments could test whether the results of the experiments reported here generalise to real-time interaction. In addition, the role of the participant whose gaze is available could be varied by designating one of them as speaker, and allowing different amounts of feedback from the listener. A follow-up experiment based on these findings could then look at full interaction, where roles would not be designated in advance. If two eye trackers were available, it would be interesting to record both interlocutors’ eye movements, but this is not essential: it seems plausible that the availability of one partner’s eye movements might lead pairs of interlocutors to adopt consistent strategies of interaction (e.g., the eyetracked person might frequently end up as the primary speaker).

This programme of research into the respective and combined effects of verbal and visual feedback could subsequently be extended in several directions. These include comparing the usefulness of visual feedback in different kinds of tasks, or using it in interaction with a virtual agent or robot. Alternatively, performance with the projected cursor could be contrasted to a setting where actual face-to-face gaze is available. Since it is relatively simple to combine a gaze cursor with a speech recording, another avenue which is clearly worth exploring is the use of gaze projection in distance learning, online tutoring, remote problem-solving, or video-conferencing (see the discussion in Stein & Brennan, 2004).

The suggestions outlined in this section constitute a rather loosely-connected set of ideas for future studies involving the gaze-projection paradigm. What I hope to have shown here and in the previous chapters is that the paradigm is a useful and versatile tool which can be adapted relatively easily to a variety of research questions and applied settings, and to the further investigation of issues of theoretical and practical relevance.

9.3 From comprehension to production and back again: The role of coordinated attention

The scope of topics covered in this thesis is broad, spanning the visual and conceptual processing of real-world scenes, the architectures and mechanisms of
language production and comprehension, how these monologue modalities interact in dialogue, as well as discussions of gaze perception and the role of joint attention. Experimentally, two paradigms were used to assess the effects of coordinated visual attention between speakers and listeners on different aspects of language use. In Experiment 1, the primary direction of influence was from comprehending an utterance to producing a similar one, and the effect was expected in the syntactic domain, as well as in eye movements. In contrast, Experiments 2-7 investigated how production-related eye movements affected the speed and accuracy of comprehension.

Though the individual experiments were designed to test specific questions, overall they appear to support one conclusion: that coordination plays an important role in smoothing potential difficulties at the point of transition between comprehension and production. In Experiment 1, syntactic alignment between a scripted confederate and a naïve participant resulted in greater similarity of eye movement patterns between comprehension and production than in the absence of syntactic coordination. In Experiments 2-7, providing an opportunity for listeners to share visual attention with a previously recorded speaker facilitated listeners’ comprehension of his object descriptions.

At the same time, joint attention was not essential for successful completion of the tasks: participants in Experiment 1 could choose between two equally correct syntactic structures in each case, and actually they frequently didn’t repeat the one used by the confederate. In the gaze-projection experiments, while access to the speaker’s perspective made it easier to locate the correct object, click accuracy was at ceiling even when the cursor attracted attention to locations where the speaker had not been looking.

It seems then that language processing is extremely flexible, opportunistic, and sensitive to context: both in comprehension and in production any additional forms of information are readily integrated as soon as they become available. However, the systems involved are so efficient that they will usually only consider supplementary information if this doesn’t require effortful and time-consuming extra processing.
I would argue that sharing an interlocutor’s perspective may make the outcome of particular processing steps – such as thematic role assignment or the identification of an object – available earlier, thereby simplifying the current processing task. Ultimately, this should make task-switches between speaking and listening in dialogue comparatively easy, and enable a smooth succession of turns between interlocutors.
References


Frazier & K. Rayner (Eds.), *Perspectives on Sentence Processing* (pp. 225-242). Hillsdale, NJ: LEA.


References


References


References


Schoonbaert, S., Hartsuiker, R. J., & Pickering, M. J. (2007). The representation of lexical and syntactic information in bilinguals:


References


References


Appendix A. Eyepriming: Experimental items

The following Table contains all combinations of prime image, prime sentence, and target image used in Experiment 1, in the format Agent – Verb – Patient.

