

Beyond words: Non-linguistic signals and the recovery of meaning

Josiah P. J. King



THE UNIVERSITY
of EDINBURGH

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

to

School of Philosophy, Psychology and Language Sciences

University of Edinburgh

2019

Declaration

I declare that this thesis has been composed solely by myself and that it has not been submitted, either in whole or in part, in any previous application for a degree. Except where otherwise acknowledged, the work presented is entirely my own.

Josiah King

7 September 2019

Abstract

Beyond the words a speaker produces, meaning can be recovered from the many ways in which the words are delivered. During everyday discourse, the way in which a speaker produces an utterance — both in their spoken delivery and accompanying movements — is an important part of the communication process. This thesis investigates the ways in which meaning is carried within these non-linguistic behaviours, focussing on how they may provide signals about upcoming message content and about speakers intentions.

Previous research has shown that the manner of spoken delivery (for instance, rates of speech, intonation, and fluency) influences listeners content-based expectations as well as pragmatic comprehension. These effects have been evidenced both post-hoc and during the moment-to-moment processing of speech. Considerably less attention has been paid on whether speakers movements and non-verbal behaviours have similar effects: Research on gesture has tended to focus on the content represented by gesture (rather than its potential to signal information about the message and/or speaker).

We focus on two ways in which non-linguistic behaviours influence comprehension: Firstly, as signals of speech planning difficulty (and so of upcoming content), and secondly as indicators of a speakers intention to deceive. The former has been studied in relation to speech disfluency, but not to gestures. The latter has been

studied extensively in relation to many linguistic and non-linguistic behaviours, but has only recently begun to be addressed with respect to the time course of the process.

Focussing on these two areas, we address the broader question about the perceptual relevance of non-linguistic signals in comprehension through a dialogue study and a series of comprehension experiments combining eye and mouse tracking techniques. Extending the Visual World Paradigm—commonly used in comprehension studies—to include a video component showing the speaker, we measure listeners' eye movements and mouse coordinates as they select objects in the on-screen display. By directly manipulating the presence of different non-linguistic behaviours, we investigate whether and when these behaviours are interpreted as signals of 1) upcoming difficulty in speech, and 2) the speaker's intention to deceive.

Our results demonstrate that, like for speech disfluency, listeners interpret the presence of representational gesturing to inform explicit predictions about upcoming referents. This follows from the findings from our dialogue study which suggest that speakers produce more of this type of gesturing relative to speech when describing shapes which are more conceptually difficult. We also show that listeners reliably perceive certain motoric behaviours of a speaker to indicate deception, and that the influence these behavioural cues have on listeners' interpretation can be detected alongside the unfolding linguistic input.

Non-linguistic cues to perceived deception are found to be robust to contexts where speakers produce a variety of cues in different modalities. However, findings point to differences in how listeners link visual and spoken cues with deception. Results indicate that when cues are present in both modalities, visual cues tend to drive listeners' biases towards interpreting an utterance as dishonest. The time course supports a view in which non-linguistic cues influence pragmatic comprehension

at an early stage, and suggests that linking visual cues with deception may be more resource demanding than it is for spoken cues.

Listeners' associations between the presence of certain non-linguistic behaviours and judgements of deception also hold in situations where there are other alternative explanations for a given cue. However, the availability of an alternative explanations for a given non-linguistic behaviour are found to influence early stages of comprehension, suggesting that listeners may engage in dynamic reasoning about the possible causes of a speaker's manner of delivery. Taken together, the results from studies presented in this thesis highlight the role of both the spoken and visual delivery of an utterance in shaping comprehension, highlighting the fact that communication is fundamentally multi-modal.

Lay Summary

During communication, speech varies with respect to how it is delivered. Speakers can change the spoken delivery of words as well as the movements they make while speaking. This thesis explores how this behaviour forms an important part of communication, and focuses on two examples of how meaning may be recovered from speakers' varying behaviour: As signals about the act of producing speech and about the speaker's intentions.

In Part I of the thesis, we look at how the act of gesturing relates to the difficulty of speaking, and approach this from both the perspective of the speaker and of the listener. In Part II, we explore if and when various behaviours are perceived as signals of the intention to deceive.

The thesis presents a series of experiments which measure participants' eye movements and mouse coordinates as they select between on-screen objects whilst hearing and viewing recorded utterances from a speaker. By varying the behaviours (e.g., fidgeting, or saying "um") in these recordings, we investigate how the delivery of language influences the comprehension of meaning.

The results of Part I demonstrate that speakers produce longer gestures (relative to speech) when speaking is more difficult, and suggest that, in turn, listeners' rely on the presence of gesture to inform guesses about what a speaker is about to describe. However, findings indicate that listeners' uptake of gesture may be

optional, used in contexts where it matters (e.g., if listeners are guessing based on fragments of descriptions) but perhaps not in contexts where subsequent speech is expected to be more useful in interpreting meaning.

In Part II, we show that certain behaviours produced by a speaker are perceived by listeners to signal that the speaker might be lying. Furthermore, we show that listeners pick up on these behaviours very quickly, when they are both auditory (an “um”) or visual (fidgeting).

The results emphasise that beyond the words used during communication, a myriad of other aspects of the way in which language is delivered can influence comprehension, highlighting the multi-modal nature of communication.

Acknowledgements

Foremost, I would like to thank Martin Corley and Hannah Rohde for their encouragement, supervision and feedback throughout. I am very grateful to Jia Loy for providing the basis on which many of the experiments in the second half of this thesis are built (and for all the discussions, feedback, and everything else she's helped with), and to Caroline McHutchison for all round general support as well as agreeing to sit in front of a camera and produce so many awkward gestures. I would also like to acknowledge the University of Edinburgh, School of Philosophy, Psychology and Language Sciences for a PhD scholarship (and a number of research grants) which helped to fund the research in this thesis.

Contents

Declaration	iii
Abstract	v
Lay Summary	ix
Acknowledgements	xi
1 Thesis overview	1
2 Background	5
2.1 Non-verbal behaviour	5
2.1.1 Kendon's Continuum	6
2.1.2 Informative vs. Communicative	9
2.2 Why people vary their non-verbal behaviour during speaking	12
2.2.1 Intentionally communicative acts	13
2.2.2 Facilitating speech production	16
2.2.3 Leakage	20
2.3 Comprehension of multi-modal language	22
2.3.1 Comprehension of gestural content	24
2.3.2 Representational gesturing vs. meta-communicative signalling	32
2.4 Methodology	37
2.4.1 Eye-tracking, mouse-tracking and the Visual World Paradigm	38
2.4.2 The current thesis	43
I Signals of speech production difficulty	45
3 Conceptual difficulty and production of speech and gesture	47
3.1 The relationship of speech and gesture	49
3.2 Experiment 3.1	54
3.2.1 Participants	55
3.2.2 Method	56
3.2.3 Coding	61
3.2.4 Results	63

3.2.5	Discussion	67
3.3	Additional exploratory research	70
3.4	Chapter discussion	83
4	Gesture as a signal of conceptual demand	87
4.1	Signals of speech planning difficulty	89
4.2	Experiment 4.1	95
4.2.1	Method	96
4.2.2	Results	105
4.2.3	Discussion	113
4.3	Experiment 4.2	119
4.3.1	Method	119
4.3.2	Results	120
4.4	Discussion	124
5	Gesture modulating listener expectations in real time	131
5.1	Signals guiding on-line expectations	133
5.2	Experiment 5.1	137
5.2.1	Method	138
5.2.2	Results	141
5.2.3	Pre-disambiguation	144
5.2.4	Post-disambiguation	148
5.3	Discussion	155
5.3.1	Pre-disambiguation	156
5.3.2	Post-disambiguation	157
5.3.3	Eye-tracking and gestures	158
5.4	Does gesturing guide listeners' predictions of upcoming message content?	159
II	Markers of deception	163
6	Gesturing informs pragmatic judgements: Interpreting non-verbal cues to deception in real time	165
6.1	Deception and delivery	167
6.2	Literal vs. pragmatic comprehension	169
6.2.1	Negation	172
6.3	Experiments 6.1 and 6.2	173
6.3.1	Abstract	173
6.3.2	Introduction	174
6.3.3	Experiment 6.1	178
6.3.4	Results	183
6.3.5	Additional Analyses of Filler trials	186
6.3.6	Discussion	191
6.3.7	Experiment 6.2	194

6.3.8	Results	197
6.3.9	General discussion	201
6.4	Chapter discussion	205
7	Competing cues: Gestural vs disfluent signals to deception	209
7.1	Perception of deception in different modalities	212
7.2	Experiment 7.1	215
7.2.1	Participants	215
7.2.2	Materials	216
7.2.3	Results	222
7.2.4	Discussion	231
8	Competing causes: Contextual effects on online pragmatic inferences of deception	237
8.1	Experiment 8.1	241
8.1.1	Abstract	241
8.1.2	Introduction	241
8.1.3	Method	247
8.1.4	Results	252
8.1.5	Discussion	256
8.2	Chapter discussion	262
III	Conclusions	265
9	General discussion	267
9.1	Signals of conceptual demand	268
9.1.1	The time course	269
9.2	Markers of deception	271
9.3	Methodological considerations	273
9.4	Conclusions	274
A	Model results for Experiments 6.1 and 6.2	277
B	Model results for Experiment 8.1	283
C	Replication of Kelly, Özyürek, and Maris (2010)	285
C.1	Experiment C.1	285
C.1.1	Method	286
C.1.2	Analysis	288
C.1.3	Results	288
C.1.4	Discussion	291
References		293

List of Tables

2.1	Example stimuli, Kelly et al. (2010)	28
2.2	Meta-communication, Clark (1996)	33
3.1	Experiment 3.1: Speaker familiarity with shapes across experimental blocks	64
3.2	Experiment 3.1: Gesture-duration model results	68
3.3	Experiment 3.1: Holds, strokes, and numbers of hands used in iconic gestures	73
3.4	Experiment 3.1: Occurrence of different gestures	76
3.5	Experiment 3.1: Model results for occurrence of other types of gestures	77
3.6	Experiment 3.1: Speech disfluency model results	78
3.7	Experiment 3.1: Gesture fluency	80
3.8	Experiment 3.1: Gestural false starts model results	80
3.9	Experiment 3.1: Gesture-onset model results	82
4.1	Experiment 4.1: Breakdown and examples of filler utterances	98
4.2	Experiment 4.1: Breakdown of mouse clicks on each object	109
4.3	Experiment 4.1: Mouse click model results	109
4.4	Experiment 4.1: Time-to-click model results	110
4.5	Experiment 4.1: Eye- and mouse-tracking model results	113
4.6	Experiment 4.2: Breakdown of mouse clicks on each object	122
4.7	Experiment 4.2: Mouse click model results	122
4.8	Experiment 4.2: Time-to-click model results	123
4.9	Experiment 4.2: Eye- and mouse-tracking model results	126
5.1	Experiment 5.1: Eye- and mouse-tracking model results, pre-disambiguation window	147
5.2	Experiment 5.1: Eye- and mouse-tracking model results, post-disambiguation window	151
5.3	Experiment 5.1: Breakdown of mouse clicks on each object	154
5.4	Experiment 5.1: Mouse click model results	154
5.5	Experiment 5.1: Time-to-click model results	155
6.1	Experiment 6.1: Breakdown of mouse clicks on each object in critical trials	185

6.2	Experiment 6.1: Breakdown of mouse clicks on each object in filler trials	189
6.3	Experiment 6.2: Breakdown of mouse clicks on each object	199
7.1	Experiment 7.1: Breakdown and examples of filler utterances	217
7.2	Experiment 7.1: Breakdown and examples of filler videos	218
7.3	Experiment 7.1: Breakdown of mouse clicks on each object and time taken to click	225
7.4	Experiment 7.1: Mouse click model results	226
7.5	Experiment 7.1: Mouse click pairwise comparisons of estimated marginal means	226
7.6	Experiment 7.1: Time-to-click model results	227
7.7	Experiment 7.1: Eye- and mouse-tracking model results	229
8.1	Experiment 8.1: Breakdown and examples of filler utterances	249
8.2	Experiment 8.1: Breakdown and examples of filler background noises	250
8.3	Experiment 8.1: Breakdown of mouse clicks on each object	254
A.1	Experiment 6.1: Mouse click model results, critical trials	277
A.2	Experiment 6.1: Time-to-click model results, critical trials	278
A.3	Experiment 6.1: Eye- and mouse-tracking model results, critical trials	278
A.4	Experiment 6.1: Mouse click model results, filler trials	279
A.5	Experiment 6.1: Time-to-click model results, filler trials	279
A.6	Experiment 6.1: Eye- and mouse-tracking model results, filler trials	280
A.7	Experiment 6.2: Mouse click model results	280
A.8	Experiment 6.2: Time-to-click model results	281
A.9	Experiment 6.2: Eye- and mouse-tracking model results	281
B.1	Experiment 8.1: Mouse click model results	283
B.2	Experiment 8.1: Eye- and mouse-tracking model results	284