For items 1-12, the prime sentence matched the prime image; for items 13-24 it did not. The structure of the prime sentence (active or passive description) and the orientation of the image (agent positioned left or right) was varied between participants. On target images, the patient was always positioned on the left. A different pseudo-randomised presentation order was created for each participant, with at least two pairs of filler items between each experimental prime-target pair. Fillers showed intransitive actions such as “the clown is pointing”, or “the burglar is crying”.

<table>
<thead>
<tr>
<th>Prime Image</th>
<th>Prime Sentence</th>
<th>Target Image</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 policeman – scold – monk</td>
<td>policeman – scold – monk</td>
<td>teacher – scold – sailor</td>
</tr>
<tr>
<td>7 pirate – chase – burglar</td>
<td>pirate – chase – burglar</td>
<td>policeman – chase - swimmer</td>
</tr>
<tr>
<td>Prime Image</td>
<td>Prime Sentence</td>
<td>Target Image</td>
</tr>
<tr>
<td>-------------</td>
<td>----------------</td>
<td>--------------</td>
</tr>
</tbody>
</table>
Appendix B.  Gaze-projection: Photos and ROIs

The following photographs of three rooms were presented on-screen in all gaze projection experiments. The yellow shapes show the region of interest (ROI) around each of the twelve objects: Any fixation falling within these outlines was counted as pertaining to that item.
Appendix C.  Gaze projection: Item selection

As mentioned in Sections 1.4.1.1. and 4.3.1, the photographic stimuli and most of the items used in the gaze projection experiments were derived from a ‘Joint Spot-the-Difference’ pilot study which I had conducted previously. In this study, 15 pairs of participants received coloured printouts of two slightly different photographs of three rooms. They could not see the other person’s printout, so they described their pictures to each other in order to find differences between the two (the example of a dialogue interaction presented in Section 1.4.1.1 is based on one pair’s description). For the pilot study, six versions of each room were set up and photographed from the same angle: each of the twelve designated items remained the same on four of the pictures, changed identity once (e.g., the iron was exchanged for a vacuum cleaner; the candles were extinguished), and once was missing entirely (e.g., no notebook on the table in the living-room photo, see Appendix B).

The stimuli for all the gaze projection experiments (Experiments 2 to 7) re-used one of the six photographs of each room, as well as most of the items. However, the pilot study did show that some of the items were hard to recognise. Therefore, items which had been mentioned by less than 12 of the 15 pairs were replaced by other comparable objects. Items which were kept from the original Spot-the-Difference study are italicised in the following Table C-1, which presents a list of the items selected per room for all the gaze projection studies.

<table>
<thead>
<tr>
<th></th>
<th>Living room</th>
<th>Kitchen</th>
<th>Bedroom</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>iron</td>
<td>duster</td>
<td>blue pillow</td>
</tr>
<tr>
<td>2</td>
<td>pink roses</td>
<td>ladle</td>
<td>plant</td>
</tr>
<tr>
<td>3</td>
<td>upright frame</td>
<td>spotty jars</td>
<td>laptop</td>
</tr>
<tr>
<td>4</td>
<td>pink cushion</td>
<td>wine at top</td>
<td>tiger</td>
</tr>
<tr>
<td>5</td>
<td>phone</td>
<td>apricots</td>
<td>t-shirt</td>
</tr>
<tr>
<td>6</td>
<td>candles</td>
<td>kettle</td>
<td>tissues</td>
</tr>
</tbody>
</table>
In addition, Experiment 2 also included a Spot-the-Difference memory task (see Sections 4.3.4.2 and 4.4.3), based on one of the other photographs of the three rooms. This meant that a maximum of eight differences could be discovered for each room: four of these were changes to an experimental item, four concerned new objects that had previously not been mentioned. The following Table C-2 presents these eight differences per room.