List of Figures

2.1	Kendon's Continuum	8
2.2	Example non-verbal behaviours	10
3.1	Experiment 3.1: Example of shapes used in critical trials	56
3.2	Experiment 3.1: Participant set-up	58
3.3	Experiment 3.1: Timeline of a sample trial	60
3.4	Experiment 3.1: Durations of speech and gesture	66
3.5	Experiment 3.1: Durations of speech and gesture by familiarity of shape	67
3.6	Experiment 3.1: Example gestures	74
3.7	False-starts in gestures	79
4.1	Experiment 4.1: Example of easy-to-name and difficult-to-name shapes	97
4.2	Experiment 4.1: Example gestures	99
4.3	Experiment 4.1: Timeline of a sample trial	104
4.4	Experiment 4.1: Eye movements over time	111
4.5	Experiment 4.1: Empirical logit transformed fixation bias	112
4.6	Experiment 4.1: Mouse movements over time	114
4.7	Experiment 4.2: Timeline of a sample trial	120
4.8	Experiment 4.2: Eye movements over time	124
4.9	Experiment 4.2: Empirical logit transformed fixation bias	125
4.10	Experiment 4.2: Mouse movements over time	127
5.1	Experiment 5.1: Timeline of a sample trial	142
5.2	Experiment 5.1: Eye movements over time for critical trials that name easy-to-name targets	145
5.3	Experiment 5.1: Eye movements over time for critical trials that name difficult-to-name targets	146
5.4	Experiment 5.1: Mouse movements over time for critical trials that name easy-to-name targets	148
5.5	Experiment 5.1: Mouse movements over time for critical trials that name difficult-to-name targets	149
5.6	Experiment 5.1: Empirical logit transformed fixation bias	152
5.7	Experiment 5.1: Empirical logit transformed mouse movement bias	153

6.1	Experiments 6.1 and 6.2: Example display	177
6.2	Experiments 6.1 and 6.2: Timeline of a sample trial	182
6.3	Experiment 6.1: Eye movements over time, critical trials	187
6.4	Experiment 6.1: Mouse movements over time, critical trials	188
6.5	Experiment 6.1: Eye movements over time, filler trials	190
6.6	Experiment 6.1: Mouse movements over time, filler trials	192
6.7	Experiment 6.2: Adaptor gestures used in critical videos	196
6.8	Experiment 6.2: Eye movements over time	200
6.9	Experiment 6.2: Mouse movements over time	201
7.1	Experiment 7.1: Timeline of a sample trial	221
7.2	Experiment 7.1: Eye movements over time	228
7.3	Experiment 7.1: Empirical logit transformed fixation and mouse movement biases	230
7.4	Experiment 7.1: Mouse movements over time	231
8.1	Experiment 8.1: Eye movements over time, fluent utterances . . .	256
8.2	Experiment 8.1: Eye movements over time, disfluent utterances .	257
8.3	Experiment 8.1: Mouse movements over time, fluent utterances .	258
8.4	Experiment 8.1: Mouse movements over time, disfluent utterances	259
C.1	Experiment C.1: Timeline of a sample trial	287
C.2	Experiment C.1: Reaction times by condition	289
C.3	Experiment C.1: Error rates by condition	290

Chapter 1

Thesis overview

Communication is more than just words. Language use varies in both how speech itself is delivered and in the visible actions which accompany it. As listeners,¹ humans effortlessly make sense of information conveyed in multiple channels to understand not only the semantic content of the words but also the intended meanings or goals of a speaker. Everyday language use occurs at a rapid pace, with approximately 3 to 5 words produced per second (Picheny, Durlach, & Braida, 1986), with disfluencies in both speech (Shriberg, 2001) and gesture (Esposito, McCullough, & Quek, 2001). The movements which accompany speech are produced for a broad range of reasons, with some movements carrying information fundamental to correctly understanding the message being communicated, and others being merely incidental (see, e.g., Ekman & Friesen, 1969; McNeill, 1992). Comprehending meaning from such a complex and varied input is achieved rapidly, suggesting that sophisticated mechanisms manage this process.

Broadly speaking, this thesis asks how meaning is recovered by listeners from the

¹In this thesis the terms ‘listener’ and ‘addressee’ (often used in gesture research) are used interchangeably

non-linguistic behaviour of a speaker. Focussing predominantly on the non-verbal behaviours produced alongside speech, we explore the ways in which non-linguistic signals aid comprehension of both the propositional content of speech and of the intended meanings and goals of the speaker. Drawing parallels with research on the manner of spoken delivery, particularly that of speech disfluency, we investigate the extent to which non-verbal behaviours are interpreted as signals of speech planning difficulty (Part I) and of a speaker's intentions, specifically the intention to deceive (Part II).

In Part I, Chapters 3 to 5 investigate the relationship of gesturing with the conceptual demands required to formulate descriptions of objects. In Chapter 3, we ask whether the production of gestures (focussing on iconic gesturing) relative to the production of speech varies according to the nameability of the object being described. We then ask a reverse of this question in comprehension: Chapter 4 explores whether listeners interpret speakers' production of different non-verbal behaviours (specifically iconic gesturing and self-adaptive movements) as signals of the nameability of upcoming referents. Chapter 5 provides an investigation into whether this association guides listeners' on-line expectations alongside the unfolding speech-gesture input. Results suggest that listeners reliably associate the presence of iconic gesturing (the content of which provides no disambiguating information), but not self-adaptive movements, with shapes which are more difficult-to-name, at least in their off-line responses.

With Part I studying how speakers' non-verbal behaviours may be interpreted as signals relating to the semantic content of a message, Part II investigates non-linguistic behaviours as perceived signals of speakers' intentions. Building on the lessons learnt in Part I, we move from listeners' predictions about upcoming semantic content—which are short-lived when studied in real-time—to situations in which a speaker's non-verbal behaviour influences listeners' global interpretation of an utterance (specifically whether it is a truth or a lie). Chapter 6

provides an overview of the research to date on non-verbal signals of deception, before presenting two experiments which investigate the time course of listeners' judgements of deception based on manner of non-verbal delivery. Results show that speakers' non-verbal behaviours can have a rapid and direct influence on listeners' judgements of deception, and we subsequently extend this investigation to situations in which listeners are faced with different non-linguistic signals presented in different modalities (Chapter 7) and to situations in which there are multiple possible explanations of why a speaker may produce a signal (Chapter 8). Findings suggest that listeners' judgements of deception based on manner of non-verbal delivery appear to take longer to establish than those based on manner of spoken delivery, although the influence of both aspects of delivery are detectable during early moments of comprehension. Additionally, the results of Chapter 8 suggest that the effects of manner of spoken delivery (specifically speech disfluency) on comprehension may be underpinned by flexible and rapid reasoning about the possible causes of a specific behaviour in the given context.

Chapter 2

Background

This chapter reviews the literature on the non-verbal behaviours that accompany speech. Section 2.1 provides an introduction to the different types of non-verbal behaviours that speakers produce, with Section 2.2 discussing the theories of why they do so. Section 2.3 then provides an overview of the literature into comprehension of speech with gesture, with a focus on studies of *how* and *when* information in the two modalities is integrated during comprehension.

2.1 Non-verbal behaviour

During communication, speakers may produce a multiplicity of motor actions. Beyond those involved in articulating speech, speakers may produce movements which are intended to communicate information to the addressee, movements which are performed self-adaptively (e.g., scratching an itch), and movements which are functional interactions with objects irrelevant to the discourse (e.g., picking up a glass). Many of these behaviours provide information about the speaker and the production of speech, information which is readily available to listeners, and

which is often informative and related to the message being communicated. The broad range of non-verbal behaviours produced during communication have been divided up in numerous ways—for instance by the relation of their meaning to speech, by the intention behind their production, or by the body part with which they are presented.

Here, we present two ways of characterising non-verbal behaviour. We first outline a traditional view of movements which are considered to be part of the act of communication and are grouped under the term *gestures*—often defined as “any visible action of any body part, when it is used as an utterance, or as part of an utterance” (Kendon, 2004, p.7). Because this view often fails to capture speakers’ use of body language, we then introduce the distinction of between *communicative* and *informative* non-verbal behaviours. Neither Kendon’s definition of gesture nor the distinction between communicative and informative captures the fact that the communicative intention behind speakers’ motor actions does not map directly to their relevance for comprehension. We therefore approach the topic by discussing non-verbal behaviours from the perspectives of both speaker (Section 2.2) and listener (Section 2.3) independently.

2.1.1 Kendon’s Continuum

One approach to a taxonomy of gestures is to view them as occurring within a continuum, across which movements vary in their relationship with spoken language—dubbed *Kendon’s Continuum* by McNeill (1992). Kendon’s Continuum is often perceived as expressing the range from fully linguistic movements to fully non-linguistic co-speech movements: At one end movements function much like a language (i.e., according to conventions and rules), and at the other movements function only to co-express with speech (see Figure 2.1). The former are those hand movements which have specific meanings and are combined according to specific

rules with other movements to convey meaning, for instance in Sign Languages. This type of gesturing is completely independent of speech, replacing it as the primary mode of communication. At the other end of the continuum, *co-speech* movements are those which are produced spontaneously alongside spoken language, and for which their meaning depends upon that of the accompanying speech. These movements complement vocalisation, adding emphasis, coordinating reference, or adding to the content of speech, thereby conveying meaning in various ways: They can establish reference deictically by locating entities and actions in space (e.g., “This one” [points]); illustrate both literally (“The man had a moustache” [gestures a twirly moustache on own face]) and abstractly (“Which do you prefer?” [two hands weighing up options]); as well as containing no meaningful content themselves but functioning to emphasise the content expressed in speech (e.g., “I did not have sexual relations with that woman” [hand beats time with onset of specific syllables]).

Kendon’s Continuum fails to capture a set of non-verbal behaviours which are prevalent during communication and can play an important role in conveying meaningful information: A speaker’s *body language*—their self-adaptive movements, postures, and facial expressions – all offer a means of meta-communication, often varying according to the intended meaning or the emotional state of the speaker (Busso et al., 2004; Gregersen, 2005). Although non-verbal behaviours such as these are not exclusively produced during acts of communication, when accompanying speech they can carry information about both the speech planning processes and the speaker’s emotional and cognitive states, thus offering indicators about the production and intention behind the spoken message. Furthermore, some of these behaviours may be produced specifically to communicate: For instance in accompanying the utterance “very clever” with crossed arms, a raised eyebrow, and a tilt of the head, a speaker may indicate sarcasm, but the same words may be uttered in earnest and accompanied by a neutral expression with a hand on the

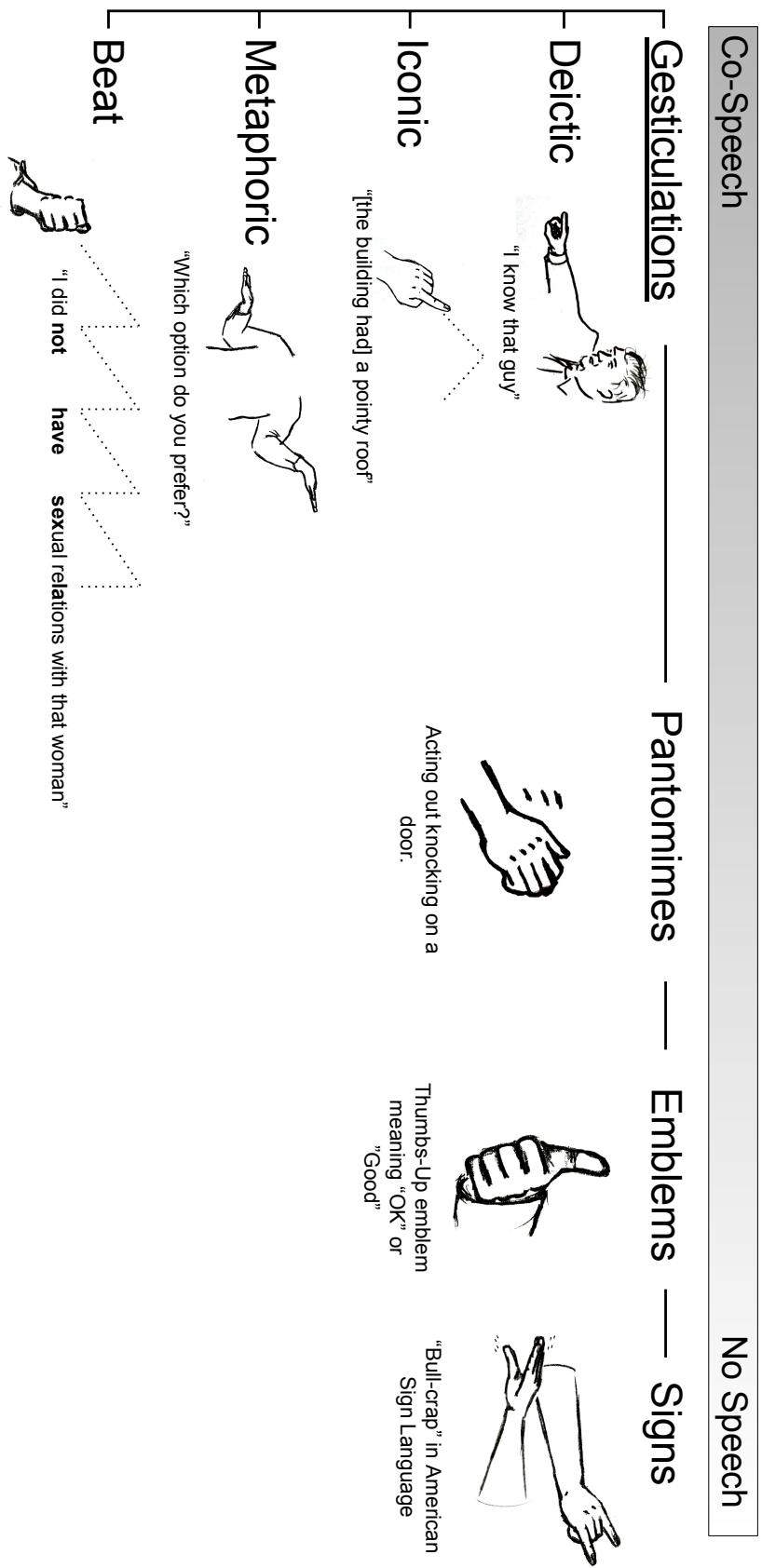


Figure 2.1: Visual representation of Kendon's Continuum: A taxonomy of gestures ranging from conventionalised sign language to co-speech gesticulations. Adapted from McNeill (2000), p.1-10.

chest (see Figure 2.2). These behaviours appear to be captured under Kendon’s definition of gestures as visible actions used as part of an utterance, but are often overlooked in favour of more stereotypically gestural hand movements.

2.1.2 Informative vs. Communicative

A possible reason for *body language* being distinguished from *gestures* in the traditional sense (i.e., those in Kendon’s Continuum) is due to non-verbal behaviours being characterized as either *communicative* or merely *informative* (Lyons, 1977). Many of our motor actions may be informative to interlocutors, without being intended to communicate. For example, in picking up a glass of water, the action indicates to the world that we are thirsty and intend to take a drink.

The distinction between communicative and informative motor actions has often been drawn between traditional gestures and body language, and much research in speech and gesture has typically been concerned with those behaviours which occur exclusively within acts of communication—i.e., focussing on a speaker’s hand movements which visually convey meaning, and less on their body language such as facial expressions or self-adaptive movements which might nonetheless reflect a speaker’s cognitive and affective states. However, this view fails to capture that the use of body language often performs an intentionally communicative function in that it may be deliberately produced in order to alter the meaning of accompanying speech (see Figure 2.2, above). Similarly, many movements which convey meaning transparently may not be intended to communicate to the listener, instead being produced for some speaker-internal reason. For example, iconic gestures—movements which represent the spatial and kinetic properties of physical, concrete items—are still produced (albeit to a lesser extent) even when the speaker is not visible to the listener (e.g., on the telephone), making it difficult to believe these movements are crucial to the message being conveyed (Alibali,



Figure 2.2: Example of honest (top) vs. sarcastic (bottom) non-verbal behaviours, from Moreno (2011)

Heath, & Myers, 2001; A. A. Cohen, 1977; Krauss, Dushay, Chen, & Rauscher, 1995).

From the speaker's point of view, the possible reasons behind production of gestures and other non-verbal behaviours are numerous. There does not appear to be a clear mapping from these reasons to a taxonomy of gesture such as Kendon's Continuum, nor to their relevance in comprehension. Non-verbal behaviours often considered to be *informative* rather than *communicative* have been suggested to have a greater influence on comprehension than the semantic content of speech. When tasked with guessing the emotions of a speaker, listeners have been found to respond to facial expressions and tone of voice more than the words themselves (a finding which has since been grossly over-generalised to the claim that only 7% of communication is verbal, see Mehrabian & Ferris, 1967).

There are many different ways to slice and dice the broad range of non-verbal behaviours: We can focus on the communicative intentions of speakers in producing non-verbal behaviours, the relevance of the information presented in behaviours to the meaning of accompanying speech, or the particular body part in which they are expressed (e.g., hand movements vs. facial expressions). There appears to be no clear mapping between these ways of categorising non-verbal behaviour. Instead, it is more sensible to consider non-verbal communication from both perspectives (speaker and listener) individually. Before turning to comprehension of non-verbal behaviour in Section 2.3, the following section discusses three theories of why a speaker may vary their non-verbal behaviour during speaking: To communicate; to aid speaking; and as unintentional displays of cognitive and affective states.

2.2 Why people vary their non-verbal behaviour during speaking

As discussed above, the non-verbal behaviours which occur during communication vary in how they combine with speech to convey meaning. Some may function as conventionalised signs (e.g., thumbs up); others may carry meaning only in conjunction with speech (e.g., describing a shape and representing its orientation in gesture but not in speech); and others may convey information more subtly still (such as a raised eyebrow indicating suspicion). In the following sections, we discuss three different explanations of why gestures and non-verbal behaviours vary in the production of language.

Firstly, we discuss the traditional view that people gesture because it provides interlocutors with an alternative modality in which to deliberately convey meaning visually rather than semantically. It is possible to imagine different contexts in which any of these movements (thumbs up, representing a shape in space, raising an eyebrow) may be performed with the intention to communicate. Secondly, we introduce the possibility that oftentimes a speaker's production of gestures might not be intentionally communicative, motivated instead by the benefits it has for the speech production processes. This has typically been studied in relation to how iconic gesturing—movements representing spatial and kinetic information—facilitates production of speech, suggesting that they may play a speaker- (as opposed to listener-) oriented role in language use. Lastly, we discuss the idea of non-verbal *leakage*—visible behaviours as unintentional displays of emotional/affective and cognitive states. There are many ways in which a speaker's non-verbal behaviours may vary without them being aware of it: For example, Vrij, Semin, and Bull (1996) found that participants tended to produce fewer movements during an interview in which they lied compared to one in which they told the truth, yet their subsequent judgements of their own behaviour indicated

that they believed the opposite to be true (i.e., they believed they moved more when they lied).

2.2.1 Intentionally communicative acts

Traditionally, those gestures which convey meaningful content to a listener directly (rather than providing information about the speaker and thereby indirectly about the message) are assumed to be produced intentionally for that purpose—i.e., to communicate. Gestures such as those which locate entities and actions in space (*pointing* or *deictic* gestures), and those which represent both literal and abstract ideas through the placement, motion and shape of the hands (*iconic* and *metaphoric* gestures for literal and abstract meanings/ideas respectively), are the two main types of gesture which tend to be viewed as being produced to communicate. Other names for this latter type include *representational*, *illustrative* and even *lexical* gestures (Hadar, 1989), in virtue of the fact that they share a transparent relationship with the content of accompanying speech.

Observationally, one can think of many such gestures for which the communicative function is obvious: When an utterance of “a shape like this” is accompanied by the speaker tracing a shape in space, the act of gesturing is at least as communicative, if not more so, than the spoken words. In a study highlighting the need to document this sort of gesture in interview transcripts, Broaders and Goldin-Meadow (2010) found evidence suggesting that speakers do produce these sort of movements, and that this influences interlocutors’ responses—e.g., the question “What was she wearing?” when paired with a gesture towards the top of one’s head indicates that the speaker is really asking if the subject was wearing a hat (Broaders & Goldin-Meadow, 2010). These sort of gestures, where the movement communicates information over and above that contained in speech, have been referred to as *complementary* (McNeill, 1992) or *non-redundant* gestures (e.g., Alibali, Evans,

Hostetter, Ryan, & Mainela-Arnold, 2009), and are often used to highlight the performative function of a message (e.g., pointing towards an open window while producing the utterance “It’s cold in here” communicates that the speaker is asking someone to close the window).

As well as conveying information about speakers’ communicative intentions, gestures may also present non-redundant message content in a literal sense—i.e., a speaker’s message may be distributed between speech and gesture. An often-cited example from McNeill (1992) of a non-redundant gesture involves a speaker describing a cartoon segment with an utterance of “She chased him out” accompanied by a gesture of her fist moving up and down in a hitting motion (the cartoon had shown one character chasing another out of a room while hitting them with an umbrella). Without the gesture, the addressee receives no information pertaining to the characters hitting one another. That people produce this sort of non-redundant gesturing is evidenced in Melinger and Levelt’s (2004) study, in which participants were tasked with describing networks of circles of different colours arranged along a path. In analysing the division of information which participants conveyed in either modality during these descriptions, iconic gestures representing spatial relationships were found to be associated with more omissions of spatial information in speech. Melinger and Levelt (2004)’s findings that speakers distribute necessary information uniquely into gesture and not into speech, at face value support a view that some gestures are intended to communicate.

Additional evidence in favour of gestures being intentionally communicative can be found in various studies which have shown that speakers tailor their use of gestures to specific listeners and situations. Speakers produce bigger gestures for situations in which the motivation to communicate clearly is stronger (Hostetter, Alibali, & Schrager, 2011). They also use more illustrative gestures (Alibali et al., 2001) and distribute more information into gestures (Gerwing & Allison, 2011) when communicating with a visible—as opposed to visually occluded—addressee.

This increase in gesturing has been shown to result from a speaker's knowledge of what the addressee sees, rather than as a response to being able to see the addressee (Mol, Krahmer, Maes, & Swerts, 2011), driving home the point that some gestures really are meant to be seen. Similarly, speakers' use of gesture has been shown to accommodate an addressee's knowledge of the spoken content: When retelling stories to new addressees, the rate, precision and size of iconic gesturing is greater than when retelling stories to addressees who have heard it already or to addressees whom the speaker knows to share their knowledge (Galati, 2014; Holler & Stevens, 2007; Jacobs & Garnham, 2007). The fact that speakers adapt their production of iconic gestures according to different audiences and contexts shows that the production of some gestures must at least in part be the result of communicative intent.

The flip-side of these studies, however, is that the production of gestures is not completely attenuated under conditions where it is clearly not a useful modality for the addressee (e.g., We frequently see people gesturing while on the telephone). Many of the gestures which people produce are redundant in one way or another, either in that the content is also expressed in speech (e.g., “a circle”[gestures a circle]), or because the visual channel itself is a redundant mode of communication. Directly contrasting the results of Melinger and Levelt (2004), So, Kita, and Goldin-Meadow (2009) found that, when describing video-taped vignettes, speakers were more likely to identify a referent in gesture when they also identified it in speech. Participants in So et al.'s study did not use gesture to provide additional content to speech, instead using the two modes of communication in parallel. This production of redundant gesturing, while still potentially useful for an addressee, is not essential for communication. The ubiquity of redundant gesturing suggests that more than simply communicative intent may be required to explain why speakers produce such movements. One explanation is that gesturing might not

always be an intentionally communicative act, but may perform a more speech-oriented role, facilitating the production and planning of spoken content. In other words, some gestures may be produced for the speaker, not for the addressee.

2.2.2 Facilitating speech production

In the previous section we discussed the traditional view of gesturing as an intentional act of communication with the aim of conveying information to an interlocutor. This view fails to explain why people produce gestures which are redundant (i.e., add no new information or are not visible to the listener). We now turn to the idea that gesturing may perform a more speaker-oriented role, suggesting that in some cases, the communicative effects (for a listener) of gesturing may simply be an epiphenomenon of the movements that speakers produce in the efforts of formulating speech.