<table>
<thead>
<tr>
<th>Living room</th>
<th>Kitchen</th>
<th>Bedroom</th>
</tr>
</thead>
<tbody>
<tr>
<td>7 notebook</td>
<td>saucepan</td>
<td>map</td>
</tr>
<tr>
<td>8 cafetière</td>
<td>pannetone</td>
<td>cactus</td>
</tr>
<tr>
<td>9 magazine</td>
<td>extractor</td>
<td>smoke detector</td>
</tr>
<tr>
<td>10 bundle of reeds</td>
<td>whisky</td>
<td>waste paper basket</td>
</tr>
<tr>
<td>11 hifi</td>
<td>egg box</td>
<td>stereo</td>
</tr>
<tr>
<td>12 straw horse</td>
<td>bucket</td>
<td>hot water bottle</td>
</tr>
</tbody>
</table>

<p>| Table C-2: Experimental items and Spot-the-Difference variants for the memory task following Experiment 2. |
|---|---|---|---|---|</p>
<table>
<thead>
<tr>
<th>item</th>
<th>difference</th>
<th>item</th>
<th>difference</th>
<th>item</th>
<th>difference</th>
<th>Type of difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 iron</td>
<td>no iron</td>
<td>duster</td>
<td>no duster</td>
<td>blue pillow</td>
<td>no pillow</td>
<td>absent</td>
</tr>
<tr>
<td>2 pink roses</td>
<td>no flowers</td>
<td>ladle</td>
<td>no ladle</td>
<td>plant</td>
<td>no plant</td>
<td>absent</td>
</tr>
<tr>
<td>3 upright frame</td>
<td>horizontal frame</td>
<td>spotty jars</td>
<td>white jars</td>
<td>laptop</td>
<td>computer</td>
<td>changed</td>
</tr>
<tr>
<td>4 pink cushion</td>
<td>brown cushion</td>
<td>wine at top</td>
<td>wine at bottom</td>
<td>tiger</td>
<td>platypus</td>
<td>changed</td>
</tr>
<tr>
<td>5 scarf</td>
<td>tangram</td>
<td>flipflops</td>
<td>new object</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 open book</td>
<td>wok</td>
<td>gloves</td>
<td>new object</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 sheepskin rug</td>
<td>bananas</td>
<td>aroma light</td>
<td>new object</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 stack of books</td>
<td>mop</td>
<td>standing books</td>
<td>new object</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendices

Appendix D.  Gaze projection: Example transcript of a speaker description

Number 1 is the kettle, in just the centre, slightly off-centre of the image, just the, in the corner of the sideboard by the books. It's a silver kettle.

Number 2 is the metal extractor fan coming down over the... the stoves.

um... Number 3 is a red mopping bucket, um in the bottom right-hand corner of the image by the washing machine.

Number 4 appears to be a cask of whisky on top of the shelf, next to the extractor fan in the centre-left of the image.

Number 5 looks like a wine rack, which is on the window sill on... in the right-hand side of the image.

Number 6 appears to be a... some sugar or tea or coffee cans... um which are on the sideboard, uh, in the sort of centre-right right-hand side of the image at the back.

Number 7 is like a ladle, uh next to the stove. A big sort of baby-blue ladle next to some fruit and veg.

um... Number 8 is a set of eggs, right in the foreground of the image, right at the bottom right... um... on the table, just in shot.

Number 9 is a... appears to be some sort of box with a red emblem on it. It's just on a shelf underneath the uh baby-blue ladle and broccoli and fruit. It's just below that on a shelf with... it's like a cream box with a red emblem.

Number 10 is a mixing bowl, in the centre, bottom centre of the image, on the table; a clear mixing bowl with some something in it, some dark things in it.

Number 11 is a red saucepan on the stove, on the... sort of slightly off-centre of the image.

And Number 12 is a... like... duster, a yellow duster, um attached to in the centre, sort of the far right of the image, um on one of the cupboards underneath the, by the sink."

(Speaker 4, trial 3, kitchen)

Total duration: 2:16 min
Appendix E.  Gaze projection: Questionnaire

At the end of the experimental session, the following questionnaire was presented on paper to listeners in Experiments 3. Adaptations of it were also used in Experiments 4-7; changes from the original version are indicated in brackets after each question. Questions concerning the speech utterance or the gaze cursor were omitted for conditions where this form of information had not actually been presented.