The idea that gestures are a result of the act of producing speech stems from a speculation that redundant gestures may simply reflect “a habit” (A. A. Cohen, 1977, p.277)—we gesture on the telephone because of a habit resulting from natural face-to-face dialogue. This idea led to a similar, but stronger, claim put forward by Clark (1996), suggesting that speech and gesture are integrated during language production to create a composite signal, meaning that it is “difficult to produce the speech without the gesture” (Clark, 1996, p.179). Clark’s hypothesis has been directly tested in several studies: Restricting participants’ ability to gesture has been found to hinder the retrieval of words with spatial content (Rauscher, Krauss, & Chen, 1996), and is associated with fewer semantically rich verbs (Hostetter, Alibali, & Kita, 2007a) and, in some studies (but not all, see Hoetjes, Krahmer, & Swerts, 2014), more disfluency (Finlayson, Forrest, Lickley, Beck, & Margaret, 2003; Rauscher et al., 1996). This approach of assessing performance in conditions where gesturing is restricted in comparison to when participants are free to gesture

has also been successful in tasks unrelated to speech: Goldin-Meadow, Nusbaum, Kelly, and Wagner (2001) found that freedom to move facilitated mental arithmetic (the majority of gestures produced were point gestures). Movements which contain no meaningful content such as rhythmic beat gestures have also been shown to facilitate speech: In Ravizza's (2003) study, participants who were asked to tap at their own pace retrieved rare words from their definitions at higher rates than those who were not asked to tap (see also Lucero, Zaharchuk, & Casasanto, 2014). Similarly, studies have linked eye-gaze with cognitive demand, aiding both memory (Glenberg, Schroeder, & Robertson, 1998) and visual imagery (Spivey & Geng, 2001). Taken together, these studies suggest a speaker-oriented role for the benefits of many motor actions produced alongside speech or other cognitively demanding computations.

The fact that gestures are part and parcel of the act of producing speech (rather than just intentionally communicative actions) is also evidenced in studies showing that speech and gesture are deeply integrated in the production of language, especially those which investigate the production of mismatching speech and gesture. Speakers encounter difficulty when producing speech which is incongruent with pointing gestures (Chieffi, Secchi, & Gentilucci, 2009) and conventionalised signs (Barbieri, Buonocore, Volta, & Gentilucci, 2009; Bernardis & Gentilucci, 2006), suggesting that words and gestures are coded by a single communication system as a single composite signal (see Gentilucci, Dalla Volta, & Gianelli, 2008). Along a similar vein, a study from Kita and Özyürek (2003) found that information in gesture is semantically coordinated with information in speech: When describing an arced trajectory, English speakers used words such as “swing” and produced a gesture with an appropriate curvature, but Turkish and Japanese speakers, for whom there is no equivalent word to “swing”, used words such as the equivalent of “go” and produced gestures with a straight trajectory. Instead of using gesture to compensate for the lexical gap in Turkish and Japanese vocabularies, gestures are

produced which parallel the spoken content, suggesting that the two modalities are manifestations of a single communication system.

As well as being tightly coupled in terms of their content, speech and gesture are linked in time. This is evidenced by findings that the relative coordination in time of speech and gesture is influenced by the familiarity of the words spoken (Morrel-Samuels & Krauss, 1992), and that stutterers' interruptions in speech tend to coincide with interruptions in gesturing (Mayberry & Jaques, 2000). Additionally, producing a manual beat gesture while speaking has been shown to influence the prosody of the accompanying word, independently of both position of pitch accent, and how the beat gesture was produced (hand, head or eyebrow) (Krahmer & Swerts, 2007). The fact that speech and gesture are so deeply integrated in language production has led to an alternative explanation for why people produce these movements: Gestures are part and parcel of the act of producing speech, rather than intentionally communicative actions.

There have been several different accounts of the mechanism by which these gestures can facilitate production of speech. One such explanation posits that gesturing increases activation of items in the lexicon, cross-modally priming the retrieval of those items (Hadar & Butterworth, 1997; Krauss, Chen, & Gotfexnum, 2000). This explanation draws on evidence such as the greater use of gesture in spontaneous as opposed to rehearsed speech (see Chawla & Krauss, 1994); the association between a gesture's duration and how long (from onset of gesturing) it takes the speaker to access the word it represents (Morrel-Samuels & Krauss, 1992); and the effects on retrieval of words with spatial content when gesture is restricted (Rauscher et al., 1996).

Another way in which these gestures may be beneficial for speaking is by preventing visuo-spatial imagery from decaying, thus providing speech production processes with higher quality information (Hadar & Butterworth, 1997). This suggested

function of gesturing is compatible with the above account, in that illustrative gestures may originate from, and maintain, visual imagery which in turn primes lexical search. The role of gesturing in maintaining visuo-spatial imagery is supported by findings that iconic (i.e., illustrative) gesturing increases when tasked with describing visual objects from memory than describing visually present objects (Wesp, Hesse, Keutmann, & Wheaton, 2001), and when tasked with describing objects which are more difficult to verbally encode (Morsella & Krauss, 2004).

Alternatively, gesturing may be beneficial for the conceptual processes preceding language production, helping speakers to package complex information into appropriate units for speech (see Kita, 2000; Kita & Özyürek, 2003). This account, known as the Information Packaging Hypothesis, is supported by evidence from studies which manipulated the conceptual load required to formulate descriptions of stimuli. In Hostetter, Alibali, and Kita (2007b), participants were tasked with describing arrays of dots in terms of the geometric shapes which connected them. In one condition, participants had to generate their own conceptualisation of geometric shapes underlying the array of dots they were presented with, whereas in the other condition a geometric conceptualisation was superimposed upon the array, and they described the array in terms of those shapes. Participants gestured more when they had to conceptualise the spatial orientation of dots themselves than when these were given to them (dots superimposed onto shapes), suggesting that increases in gesturing are a result of the conceptual demands required to organise information into possible linguistic formulations.

The above studies provide a growing body of evidence suggesting that the demands required to plan and produce speech both influence and are influenced by the use of gesture: Speakers produce more gestures under increased cognitive load (Hostetter et al., 2007b; Morsella & Krauss, 2004; Wesp et al., 2001) and perform tasks better when free to gesture (Hostetter et al., 2007a; Rauscher et al., 1996; Ravizza, 2003). Whether gestures' facilitatory effects on speech production are

due to priming across modalities, helping to package information into units for speech, maintaining visuo-spatial imagery, or even by some underlying mechanism common to both speech and gesture (not discussed in depth here, but see Chu & Kita, 2016; Kita, 2014; Kita & Özyürek, 2003), the common thread in all these accounts is that gesturing can be advantageous to speaking. By these accounts, many gestures are produced because they facilitate production of speech—i.e., regardless of whether they are redundant for an addressee, they are not redundant for the speaker.

Having discussed how gesturing appears in some instances to be an epiphenomenon of the act of producing speech, the following section turns to a further way in which non-verbal behaviours may be by-products of speaker-internal processes, reflecting unconscious displays of the speaker’s cognitive and emotional states.

2.2.3 Leakage

In Section 2.1 we discussed how non-verbal behaviours occurring during communication vary from traditional gestures representing meaningful content to other aspects within a speaker’s motor control such as self-adaptive movements, postures, and facial expressions, often grouped together under the term *body language*. Although these behaviours are not limited to acts of communication, when produced alongside speech they form a vital part of communication, conveying meaning indirectly via information about a speaker’s cognitive and emotional states while producing an utterance. Sometimes, these behaviours may be part and parcel of the act of communication: Anecdotally, a furrowed brow accompanying the utterance “her name is [pause]” may be used (either intentionally or automatically) to communicate recall difficulty (see also Figure 2.2 in Section 2.1). However, it may also be unconscious displays of a speaker’s cognitive and affective states during the production of speech (Ekman & Friesen, 1969). This idea of *non-verbal*

leakage can show how the visual channel conveys far more meaning than just that which is directly represented by a speaker's gestures: A range of other visible behaviours offer information pertaining to a speakers' message indirectly, via cues to speaker-internal states.

Research into body language has tended to be concerned with its relationship with a speaker's emotions. A whole field of research has been devoted to the categorisation and recognition of emotions in facial expressions and actions (see Ekman, 1992; Ekman & Friesen, 1969, 1975), with a growing set of established mappings from emotions to distinct expressions (anger, disgust, fear, happiness, sadness, and surprise to name the basic six). Additionally, evidence suggests there are typical body languages associated with both anxiety (Daly, 1978; Gregersen, 2005), shame (Keltner, 1995; Keltner & Harker, 1998), and reticence (Burgoon & Koper, 1984). In a similar vein, evidence suggests that there are specific facial expressions which are associated with descriptors relating to increased cognitive load such as "thinking", "concentrating" and "confusion" (Rozin & Cohen, 2003) suggesting that these non-verbal behaviours may vary systematically with cognitive load (or at least may be perceived to do so).

Despite learning to regulate displays of emotions (and emotions themselves, e.g., Cole, 1986), it is suggested that people have less control over emotionally driven non-verbal behaviours (Ekman, O'Sullivan, Friesen, & Scherer, 1991; Sporer & Schwandt, 2006), and so often *leak*. In a 1989 study, Babad et al. asked judges to separately assess the verbal and non-verbal channels of recordings of teachers talking to high- and low-expectancy students. When talking to low-performing students, Teachers were rated as displaying more negative affect in facial expressions and body language, while ratings based on verbal channels perceived them to be more didactic, suggesting that although controlling their speech, teachers' non-verbal behaviours 'leaked'.

Sections 2.2.1 through 2.2.3 have shown how, across all different types of gestures and non-verbal behaviours which occur during communication, the cause of these behaviours may vary: some may be intended to communicate, others to aid the production of speech, and others still may be unintentional and at times revealing displays. The following section turns to the comprehension of speech and gesture, and discusses how a speaker's gestures and non-verbal behaviours might aid listeners' understanding of the speaker's message, regardless of whether they are intended to communicate, or to aid speaking, or are not produced intentionally at all.

2.3 Comprehension of multi-modal language

Above, we have discussed in brief three roles that non-verbal behaviours may have for a speaker, listed below:

- To intentionally convey information to an addressee
- To aid the planning and production of speech (and other cognitive tasks)
- As unintentional displays/leakage of emotional and cognitive states

In comprehension however, the purpose for which a speaker produces these behaviours is less important. Any gestures and non-verbal behaviours which are not intended to convey meaning may still communicate by virtue of the potential benefits they have for the addressee, both as an aid to comprehension of speech and as a source of information in themselves.

A growing body of work has shown that gesturing plays an important role in comprehension. In a 2011 meta-analysis of 63 studies on comprehension of speech

with gesture (compared to speech alone), Hostetter found that gesturing improved the immediate comprehension of a message, strengthened a message’s memorability, and improved understanding enough to result in greater learning and effective application of a message’s content.

There are many possible ways in which gesturing might facilitate comprehension. Firstly, the act of gesturing appears to alter the quality and content of spoken descriptions, meaning that, for a listener, utterances accompanied by gesturing are likely to be of a better quality, both in delivery—for instance having fewer filled pauses (Rauscher et al., 1996), and in content—having more semantically rich verbs (Hostetter et al., 2007a). In this way, any benefits gesturing has on production of speech (see Section 2.2.2) correspond to improvements in the audio stream presented to a listener.

Gesturing may also help in capturing the attention of an addressee. High rates of gesturing generally result in positive attitudes towards that speaker regarding aspects such as their competence (Maricchiolo, Gnisci, Bonaiuto, & Ficca, 2009) and likeability (Kelly & Goldsmith, 2004). In a real-world example, the number of views online that TED-talks (invited speakers giving talks on various topics) receive has been claimed to be directly correlated with the number of hand gestures made in a talk (see <https://www.scienceofpeople.com/secrets-of-a-successful-ted-talk/>). At the sentence level, gestures can also direct addressees’ attention to specific words: Rhythmic beat gestures have been shown to help focus attention on important information in speech, for instance in identifying the subject in ambiguous German sentences (Holle et al., 2012), and improving subsequent word recall (Igualada, Esteve-gibert, & Prieto, 2017).

Aside from improving the quality and content of the audible channel (in facilitating spoken descriptions) and maintaining and directing an addressee’s attention, gestures and non-verbal behaviours are themselves a valuable source of information

available to addressees. Gestures can be meaningful in that they can directly represent meaningful content similar to that conveyed in speech (albeit visually rather than lexically). Additionally, the mere occurrence of certain non-verbal behaviours is informative: Beginning to produce a large iconic gesture may indicate that formulating verbal descriptions of an upcoming referent is conceptually demanding (e.g., Hostetter et al., 2007b), and producing self-adaptive movements may indicate that the speaker is experiencing anxiety (e.g., Gregersen, 2005).

An important distinction here is that non-verbal behaviours can be meaningful to a listener in two ways. Firstly, a behaviour may directly convey meaningful content in a similar way to the semantics of speech—through conventions, visual representation or deixis. Secondly, behaviours may signal information about *the act of communicating*—about the speaker, message or discourse. The distinction here is between the representational information (if any) conveyed by a behaviour, and the information conveyed by *the fact that the speaker is producing that type of behaviour*. Research into how people understand and process language in multiple modalities has tended to focus on the former, investigating how gestural content influences comprehension of a speaker’s message or intended meaning. Section 2.3.1 provides an overview of this research, before Section 2.3.2 discusses the second way in which meaning may be perceived in non-verbal behaviours: as meta-communicative signals.

2.3.1 Comprehension of gestural content

In Section 2.1 we introduced Kendon’s Continuum, as a means of apportioning the ways in which gestures carry meaning alongside speech: A gesture’s content can convey meaning via conventions or norms, representationally, or deictically (see Figure 2.1, Section 2.1). We also explained in Section 2.2 how, beyond the content conveyed in speech, a gesture’s meaning can be either redundant or non-redundant.

Non-redundant gestures specifically have been found to have a greater effect on comprehension: In Hostetter's meta-analysis, the facilitatory role of speech with gesture relative to speech without was greater for gestures which conveyed task-relevant information not expressed in speech than for gestures which conveyed the same information as was expressed in speech. Moreover, recent evidence suggests that listeners are more likely to fixate on speakers' gestures when they expect them to be non-redundant (Yeo & Alibali, 2017).

The meaning conveyed by non-redundant gestures may express content which is implicit in speech, such as the speaker's intended meaning in producing an utterance, thereby aiding listeners' comprehension not just of the literal message but of intended meaning. For example, both adults and children have been shown to better understand indirect requests (e.g., "It's getting hot in here" as a request that someone open a window) when accompanied by a relevant pointing gesture (e.g., pointing to the window, see Broaders & Goldin-Meadow, 2010; Kelly, 2001; Kelly, Barr, Church, & Lynch, 1999). Further evidence that comprehension is influenced by gestures which convey content different to that of speech can be found in studies which show that listeners' understanding of speech is impeded by gestures which are directly incongruent with spoken content (see e.g., Goldin-Meadow & Sandhofer, 1999; Kelly & Church, 1998; Kelly et al., 2010).

Even so-called redundant gestures which co-express the same content as speech are not truly redundant for listeners: The visual modality is particularly useful for conveying certain spatial and kinetic information. When speech and gesture are both used to reference the same object or action (e.g., "a triangle" [gestures triangle shape]), the gesture offers the addressee extra information about the factors such as the orientation and relative size of the object. Gesturing's capacity for easily conveying spatial information is supported by Hostetter's finding that gestures relating to topics about movements have a greater facilitatory influence on comprehension than those about abstract topics. This has also been tested

directly in a study by Driskell and Radtke (2003), in which speakers were tasked with conveying a target word to a listener without using that word—much like the popular board game *Articulate*TM. When speakers were allowed to gesture, listeners required fewer guesses before identifying the correct word. Importantly, this beneficial effect of gesturing on comprehension was present for spatial (*under, short*) and movement (*hold*) terms, but not for non-spatial terms (*warm*).

The growing body of evidence suggests that gesture does influence comprehension, with Hostetter's (2011) meta-analysis suggesting a set of moderators emerging across studies: Gestures appear to influence comprehension more when the topic is about spatial or motoric information; when gestures express unique information that is not in speech; and for audiences of children more than adults. Although these studies offer convincing evidence that gesturing does facilitate comprehension, there has been little research into exactly *when* during the time course of the comprehension processes gesture contributes to understanding, and how it interacts with the comprehension of spoken language.

Comprehension of gestural content in real-time

Co-speech movements may have benefits for both parties in a conversation: From providing an alternative modality in which a speaker can convey information, in turn aiding an addressee's comprehension of the message, to facilitating the production of speech, resulting in improvements to the audio stream for a listener. Many studies, such as those reviewed in Hostetter (2011), have investigated the effects of gesture on either message recall, the ability to use information from the message effectively, or via some after-the-fact measure of comprehension of the message. Such measures of comprehension tend to be either questions (for example, Beattie & Shovelton, 1999a, 1999b; Holler, Shovelton, & Beattie, 2009; Kelly, 2001; Kelly et al., 1999) or tasks such as matching an object with the

content of the message (see Driskell & Radtke, 2003; Krauss et al., 1995), both of which are conducted after the presentation of the stimulus.

One exception is a study by Kelly et al. (2010) which relied on a measure of reaction time. Kelly et al. (2010) presented participants with words paired with pantomime gestures, for which both modalities varied in their semantic congruency with a previously seen video of an action being performed. Speech and gesture either both expressed information which matched the action in the initial video (e.g., “dial” [dialling gesture] following a video of someone dialling a phone), expressed incongruent information in one modality, or (in filler trials) expressed incongruent information in both modalities. Incongruency could be either weak or strong (see Table 2.1). Tasking participants with responding (via key-press) to whether or not information (in either speech or gesture) matched or mismatched with the video before, Kelly et al. (2010) found that incongruencies in either modality resulted in slower responses, and that the strength of the incongruency influenced the number of incorrect responses (stronger mismatches resulting in more errors). This result held even when participants were asked to respond only to whether speech (and not gesture) matched the previously seen action presented in the video (Experiment 2 Kelly et al., 2010), suggesting that the integration of the two modalities is obligatory. The main findings from Kelly et al. (2010) are replicated in Chapter C, but it is only recently that studies have begun to measure directly the time-course of the processes involved in understanding multi-modal language.

Studies of the comprehension of spoken language suggest that listeners are able to predict one another’s material, drawing on information at various levels of the input to pre-activate upcoming content alongside the moment-to-moment processing of speech (see Kuperberg & Jaeger, 2016; Kutas & Federmeier, 2011). Because pre-activation during comprehension is not necessarily achieved with any conscious awareness of the listener, investigation requires experimental paradigms which allow on-line measurements of behavioural or neurophysiological effects,

Table 2.1: Examples of weak and strong incongruity speech-gesture pairs in Kelly et al. (2010).

Incongruent Modality	Level of Incongruence	Example
Gesture	Weak	“dial” [typing gesture] following a video of a phone being dialled
Speech	Strong	“twist” [dialling gesture] following a video of a phone being dialled

from which it is possible to make inferences about the predictions, hypotheses and judgements which listeners hold about the unfolding language.

One method of investigating language comprehension in real-time is to study Event Related Potentials (ERPs)—measuring systematic patterns in electrical activity at the scalp. The first ERP language studies used a paradigm which manipulated the load required to integrate a word into the context of a preceding sentence, to elicit what is known as an N400 effect in the brain. The first instance of this was a study by Kutas and Hillyard (1980), in which the semantic incongruity of a sentence-final word (e.g., “He spread the warm bread with *socks*”) relative to a congruent control word resulted in an N400 effect. Since then, the N400 effect has been found to be triggered by the semantic congruency of target words or pictures with various contexts, be it sentences, discourses (Van Berkum, Hagoort, & Brown, 1999), world knowledge (Hagoort, Hald, Bastiaansen, & Petersson, 2004), or preceding pictures (McPherson & Holcomb, 1999).

Investigations of language comprehension in real-time (such as those above) have tended to focus on comprehension of verbal utterances in isolation, without involvement of any visible information about the speaker. Some studies, however, have used these methods to investigate comprehension of speech and gesture,

with results suggesting that gestural content can have a rapid and direct effect on comprehension (as indicated by the results of Kelly et al. 2010). In gesture research, ERP effects have been measured when participants are presented with words that vary in their semantic relationship to preceding gestures (Kelly, Kravitz, & Hopkins, 2004); gestures in relation to preceding cartoon images (Wu & Coulson, 2005); and speech-gesture mismatches in relation to sentence context (Özyürek, Willems, Kita, & Hagoort, 2007).

In Kelly et al. (2004), participants viewed a scene in which a speaker stood behind a table upon which there were two objects. ERPs for spoken adjectives relating to one object were found to be influenced by whether or not a gesture immediately preceding speech (and held during speech presentation) contained congruent information about the size and shape of that object. Relative to trials in which gesture and speech conveyed the same meaning, Kelly et al. found an N400 effect when gestures strongly mismatched the speech, suggesting integration of the semantic content in the two modalities. Kelly et al. also found earlier effects of two types of mismatch: Gestures which strongly mismatched speech (these gestures were in fact directed toward the other object on the table) and gestures which complemented speech (gestures directed towards the correct object but representing a different dimension i.e., “tall” with a [thin] gesture). Kelly et al. claim that these results show that the semantic content of gestures influences both the early “sensory/phonological” processing of linguistic information as well as the later semantic processing.

ERPs have also been measured in response to more naturalistic stimuli, rather than pairs of gestures and words presented sequentially as in Kelly et al. (2004). Holle and Gunter (2007) investigated whether a gesture supporting either the dominant or subordinate meaning of a homonym moderated an N400 effect elicited at a subsequent disambiguating word. For example, a gesture presented alongside the homonym “ball” could support either of the possible subsequent target words

“game” or “dance”. Holle and Gunter found evidence suggesting that participants used the content of the gesture to disambiguate speech: The N400 at target words was larger when the gesture supported the alternative meaning. Holle and Gunter’s finding supports Kelly et al.’s (2004) claim that gestural content is integrated rapidly with the processing of speech, informing listeners’ predictions of upcoming content.

While Kelly et al. (2004) and Holle and Gunter (2007) measured ERPs time-locked to speech targets in a context partially defined by preceding gestures, Özyürek et al. (2007) investigated responses to gestures more immediately by manipulating the congruency of speech and gestures with a preceding sentential context. Özyürek et al. presented participants with audiovisual stimuli such as those in examples A and B below, in which incongruity is presented in gesture (A) or speech (B).

A “He slips on the roof and rolls down [walking gesture] the other side”

B “He slips on the roof and walks to [rolling gesture] the other side”

The N400 effects elicited from such stimuli were similar despite the cross-modal incongruency. Özyürek et al. found that when information which mismatches with preceding sentence contexts is presented in speech, in gesture, or in both modalities, the latencies, amplitudes, and topographical distributions of electrophysiological responses were similar, suggesting that during language comprehension, listeners simultaneously incorporate information from different domains/modalities to interpret meaning.

Gestural imprecision and noise

The line of research described above has proved fruitful in establishing that the semantic content conveyed by gestures can have a direct effect on comprehension,

and that this happens concurrently with the processing of linguistic input. In every day conversation, however, the semantic content of a gesture may be less clear to a listener. This may be due to imprecision or ambiguity in how a gesture is presented, or to the amount of noise in the visual channel.

Studies have demonstrated that people often misinterpret the specific meaning in illustrative gestures (Feyereisen, Van de Wiele, & Dubois, 1988; Hadar & Pinchas-Zamir, 2004; Krauss, Morrel-Samuels, & Colasante, 1991). When presenting participants with a series of gestures taken from longer narratives, and tasking them with choosing from two possible lexical affiliates for each gesture, Krauss et al. (1991) found that lexical affiliates which were objects and descriptions were correctly selected only 55% of the time, with actions and locations being correctly selected 69% of the time. In a further experiment, Krauss et al. found that judgements of whether a gesture referred to an action, location, object name or description, were derived almost entirely from the semantic content of accompanying speech (audio-video condition did not differ from audio-only, but video-only resulted in poorer judgements). Findings suggest that interpretations of gestures are relatively imprecise, perhaps indicating that naturally occurring gestures tend to lack the precision of spoken language, and may often be more lax than those used in experimental settings (leading to low or medium ratings of inter-coder reliability, see Eisenstein & Davis, 2004).

An additional factor is that during everyday communication the visual modality is full of noise—speakers produce numerous movements for a wide variety of reasons, and types of gestures might not be mutually exclusive (e.g., a beat gesture may also contain some illustrative content, see McNeill, 1992). This may make it more difficult to distinguish meaningful content represented by gestures. Evidence suggests that visual noise interferes with comprehension. Holle and Gunter's (2007) study (described above) also included a version which included a situational context where a speaker produced a lot of meaningless movements (unrelated

self-adaptive gestures), and found that the facilitative influence of gesture on addressees' disambiguation of homonyms was attenuated when the speaker made more of these movements (Experiment 3). Holle and Gunter's findings suggest that the integration of gesture and speech depends partially on how clear the visual channel is. Moreover, Obermeier, Kelly, and Gunter (2015) developed Holle and Gunter's paradigm and found a speaker specific weakening of the integration of gesture with speech, suggesting that addressees adjust the influence gesture has on comprehension to accommodate for different gesturing styles of speaker. In the real world, comprehending meaning in gestures requires navigating a visual channel which is full of noise for meaningful movements which may be lacking in clarity. In many cases, a speaker's gestures and non-verbal behaviours convey meaning in a different way, as signals about the act of communication itself.