Open questions were always followed by sufficient blank space for participants to write their reply; rating scales and multiple choice options are indicated.

1. Imagine that someone asked you to describe what you did and what happened in the study you just took part in. What would you tell them?

2. How useful did you find the speaker’s descriptions?

   not at all useful  somewhat useful  extremely useful
   1                2                3                4                5                6                7

3. How clear did you find the speaker’s descriptions?

   not at all clear  somewhat clear  extremely clear
   1                2                3                4                5                6                7

4. How useful did you find the green dot?

   not at all useful  somewhat useful  extremely useful
   1                2                3                4                5                6                7
[Additional question for Exp. 5, cursor-only condition:] How clear did you find the green dot?

not at all clear | somewhat clear | extremely clear
---|---|---
1 | 2 | 3 | 4 | 5 | 6 | 7

[Additional question for Exp. 5, speech-only and cursor-only conditions:] How helpful do you think it would have been to see where the speaker was looking/...to hear what the speaker was saying?

not at all helpful | somewhat helpful | extremely helpful
---|---|---
1 | 2 | 3 | 4 | 5 | 6 | 7

5. How tightly would you say the dot was linked to what the speaker was saying? [alternatively, for Exp. 5, speech-only condition:] How tightly do you think the speaker’s eye movements would be linked to what they were saying?

not linked at all | somewhat linked | very strongly linked
---|---|---
1 | 2 | 3 | 4 | 5 | 6 | 7

6. To what extent did you ignore the dot and concentrate on the speech?

didn’t ignore dot at all | sometimes ignored dot | ignored dot completely
---|---|---
1 | 2 | 3 | 4 | 5 | 6 | 7

7. To what extent did you try to follow the dot?

didn’t follow dot at all | sometimes followed dot | followed dot all the time
---|---|---
1 | 2 | 3 | 4 | 5 | 6 | 7

8. To what extent did you find the dot distracting from your task?

not at all distracting | somewhat distracting | extremely distracting
---|---|---
1 | 2 | 3 | 4 | 5 | 6 | 7
[Additional questions for Exp. 6, recall task:]

The next questions refer only to the second part of the study, where you tried to remember the objects you had clicked.

How useful did you find the green dot in the memory task?

<table>
<thead>
<tr>
<th>not at all useful</th>
<th>somewhat useful</th>
<th>extremely useful</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

To what extent did you ignore the dot in the memory task?

<table>
<thead>
<tr>
<th>didn’t ignore dot at all</th>
<th>sometimes ignored dot</th>
<th>ignored dot completely</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

To what extent did you try to follow the dot in the memory task?

<table>
<thead>
<tr>
<th>didn’t follow dot at all</th>
<th>sometimes followed dot</th>
<th>followed dot all the time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

To what extent did you find the dot distracting from the memory task?

<table>
<thead>
<tr>
<th>not at all distracting</th>
<th>somewhat distracting</th>
<th>extremely distracting</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

How tightly would you say the dot in the memory task was linked to what the speaker had said in the clicking task?

<table>
<thead>
<tr>
<th>not linked at all</th>
<th>somewhat linked</th>
<th>very strongly linked</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

How similar would you say the dot in the memory task was to the dot in the clicking task?

<table>
<thead>
<tr>
<th>completely different</th>
<th>somewhat similar</th>
<th>identical</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

9. What do you think this study was investigating?
10. Any further comments on any aspect of the experiment? [Question omitted for Exp. 5 - 7]

[Additional Question for Exp. 6 and 7]

In general, how do you think a speaker’s eye movements would be related to what they are saying when describing objects? (N.B. Don’t think too much about the dot you saw, as it did not always reflect the speaker’s true eye movements).

11. [Experiment 3:] Actually, this experiment compared two different “dot” conditions, which did not always correspond to what the “dot” was said to be in the instructions.