2.3.2 Representational gesturing vs. meta-communicative signalling

The general approach to investigating how non-verbal behaviours influence comprehension of a speaker's message (using both on-line and post-factum measures) has been to investigate how the semantic content of gestures interacts with that of accompanying speech (Section 2.3.1). In everyday communication, however, interpreting the semantic content of a gesture may be more difficult. This may in part be due to the fact that gesturing occurs in a continuous physical space, leading to imprecision and ambiguity in gesture content, and because of the multitude of other movements a speaker may produce during communication. An alternative way in which speakers' non-verbal behaviours might carry meaning is by signalling *meta-communicative* information about the speaker, the message, or the dialogue.

‘Meta-communication’

Clark (1996, p.241) drew a distinction between communication and *meta-communication*: Beyond the topic of the conversation, interlocutors are constantly engaging in communication about the process of the dialogue—about what each interlocutor is doing across a hierarchy of levels. Consider, for example, the exchange of utterances in Table 2.2 (from Clark, 1996, p.245).

Table 2.2: Example of meta-communication from Clark (p.245, 1996)

Utterance	Level 1	Level 2	Level 3
D: “Forty-nine Skipton Place”	The address is Forty-nine Skipton place	Do you understand that the address is Forty-nine Skipton place?	
J: “Forty-one”	I ratify that the address is Forty-one	I heard “Forty-one”	Did you present “Forty-one”?
D: “Nine. nine”			No. the “one” is “nine”
J: “Forty-nine”	I ratify that the address is Forty-nine	I heard “Forty-nine”	

Beyond the words spoken in this dialogue, Darryll (D) and June (J) are engaging in several levels of what Clark termed *meta-communication*. In producing the utterance “Forty-one”, June is communicating that she heard the number forty-one, and posing a query back to Darryll about whether he had presented the number forty-one in his initial utterance. According to Clark, meta-communication is recursive, occurring on several levels—i.e., we communicate about communicating

about communicating. The levels of meta-communication are also given in Table 2.2.

Speakers' motor actions can also communicate about the act of communication. Bavelas, Chovil, Lawrie, and Wade (1992) identified a set of gestures which appeared to function as a means of coordinating and managing conversation, for example in indicating previous contributions to the discourse (e.g., flicking the index finger across towards the interlocutor as if to say "as you just said"). Additionally, in Holler and Wilkin's (2011) study, addressees were found to mimic a speaker's gesture of Tangram figures. In mimicking gestures, addressees both convey information about the Tangram figure, but their iconic gesturing is also used to convey their understanding of the information just conveyed by the speaker—an act comparable to June's utterances in the example exchange in Table 2.2.

Non-linguistic signals

Many of a speaker's non-linguistic behaviours may signal meta-communicative information by tending to occur in a systematic manner, according to, for example, the cognitive load and emotional states of the speaker. These behaviours need not be intentionally produced by a speaker for them to be beneficial for a listener.

Speech disfluency is one such non-linguistic signal which has received a lot of attention in research. Disfluencies in speech have been suggested to vary according to lexical and conceptual factors involved in producing an utterance: For instance, speech has been found to be more disfluent when producing less-preferred syntactic structures (Cook, Jaeger, & Tanenhaus, 2009) or discourse-new expressions (Arnold, Losongco, Wasow, & Ginstrom, 2000), or choosing from a larger range of expressive alternatives (Schachter, Christenfeld, Ravina, & Bilous, 1991; Schachter, Rauscher, Crone, & Christenfeld, 1994). Disfluencies also tend to occur more frequently at

beginnings of phrases or utterances (Barr, 2001; Boomer, 1965; Shriberg, 1996), at major discourse boundaries (Swerts, 1998; Watanabe, 2002), and before longer utterances (Oviatt, 1995).

In turn, specific disfluencies in the audio stream presented to listeners have been found to influence comprehension. For example, Corley, MacGregor, and Donaldson (2007) found that an N400 effect associated with unpredictable vs. predictable words was reduced when a disfluency preceded the target word, suggesting that disfluency in some way prepares listeners for less familiar words. Similarly, following a filled pause (“um” and “uh”), listeners’ eye-gaze and mouse movements have been found to be directed towards discourse-new (Arnold, Tanenhaus, Altmann, & Fagnano, 2004; Barr & Seyfeddinipur, 2010) and unfamiliar (Arnold, Hudson Kam, & Tanenhaus, 2007) objects, as well as implicit judgements that the speaker is dishonest (Loy, Rohde, & Corley, 2017) (although whether this is a valid association for listeners to draw is less clear).

Several studies have indicated parallels in the production of speech disfluency and certain forms of gesture: For example, an increase in disfluency found when producing less preferred syntactic structures was matched by an increase in gesturing (excluding self-adaptive movements) in Cook et al. (2009). Likewise, Butterworth and Beattie (1978) proposed a set of non-verbal behaviours which they termed *speech focussed movements* (those which are rhythmically timed with and reflect the meaning of speech), suggesting that these movements tend to parallel vocal hesitations as indicators of planning in speech. The studies discussed in Section 2.2 suggest that iconic gestures may also vary according to the demands of producing speech: Increased rates of iconic gestures have been associated with descriptions of a) spatial or motor information (for a review see Alibali, 2005), b) objects from memory (Wesp et al., 2001), and c) less describable and more conceptually demanding referents (see e.g., Alibali, Kita, & Young, 2000; Hostetter et al., 2007b; Morsella & Krauss, 2004).

It stands to reason that listeners' sensitivity to speech disfluency to inform comprehension may also extend to the presence of different types of non-verbal behaviour. For example, the presence of iconic gesturing (regardless of what content it conveys) may inform listeners about the speaker's speech production processes in that an upcoming referent might have particularly salient spatial or dynamic properties, is no longer present, or is hard to describe verbally. Similarly, speakers' non-verbal leakage of affect (see Section 2.2.3) can inform listeners' perception of speakers' emotional and cognitive states while producing an utterance, in turn influencing comprehension of an utterance. Research into detection of deception points towards this being the case: Meta-analyses have revealed that, along with speech disfluencies, listeners reliably interpret certain non-verbal behaviours—namely postural shifts, and increased arm, foot and leg movements—as signals that a speaker is being deceitful (and that the accompanying utterance is not true) (see Hartwig & Bond, 2011; Zuckerman, DePaulo, & Rosenthal, 1981). We will revisit this evidence in Chapter 6.

Research into how manner of spoken delivery influences comprehension suggests that listeners are sensitive to speech disfluencies in an on-line manner—e.g., anticipating less familiar objects or words (Arnold et al., 2007; Barr & Seyfeddinipur, 2010; Corley et al., 2007) during the processing of speech. The influence of non-verbal delivery as signalling meta-communicative information, has not received the same attention. To our knowledge, no studies have investigated whether listeners interpret speakers' non-verbal behaviours as signals of difficulties in planning and producing speech. Furthermore, the one aspect of meta-communication which has been widely researched in relation to speakers' manner of non-verbal delivery—perception of deception—has only been studied in after-the-fact judgements or explicit beliefs about non-verbal signals (e.g., Vrij & Semin, 1996; Zuckerman, Koestner, & Driver, 1981) (unlike the time course over which manner of spoken delivery influences judgements of deception; Loy et al. 2017). With this in mind,

this thesis aims to obtain a more informed view of if and when speakers' non-verbal behaviours are interpreted as signals about the speaker and speech production process.

2.4 Methodology

Thus far, we have discussed how non-verbal behaviours can serve numerous functions during everyday communication, performing various roles beyond the straightforward purpose of providing a mode in which a speaker can intentionally convey meaning. Specifically, we have explored the numerous advantages that gesturing has for the production of speech (e.g., at the planning level: Kita 2000; and in assisting recall: Wesp et al. 2001), as well as the idea that non-verbal behaviours can *leak*, reflect a speakers' cognitive and emotional states whilst speaking. In turn, we have seen how the content conveyed by gestures has been shown to facilitate comprehension (see Hostetter, 2011), and is integrated into comprehension simultaneously with speech (e.g., Kelly et al., 2004; Özyürek et al., 2007). Finally, we have drawn on research in the manner of spoken delivery, and discussed how speakers' non-verbal behaviours might be meaningful to a listener beyond what a gesture depicts, as meta-communicative signals about a speaker and their message. This thesis is concerned with this last point, focussing on how different types of non-verbal behaviours may be associated with, firstly, the conceptual demands required to formulate an utterance, and secondly, the intention to deceive. The former question, to our knowledge, has not been studied in any respect, while the latter has received extensive attention in relation to various non-verbal behaviours and subsequent judgements of deception, but research is lacking with respect to the time-course of these judgements.

In order to investigate if and when non-verbal signals influence the comprehension

process, many of the studies presented here make use of eye- and mouse-tracking methodologies. These techniques have been extensively used to investigate the mechanisms underlying comprehension of spoken stimuli, with very few studies including visual displays of the speaker during stimulus presentation. The following section discusses these approaches to measuring the time course of comprehension, with a focus on the few studies which have attempted to use these methods to investigate comprehension of gestures specifically.

2.4.1 Eye-tracking, mouse-tracking and the Visual World Paradigm

Eye-tracking has been employed in many studies to measure how listeners' visual attention is coordinated at a given point during the presentation of a spoken utterance. Advances in eye-tracking technologies in the 1970s and 80s brought about an explosion in its use in psychological research. Studying the cognitive processes involved in a variety of tasks such as visual search and scene perception has been made possible by eye-tracking, and in the field of language this began with studies in reading (for an overview, see Rayner, 1998), and more recently has expanded into investigating the processing of spoken language. Much of the research into spoken language processing has followed from the development of the *visual world paradigm* (Cooper, 1974; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995), in which listeners' gaze behaviour towards a display of various objects is measured while they hear speech, allowing the analysis of patterns of eye movements *over the course of* spoken utterances. From these patterns it is possible to draw inferences about how listeners' anticipations, evaluations and judgements of speech unfold over the course of an utterance.

The first studies to use the visual world paradigm simply showed that eye movements are drawn towards objects in a display which are referenced by the

accompanying speech: Upon hearing the word “lion” in a narrative, listeners are more likely to fixate upon a picture of a lion within a display compared to other unrelated pictures (Cooper, 1974). These studies demonstrated a systematic relationship between linguistic processing and visual processing. By manipulating how the objects in the display relate to the linguistic input, it is possible to draw conclusions about how and when listeners’ develop hypotheses based on the speech stream. As an example, in a 1998 study, Allopenna et al. used the visual world paradigm to study phonological processing. Allopenna et al. found that, upon hearing a word (e.g., *beaker*), participants tended to fixate toward the picture of the beaker, but also fixated more initially to an object with the same onset and vowel (e.g., *beetle*) and then to an object which rhymed (*speaker*) compared to an unrelated object (*carriage*). The visual world paradigm has been employed in studies approaching many different aspects of comprehension (for an overview, see Huettig, Rommers, & Meyer, 2011), from phonological processing (Allopenna et al., 1998; Marslen-Wilson, 1987), to the effects of semantics (Huettig & Altmann, 2005; Moss & Marslen-Wilson, 1993) and syntax (Dahan & Tanenhaus, 2004; Kamide, Altmann, & Haywood, 2003), pragmatic inferencing (Grodner, Klein, Carbury, & Tanenhaus, 2010; Huang & Snedeker, 2009), and even the preactivation of visual shapes (Dahan & Tanenhaus, 2005). The visual world paradigm has also opened the door to research into how the delivery of spoken language influences comprehension. This approach has been especially fruitful in investigating how listeners respond to disfluencies in speech. In one such study by Bailey and Ferreira (2007), disfluencies within a syntactically ambiguous utterance were found to cue listeners towards more complex syntactic constituents—a result which parallels previous research suggesting that speakers do produce disfluencies in such situations (Clark & Wasow, 1998). Visual world studies have found similar effects of disfluency on both semantic prediction (Arnold et al., 2007, 2004) and pragmatic judgements (Loy et al., 2017).

More recently, the recording of participants' mouse movements as a means of tracking the on-going development of their decisions to click on objects has gained interest in the field of language research. As with eye-tracking, by recording positions and trajectories of the cursor relative to specific responses on a screen, it is possible for researchers to study the influence of various experimental manipulations on the decisions participants make during these experiments. Mouse-tracking can be integrated into visual-world paradigms with comparative ease, offering a further means of tracking the time-course of lexical activations during real-time comprehension. For example, Spivey, Grosjean, and Knoblich (2005) tasked listeners with responding to spoken instructions such as "Click the candle", while viewing displays which depicted the target (candle) and a distractor. Listeners' mouse trajectories showed a marked attraction towards distractors which shared the same phonological onset with the target word (e.g., candy) compared to distractors which did not (e.g., pickle).

Studies have demonstrated that mouse movements can capture the same effects of spoken language as eye movements: The tendency towards more difficult-to-name objects which studies have found in listeners' gaze behaviour (Arnold et al., 2007; Barr, 2001) has also been found in mouse movements (Barr & Seyfeddinipur, 2010). Similarly, Farmer, Cargill, Hindy, Dale, and Spivey (2007) found that trajectories of listeners' mouse movements were influenced by visual context towards certain syntactic representations—an effect previously evidenced in eye movements (Tanenhaus et al., 1995). Mouse-tracking offers more than just an equivalent to eye-tracking: Eye fixations occur within a Poisson distribution (at any time, listeners are either fixating upon an object or they are not), and therefore it is possible that when aggregating over all trials in an experiment, the resulting patterns of fixations are not representative of different activations competing simultaneously for attention, but simply the averaged effects of individual trials in which a single activation occurs. In this respect, mouse-tracking offers a clearer

picture: Analysis of the curvature of trajectories can show attraction to objects in a display on a trial-by-trial basis by virtue of the fact that the measure captures continuous measurements across the display.

Although eye- and mouse-tracking methodologies such as the visual world paradigm have become widely used in research into the effects of manner of spoken delivery, approaches using these techniques to study manner of non-verbal delivery have barely gotten off the ground. Integrating visual information about a speaker into eye-tracking paradigms face two major obstacles. Firstly, matching specific non-verbal behaviours to a variety of utterances is difficult due to the coordination of speech with visible mouth-movements. This is a problem common to any study investigating comprehension of speech and gesture, and approaches have varied: Experimental stimuli have shown the speaker from the neck down (e.g., Kelly et al., 2010; Özyürek et al., 2007); shown the speaker with a nylon stocking on their head (Holle & Gunter, 2007; Holle et al., 2012); or cleverly shown the speaker conveniently obscuring their mouth with a sheet of paper (Saryazdi & Chambers, 2017).

Secondly, the saliency of a video in a display is likely to detract participants' visual attention away from the other objects in a display which are more often than not the objects of interest in terms of the comprehension process. One method of avoiding this problem is to reduce the salience of gestures by coordinating the presentation of speech and gesture such that either gesture occurs prior to the critical periods in the speech stream, or use mainly static gestures to avoid movement during the critical period. This approach has been somewhat successfully employed by Louwerse and Bangerter (2010) who presented participants with a disembodied arm of a speaker pointing towards a row in an adjacent array alongside referring expressions which unambiguously referred to one object by describing three features. Louwerse and Bangerter found that listeners' uptake of ambiguous but informative pointing gestures was similar to the effects of verbal

location descriptions, in allocating their attention to the appropriate sub-domain in the display.

Another method for dealing with the visual salience of gestures interfering with investigations of the comprehension processes is to design experiments in which participants' attention to gestures is treated as the dependent variable, manipulating content and delivery of speech. Although this method allows researchers to draw conclusions about how and when listeners allocate attention to gestures, it is limited in its capacity for studying the effects of these gestures on comprehension of the speakers' message. Yeo and Alibali (2017) employed this strategy and found that addressees were more likely to fixate at least once upon a speaker's gestures when the speaker was disfluent, and when the gesture was expected to be non-redundant (although these factors bore no relation to the time spent fixating upon gestures). However, the overall time participants spent fixating upon gestures in Yeo and Alibali's (2017) study was low (about 10% of the time), a result which patterns with findings from Gullberg and Holmqvist (2006), reporting low durations of gesture-directed gaze even when provided with extra motivation to attend to gesturing via a social cue (speaker's gaze).

On the basis that addressees do not attend to gestures very much, it may be that the problem of their visual salience is being over-stated, and inclusion of visual information about a speaker will not detract too much from fixation biases towards some objects in a display over others. The possibility of still capturing any potential effects of gesturing with this method is predicated on the suggestion from Gullberg and Kita (2009) that most gestures are perceived through peripheral vision. To our knowledge, only two studies have attempted variations of this approach: In one, Silverman, Bennetto, Campana, and Tanenhaus (2010) placed a video component in the center of a visual world paradigm with the aim of investigating differences between adolescents with autism and typical controls' abilities to integrate information across speech and gesture. In the second, Saryazdi

and Chambers (2017) used a large format display showing a video including both speaker and objects to study whether gestures representing size and shape influenced fixations to objects in the video. Notably, Saryazdi and Chambers found the presentation of a gesture increased the speed with which listeners visually identified objects, and this effect was evident within 600 ms of the onset of the critical noun. However, this facilitatory effect was only apparent for identification of smaller objects. One possible explanation of this is that participants in Saryazdi and Chambers' study may not have been extracting information about size and shape from the gestures, but were responding to an association between the occurrence of grasping gestures and reference to smaller objects.

The upshot of Saryazdi and Chambers' findings is that they show the potential of this line of research: The presentation of a gesture increased (rather than detracted from) listeners' fixations towards the target-object during early moments of comprehension. The current thesis develops this further, integrating a video component into a standard visual world paradigm (e.g., in which referenced objects are separate from the video). This makes counterbalancing the positions of objects within the display more straightforward, as well as reducing a possible confounding effect of participants misinterpreting the onset of gestures as points directed towards an object in the video.

2.4.2 The current thesis

As previously mentioned, this thesis approaches the question of how meaning is recovered from non-linguistic behaviours in two ways: As signals of conceptual demand (Part I) and as markers of deception (Part II). Chapters 4 to 8 present a series of experiments employing visual world paradigms in which we manipulate the non-linguistic behaviours presented along with speech.

In the majority of these experiments (Chapters 4, 5, and 6) the inclusion of videos of a speaker in the visual displays presented to participants allows us to manipulate the presence of different types of non-verbal behaviour accompanying spoken utterances. In Chapter 7 we manipulate both the presence of non-verbal behaviours in video and the presence of disfluency in audio. Finally, Chapter 8 focusses on non-linguistic behaviours in the audio channel, manipulating the presence of speech disfluency and the availability of other causes of disfluency. In this way, Chapters 4 to 8 aim to investigate whether and when different non-linguistic signals biases listeners' eye and mouse movements towards those objects in the display which a) are more difficult-to-name and b) indicate that an utterance is perceived to be a lie (in that the object is not the one referred to by the speaker). First, Chapter 3 presents a study of the production of speech and gesture, varying the conceptual demands required to formulate descriptions of referents.

Part I

Signals of speech production difficulty

Chapter 3

Conceptual difficulty and production of speech and gesture

In Part I of this thesis, we investigate how non-linguistic delivery varies with the conceptual demands required to formulate spoken descriptions, with an eye to how this may influence comprehension. Here, we focus on the manner of non-verbal delivery—i.e., on how the movements produced alongside speech can signal information about the speech production process. A parallel strand of research on manner of spoken delivery has been conducted with the following line of reasoning:

1. When speaking involves greater cognitive load, speakers produce more disfluency (when producing, e.g., less familiar, low frequency words, Arnold et al. 2000; Schnadt and Corley 2006; less preferred syntactic structures, Cook et al. 2009; and descriptions of new referents Barr 2001).
2. Disfluencies can therefore provide insight into the speaker’s mind, acting as possible signals of speaker effort.
3. Accordingly, comprehension is sensitive to the presence of disfluency:

Following disfluent speech, listeners tend to anticipate upcoming referents to be less familiar (Arnold et al., 2007; Corley et al., 2007), new to the discourse (Arnold et al., 2004), or new to the speaker (Barr & Seyfeddinipur, 2010).

Part I investigates whether this form of non-linguistic signalling applies to the domain of non-verbal delivery.

This chapter explores how the production of gesture alongside speech is influenced by the conceptual demands of a referent (equivalent to 1 in the example above). In doing so, it provides a baseline for the following chapters (4 and 5) which present comprehension studies investigating listeners' sensitivity to gesture when anticipating upcoming referents (parallel to 3, above).

A growing body of work suggests that when speech planning becomes more difficult, speakers tend to produce more gestures (see e.g., Alibali et al., 2000; De Ruiter, 1998a; Kita, 2000). Previous research pins this increase either on gestures helping the speech planning processes (for instance, in helping the speaker to *package* more complex information into appropriate units for speech, e.g., Kita 2000), or on some of the communicative load being traded off from speech to gesture (e.g., De Ruiter 2006). These offer two contrasting views on how the relationship between speech and gesture may signal information about the speech planning process. The former holds that the amounts of speech and gesture co-vary, increasing in parallel. The latter view suggests the inverse: A negative relationship between speech and gesture indicating a transference of communicative load to whichever modality is least difficult.

Studies attempting to discern between these two views have varied in their conclusions: Underspecification of referents in speech has been associated with underspecification in gesture (co-varying, So et al. 2009), but omissions of spatial content in speech are found more in speakers who gesture than those who do not

(trading-off, Melinger and Levelt 2004). These mixed findings may in part be due to the varied approaches to how gestures are measured. In the present study we measure the relative durations of gestures and speech directly, in order to establish whether these metrics co-vary (as would be predicted by the packaging account) or correlate inversely (as would be predicted by a trade-off). Pairs of participants took part in a shape-matching game, alternating in the roles of director and matcher. Directors saw two shapes (one easy, one difficult) for two seconds, and subsequently described them to their partner. In contrast to the trade-off account, speech duration and gesture duration were found to increase in parallel. Moreover, for objects which were more difficult to verbally encode, gesture duration increased at a higher rate than speech duration. Findings support the view that production of speech and gesture co-vary, but indicate that the relationship is somewhat more nuanced, and is dependent upon conceptual load.

3.1 The relationship of speech and gesture

Research suggests that speakers tend to produce more gestures when describing spatial or motor information (see Alibali, 2005), or referents which are more difficult to verbally encode (Morsella & Krauss, 2004). A traditional view is that this increase in gesturing is due to gestures which are produced to convey meaning. On this view, gesture production might increase because some of the communicative load is being traded off from speech to gesture (Bangerter, 2004; De Ruiter, 2006; Melinger & Levelt, 2004). Alternatively, gestures might have more direct benefits for speech (Kita, 2000; Krauss et al., 2000; Rauscher et al., 1996), helping with (rather than substituting for) the production of spoken words (for a full review see Section 2.2.2).

Explanations of the way in which gesturing may benefit speech production have

been varied. One hypothesis holds that gesturing increases the activation of relevant items in the mental lexicon, thus facilitating access (Krauss et al., 2000). Another suggestion is that gesturing prevents visuo-spatial imagery from decaying, providing speech production processes with higher quality information (Hadar & Butterworth, 1997). In a similar vein, the Information Packaging Hypothesis (Hostetter et al., 2007b; Kita, 2000; Kita & Özyürek, 2003) maintains that gesturing helps speakers to package complex information into appropriate units for speech. These positions are consistent with evidence that the production of gestures has been shown to facilitate working memory (Morsella & Krauss, 2004; Wesp et al., 2001), conceptual planning (Melinger & Kita, 2007), and even mental arithmetic (Goldin-Meadow et al., 2001). According to these views, complex information is articulated in full, but the difficulty in doing so is partially alleviated by gesturing. The resulting prediction is that speech and gesture increase in parallel, rather than trading off against one another.

A 2009 study by So et al. found evidence which patterned with this prediction. Participants were asked to describe scenes from videotaped vignettes (e.g., a man giving a woman a basket), and their uses of both speech and gesture to indicate characters in the scene were measured. So et al. found that speakers more often used a gesture to identify a referent if that referent was also specified in speech. So et al. viewed their results as evidence in support of an account of speech and gesture going ‘hand-in-hand’—i.e., the two modalities co-varying.

In contrast to So et al.’s (2009) evidence for the co-variance of speech and gesture, other studies point to the use of gesture to compensate for underspecification in spoken descriptions. For example, in a communication task about spatial arrangements of connected dots of different colours where it was possible to determine the minimal content necessary to uniquely identify each stimulus, Melinger and Levelt (2004) examined whether omissions in speech were accompanied by compensatory

gestures. Melinger and Levelt found that people who gestured made more—and different—omissions in speech than people who did not.

Melinger and Levelt's findings pattern with other studies (Bangerter, 2004; De Ruiter, 2006; der Sluis & Krahmer, 2007) in which the informational content of speech is found to inversely correlate with the amount or precision of gestures. This evidence contrasts directly with views that gesturing is produced to aid speech production, suggesting instead that gestures carry meaning and can be used to compensate for difficulty in speech.