In your case, do you think the dot was…

a) … recorded when the speaker was remembering the object descriptions (i.e. practice 2)? □

b) … recorded when the speaker was actually making the object descriptions (i.e. practice 1)? □

[Experiment 6:] Actually, this experiment compared four different “dot” conditions. For some participants, the dot is played back exactly as it was recorded while the speaker spoke (we call this “on time”). For other participants, the dot is 5 seconds ahead of the speech (“fast”), or 5 seconds behind the speech (“slow”). For a final set of participants it is completely disconnected (“random”).

Please try to guess which condition you were in:

Clicking task:

   on time □   fast □   slow □   random □

Memory task:

   on time □   random □
[Experiment 7 (and 5, speech & cursor condition):] Actually, this experiment compared five different “dot” conditions. For some participants, the dot is played back exactly as it was recorded while the speaker spoke (we call this “on time”). For other participants, the dot is 1 second ahead of the speech (“1s fast”), or 2 seconds ahead (“2s fast”). Finally, some participants see a dot which is 1 second behind the speech (“1s slow”), or 2 seconds slow (“2s slow”).

Please try to guess which condition you were in:

- on time
- 1s fast
- 2s fast
- 1s slow
- 2s slow

12. Why? (Please provide a short explanation of your answer in 11.)

Finally, please provide a few items of demographic information:

13. Gender: female □ male □

14. Handedness: left □ right □ ambidextrous □

15. Age (in years): ____

16. Which part of the country did you grow up in?

17. Main occupation (if student, what subject?):

18. How many eyetracking studies have you participated in previously? (please tick one)

- none
- one
- 2-4
- 5-10
- > 10

Thanks again for your participation!
Appendices

Appendix F.  Gaze projection: Recognition task

Following Experiment 4, the following multiple-choice items were included in the post-experimental questionnaire as a pilot study for a future experiment:

Now please think about what the speaker actually said. The following is a list of words which could be used to describe the objects in the three rooms. Please put a tick next to the word your speaker used for each object. If you think they used a completely different expression, please tick “other” and write it down. Don’t worry if you can’t remember - just guess!

Living room:

☐ cafetiére  ☐ coffee pot  ☐ coffee jug  ☐ coffee percolator  ☐ other:

☐ candles  ☐ tray of candles  ☐ tea-lights  ☐ row of candles  ☐ other:

☐ cushion  ☐ pillow  ☐ other:

☐ deer-like thing  ☐ toy reindeer  ☐ ornament  ☐ ram  ☐ horse toy  ☐ giraffe  ☐ other:

☐ bunch of roses  ☐ roses  ☐ vase  ☐ flowers  ☐ other:

☐ hifi system  ☐ cd-player  ☐ stereo  ☐ box  ☐ square  ☐ hifi  ☐ other:

☐ magazine  ☐ book  ☐ other:

☐ notebook  ☐ notepad  ☐ rectangular shape  ☐ other:

☐ cordless phone  ☐ telephone  ☐ phone  ☐ other:

☐ picture frame  ☐ photo frame  ☐ photo  ☐ picture  ☐ other:

☐ bunch of twigs  ☐ bundle of twigs  ☐ bunch of sticks  ☐ straw  ☐ wicker things  ☐ rushes  ☐ other:
**Kitchen:**

- **bowl**
- **mixing bowl**
- **other:**

- **mopping bucket**
- **bucket**
- **mop bucket**
- **other:**

- **duster**
- **cloth**
- **tea-towel**
- **other:**

- **box of eggs**
- **eggs**
- **tray of eggs**
- **set of eggs**
- **other:**

- **cooker-hood**
- **kitchen extractor**
- **extractor fan**
- **metal fan**
- **other:**

- **ladle**
- **spoon**
- **piece of cutlery**
- **other:**

- **pan**
- **pot**
- **saucepan**
- **other:**

- **biscuit tins**
- **canisters**
- **cans**
- **jars**
- **tin pots**
- **tins**
- **other:**

- **bottles of whisky**
- **bottle of whisky**
- **cask of whisky**
- **whisky bottle**
- **whisky**
- **other:**

- **wine rack**
- **bottle of wine**
- **wine holder**
- **other:**
Bedroom:

- waste paper bin
- bin
- vase
- rubbish bin
- waste paper basket
- cactus
- cactus plant
- cacti pot
- pot with a cactus
- plant pot
- candle
- other:

- smoke alarm
- smoke detector
- fire alarm
- jar
- alarm
- other:

- laptop
- television
- computer
- computer monitor
- tv
- television screen
- other:

- pillow
- cushion
- other:

- cd-player
- radio
- stereo
- other:

- tiger cuddly toy
- tiger teddy-bear
- toy tiger
- tiger-toy
- tiger
- stuffed tiger
- other:

- box of tissues
- tissue box
- set of tissues
- other:

- shirt
- blouse
- t-shirt
- other:

- palm
- plant
- palm tree
- other:
Appendix G.  Gaze projection: Recall task

As mentioned in Section 8.1.2.1, the main clicking task of Experiment 6 was followed by a surprise recall test, similar to the one conducted after Experiment 2. There, some participants had struggled to recall any items at all, irrespective of the cursor condition they had been in. In the current study, the gaze cursor therefore reappeared on the blank screen to serve as a memory cue (Kent & Lambert, 2008).

Design, Predictions, and Procedure

The main purpose of the recall task was to discover whether seeing a helpful cursor during encoding (i.e. in the main experiment) also made it easier to recall the encoded items later. The number of recalled items was therefore compared between participants who had seen an informative (i.e. natural, ahead, or delayed) gaze cursor and those who had seen a random dot – for the purposes of this analysis, participants who had seen one of the shifted cursors in the main experiment were treated as if they had seen a natural cursor (the timing of when exactly they looked at the object was not expected to have an effect).

During the recall task itself, half of the participants across conditions were presented with a helpful cursor, which used exactly the same movement file as the natural gaze cursor in the main experiment. The other half saw the random dot. Depending on the type of cursor (‘helpful’ vs. random) they had been seen in the main experiment, this created a 2 × 2 design for the recall task; either the new cursor was the same as the one seen in the first task, or different. To put it differently: of the six participants in each encoding condition (natural, ahead, delayed, or random in the main experiment), three were shown a natural cursor in front of a blank screen in the recall task, while the other three saw a random recall cursor. In this way, twelve participants overall (three per condition) experienced the same cursor during encoding and during recall, while the other twelve saw the other kind of cursor. This made it possible to test whether any improvement in recall from seeing the gaze cue was due to the fact that it directed the participant’s attention back to the
original location of the item (natural cursor) or to the original location of encoding (identical cursors during encoding and recall).

All participants who saw the natural recall cursor were expected to benefit from it, as it always highlighted the locations where the objects that they were trying to recall had been. Participants who encoded with a random cursor in the main experiment might also be expected to find this recall condition helpful, if their memory for the item contained a spatial index of its actual location. In contrast, a random recall cursor could only potentially benefit participants who had seen the same random cursor in the main experiment, if their memory for the items was indexed with regard to where they had first heard about them (Spivey & Geng, 2001; Spivey et al., 2004; F. Ferreira et al., 2008; Hoover & Richardson, 2008). Finally, if this kind of location-based encoding proved helpful, participants who saw a natural cursor in both tasks should receive an extra memory boost – they would be seeing a recall cursor indexing both where the object had been, as well as where they had been looking at the time of its description.

For all participants, the order in which they were asked to recall objects from the three rooms was the same during the recall task as it had been in the main experiment. Before each trial, a one-word reminder of the room in which they were trying to recall objects appeared on-screen for 500 ms. The screen then went blank, apart from the cursor. Each recall trial lasted a maximum of three minutes, or until the participant pressed a button to indicate that he or she could think of no more objects.

**Results and Discussion**

As in Experiment 2, recall accuracy was measured by comparing the number of correctly named objects between conditions. Despite the fact that this time the gaze cursor had been available as a memory prompt, the number of items recalled still showed large individual differences, varying from three to eleven items recalled (out of twelve) per trial. On average, participants recalled 23.79 out of the total of 36 items (66% ; SD = 4.53), two more than in Experiment 2 (M = 21.5, SD = 4.54).

34 Note that participants in Experiment 1 heard descriptions by six different speakers, whereas the current study just used one. Any difference between the studies could therefore also have been caused by Speaker 4 choosing particularly memorable
Although this suggests that the gaze cursor may have proved somewhat helpful, the difference between experiments is not significant ($t(34) = -1.43, p = .16$). The following table provides an overview of the number of items correctly recalled after Experiment 6, per combination of encoding and recall condition.

<table>
<thead>
<tr>
<th>Encoding cursor</th>
<th>Recall cursor</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>natural</td>
<td>M 22.67 SD 7.51</td>
<td>M 20.67 SD 4.16</td>
</tr>
<tr>
<td>5s ahead</td>
<td>M 21.00 SD 6.08</td>
<td>M 28.00 SD 0.00</td>
</tr>
<tr>
<td>5s delayed</td>
<td>M 25.33 SD 4.16</td>
<td>M 21.67 SD 2.89</td>
</tr>
<tr>
<td>random</td>
<td>M 26.67 SD 2.08</td>
<td>M 24.33 SD 4.16</td>
</tr>
<tr>
<td>Total</td>
<td>M 23.92 SD 5.13</td>
<td>M 23.67 SD 4.08</td>
</tr>
</tbody>
</table>

The main question of whether the usefulness of the gaze cursor seen during encoding and/or during recall affected the number of items recalled was addressed in an analysis of variance with the within-participants factor TRIAL (3 levels), and the between-participants factors ENCODE (2 levels: helpful vs. random) and RECALL (2 levels: natural vs. random). It showed a main effect of TRIAL ($F(2, 40) = 6.00, p < .01$), with more items recalled from later than from earlier rooms (room 1: $M = 6.83$, room 2: $M = 8.08$; room 3: $M = 8.67$). There was also a marginal interaction of TRIAL and ENCODE ($F(2, 40) = 2.92, p = .065$), due to a greater increase in number of items recalled if participants had seen a random cursor during encoding. However, none of the effects of main interest turned out to be significant (all $p > .3$): Overall, participants recalled no better if they had seen a helpful dot during encoding than if it had been random, and the type of cursor seen during recall had no effect either. Most importantly, there was no interaction; whether the cursor was the same or different during encoding and recall seems to have been irrelevant. Similar analyses with four levels of ENCODE (corresponding to the four conditions in the main descriptions. In fact however, the two participants who heard his descriptions in Experiment 1 recalled substantially fewer items (cursor: 18, no cursor: 16) than the average of 21.5 items.
experiment) or with a single combined match vs. mismatch factor showed the same pattern of effects.

In fact, several findings indicate that participants did not make much use of the cursor in the recall task, although it could potentially have been extremely helpful. The lack of significant improvement in recall performance compared to Experiment 2, where there was no cursor, has already been mentioned. In addition, the means in the table above show no recall boost for participants who saw the natural cursor in both tasks. This would have been expected if participants used the locations pinpointed by the informative cursor as a hint to the object that had been there. Finally, two results from the section of the post-experimental questionnaire dealing with the recall task confirm that participants were rarely even aware of the potential usefulness of the gaze cursor: while cursors during the encoding and recall tasks were rated as highly similar by participants who saw the identical cursor twice ($M = 5.33$ on a 7-point scale, $SD = 1.72$), the mean rating from participants for whom they were radically different was not significantly lower ($M = 4.5$, $SD = 1.45$; $p = .21$). Even more strikingly, there was practically no difference between conditions in mean ratings of the cursor’s usefulness: $M = 2.67$ ($SD = 1.50$), and $M = 2.42$ ($SD = 1.56$), for matching and mismatching conditions, respectively ($p = .69$).³⁵

Clearly, the recall task again failed to differentiate between cursor conditions. As in Experiment 2, likely reasons are the large differences in participants’ memory ability and motivation, as well as their apparent lack of awareness of the fact that it might help them if they followed the gaze cursor.

---

³⁵ The answers to the other recall questions showed a similar lack of effects and will therefore not be discussed further.