It is important to note that whether speech and gesture go hand-in-hand or trade off against one another is a separate issue from the question of whether gesture facilitates speech planning. Regarding the former, the question is really just about the direction of the relationship between speech and gesture: When people use one modality more, does their use of the other modality increase (hand-in-hand) or decrease (trade off)? To investigate these two possibilities, several studies have measured gesture production while manipulating the effort required to formulate spoken descriptions: In situations where verbal referring is more difficult, the trade off account predicts more gesture and less speech. One such study by De Ruiter (1998a) found that rates of gesture relative to speech did not change depending on whether speakers were describing simple arrangements of shapes and vertical/horizontal lines, or random arrangements of shapes and diagonal lines. Although this finding contrasts with predictions of a trade off account, as Morsella and Krauss (2004) note, De Ruiter's manipulation varied stimulus complexity and not *describability*. In an experiment designed to tease apart these two attributes, Morsella and Krauss (2004) concluded that while visual complexity did not affect gesture rates, verbal codability did, with participants producing higher rates of gesturing for harder-to-name pictures (squiggles) than easy-to-name pictures (familiar objects). However, Morsella and Krauss's study measured only the proportion of time in a trial that participants spent gesturing (rather than

in relation to the time spent speaking) thus failing to capture the relationship between gesture and speech.

More recently, De Ruiter, Bangerter, and Dings's (2012) study found evidence which patterned with an account of speech and gesture going hand-in-hand: Participants' uses of pointing and iconic gestures were associated with the number of locative and feature descriptions respectively. Moreover, De Ruiter et al. found little evidence suggesting that gestures varied with experimental manipulations aimed to make speaking more difficult: Rates of iconic gestures were unchanged depending upon whether referring to simple, humanoid or abstract Tangrams, or whether referring to a novel or a repeated referent. De Ruiter et al. found that verbal codability affected participants' use of speech (both the lengths of descriptions and the times taken to initiate speech), but not their use of gesture. This contrasts with the idea that gesture facilitates the planning and production of speech, which would predict more gesturing when formulating spoken descriptions is more difficult (i.e., greater increases in gesturing when referring to abstract Tangrams or novel referents).

One reason for the differing findings between studies may lie in how gesture, and speech, have been measured. Unlike speech—where distinct phonemes and words offer comparatively clear means of measuring utterance length and duration—multiple pieces of information (and multiple types of gesture) may be produced in the time between the raising and lowering of hands. Tending towards *rate* measures (i.e., gesture proportional to speech), the predominant strategy of previous studies has involved measuring the number of discrete gestures produced when speaking (e.g., De Ruiter et al., 2012; Gerwing & Allison, 2011; Hoetjes, Koolen, Goudbeek, Krahmer, & Swerts, 2015; Hostetter et al., 2007b). However, in the literature, the definition of a gesture has varied widely, from “illustrat[ing] a particular feature of the target (e.g., shape)” when describing Tangram figures (De Ruiter et al., 2012, p. 238) to change in any one of “shape and placement of the hand, trajectory of the

motion” when identifying referents in a narrative (So et al., 2009, p. 118). There is further variation in the denominator of these rate measures: Counts of discrete gestures according to varying criteria have then been averaged per trial (Morsella & Krauss, 2004), per minute (Mol et al., 2011), per 100 words (Gerwing & Allison, 2011; Hoetjes et al., 2015; Hostetter et al., 2007b; Masson-Carro, Goudbeek, & Krahmer, 2015), per *feature description* (De Ruiter et al., 2012) or per *semantic attribute* (Hoetjes et al., 2015).

A second explanation of why findings differ may be that much of the research into gestures has involved studies in which a single speaker’s gestures are evaluated under various conditions. Experimental paradigms have tended towards those in which participants produce speech and gesture either to an imagined future addressee (e.g., Morsella & Krauss, 2004; Wesp et al., 2001) or to an addressee who is present but in a comparatively passive role (e.g., Bangerter, 2004; De Ruiter et al., 2012; Hoetjes et al., 2015; Holler & Stevens, 2007). Both of these designs may fail to capture the dynamic process of conversation; and it may be that participants view the task demands very differently from those of natural dialogue. The production of speech during dialogue is often considered to be an element of a joint activity (Clark, 1996), and the production of gesture is no exception to this; gesture production increases during dialogue, even when the interlocutor is visually occluded (Bavelas, Gerwing, Sutton, & Prevost, 2008).

The present study re-examines the relationship between speech and gesture, using a director-matcher paradigm in which pairs of participants alternate roles thereby each making conversational contributions in describing visual images which their partner has to match from an array of presented possibilities. To avoid the complexities involved with defining a single ‘gesture’ (and to some extent, a ‘word’), we propose a duration-based approach in which the relative durations of speech and gesture are measured directly. By recording the durations for which a speaker conveys (or attempts to convey) information via different channels, we aim

to establish whether, and how, speech and gesture co-vary. Images are either easy or difficult-to-name, rendering conceptual planning of each utterance relatively easy (or difficult), and participants encounter these images in multiple trials as both speaker and a listener.

If speech and gesture go ‘hand-in-hand’ we expect there to be a positive relationship between speech and gesture duration, with an inverse relationship indicating a trade off between modalities. Moreover, if gesturing is in part produced to support the production of speech (through lexical facilitation, through supporting imagery, or by packaging information) this relationship should be affected by conceptual difficulty (with more gesturing relative to speech for images which are difficult-to-name relative to those which are easy-to-name). Additional analyses (Section 3.3) will explore how conceptual difficulty of referents influences the production of other types of gestures, of disfluencies in both speech and in gesture, and of the relative timings of the two modalities.

3.2 Experiment 3.1

To distinguish hand-in-hand vs. trade-off accounts of the relationship between speech and gesture, Experiment 3.1 manipulates the complexity of a set of objects that participants must describe. We measure the relative durations of iconic gesturing and spoken descriptions. A hand-in-hand account suggests a positive association between the two, whereas a trade-off account suggests an inverse relationship (as gesturing takes over some of the communicative load).

Pairs of participants engaged in a collaborative matching game, in which they were tasked with matching two target shapes seen by one participant from a set of six shapes seen by the other participant. The entire experiment was recorded by two cameras capturing audiovisual data of both participants. Shapes varied

in *nameability*—they were either easy or difficult-to-name, and participants took turns in the roles of *director* and *matcher*. Each shape was presented 4 times throughout the experiment, meaning that descriptions were elicited under varying conditions of familiarity of a shape for a speaker. For a given trial, a speaker could have: a) over two or more previous trials, both described the shape themselves and had the shape described to them; b) described the shape in a previous trial; c) had the shape described to them in a previous trial; or d) experienced no previous trial in which that shape was described by either participant. In a subset of trials in the second half of the experiment, familiarity was directly controlled such that the difficult-to-name shape was repeated in consecutive trials.

Referent nameability and familiarity allowed us to investigate how the relationship between durations of speech and gesture vary with the conceptual difficulty of generating a description. If gesturing is in part produced to support the production of speech, there should be more gesturing relative to speech for images which are difficult-to-name (relative to those which are easy-to-name) and which are new (relative to those which have been described/heard previously).

3.2.1 Participants

Forty-four participants were recruited from the University of Edinburgh student community, and took part in exchange for £7 each. Consent was obtained in accordance with the University of Edinburgh’s Psychology Research Ethics Committee guidelines (ref number: 110-1617/1).

3.2.2 Method

Stimuli

Eighty shapes were used as target shapes (shapes which the director was tasked with describing) in the experiment. This set consisted of 40 easy-to-name shapes and 40 difficult-to-name shapes. Easy-to-name shapes were two-dimensional geometric shapes, for example: circle, diamond, heart, star. Each of these shapes was edited to create a difficult-to-name variant, by rotating and/or mirroring a section of the shape (see Figure 3.1) such that the name wasn't a single lexical item. 20 of these shapes (10 easy-to-name and 10 difficult-to-name variants) were used as target shapes in critical trials. Forty shapes (20 easy-to-name and 20 difficult-to-name variants) were used as target shapes in filler trials. A further set of 20 'matcher-only' shapes (10 easy-to-name and 10 difficult-to-name variants) were included in the matchers' arrays but were never seen as target shapes.

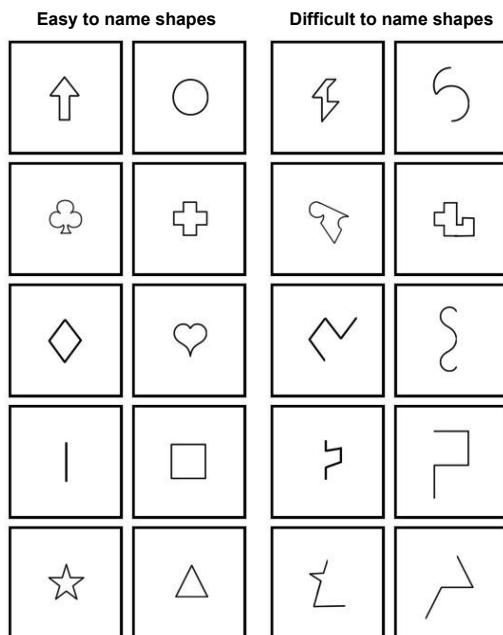


Figure 3.1: Shapes used in critical trials in Experiment 3.1

Experimental blocks

The experiment consisted of two blocks, each containing 40 trials (20 critical and 20 fillers). Critical trials always presented to the director one easy-to-name shape and one difficult-to-name shape (a difficult-to-name variant of a different easy-to-name shape). Filler trials presented either two difficult-to-name or two easy-to-name shapes.

In each experimental block, every critical shape was presented twice. In the first block, trials were randomly ordered, with the constraint that no shape was repeated in consecutive critical trials. Although the majority of the shapes were described once by each participant in the first block, the probability that a given shape was described twice by the same participant was 25%.¹

In the second block, each shape was described once by each participant. Pairs of consecutive critical trials alternated with pairs of filler trials, and the difficult-to-name shape was repeated in consecutive critical trials. This meant that for critical trials where participant B was the director, they were tasked with describing the difficult shape which had just been described to them in the previous trial by participant A. Between blocks, participants were given the option of a short break.

Procedure

Participants sat facing one another with an unobstructed space between them. Each participant had a monitor (with a resolution of 1280×1000) and a mouse on a table to their left, positioned such that they could not see what was on their partner's monitor. The set-up was designed to encourage face-to-face dialogue, and to discourage participants from leaving their hand resting on the mouse whilst

¹Whilst the intention was to make it so that each participant described each critical shape once, a typing error in the experiment script resulted in these distributions.

speaking (the position being uncomfortable for a right-handed mouse user), thus leaving both arms free to gesture. Audio and video was recorded by two cameras positioned to the right of each participant, facing their partner. Figure 3.2 shows an example set-up for a participant (a still from one of the cameras).



Figure 3.2: Example set-up of a participant in Experiment 3.1

Taking turns in the roles of director and matcher, participants were tasked with collaboratively matching the two shapes seen on the director's monitor from a set of six possibilities on the matcher's monitor. Participants were asked to restrict their communication to within a 10 second time-window during which no images were present on either screen. The aim of this was to encourage participants to look at their partners during communication, and not at their screens.

Figure 3.3 shows the procedure for a given trial. Each trial began with messages displayed on both participants' monitors informing them of their role for the trial (director or matcher), and that the experimenter would proceed to the task. Once the experimenter deemed both participants ready, the task began. While the

matcher's monitor remained blank, the two shapes (each measuring 325×325 pixels) were presented on the director's screen centered vertically and positioned horizontally such that the midpoint of each shape was 25% of screen-width in from the outer edges of the display. Shapes were randomly positioned (left vs. right) for each trial. After 2000 ms, the director's monitor went blank, and the 10 second communication window began. A 3 second countdown followed by the sound of a bell marked the end of this 10 second window and the time at which participants should stop communicating. After this window, on the matcher's screen a 2×3 array of shapes was displayed and the cursor was centered and made visible. This array included both target shapes for that trial, two randomly selected filler shapes matching the nameability of the target shapes, and two randomly selected from the remaining filler and 'matcher-only' shapes. Positions of the shapes in the array were randomly assigned for each trial. Using their mouse, matchers clicked on the two shapes they believed best matched the descriptions they had been given. Upon each mouse click on a shape, the shape was highlighted red. Once either matcher had clicked on two shapes or 10 seconds had elapsed, the array disappeared and feedback was displayed on both director and matcher screens.

As a pair, participants were awarded points for successful matching, scoring five points for each shape successfully matched. Feedback was given by a sound effect (buzzer, bell, or two bells for zero, one or both shapes matched) and the cumulative score displayed on-screen. To increase motivation, the highest scoring pair received £40, and participants were informed of this beforehand. A high-score table was shown prior to the experiment, and participants added their score to the table after they had played.

Gesturing was permitted but not explicitly encouraged, as participants were told that during this period, they were "both allowed to talk, gesture, ask questions, and so on". At the end of the experiment, participants were asked to complete a short questionnaire about their experience during the game. This questionnaire

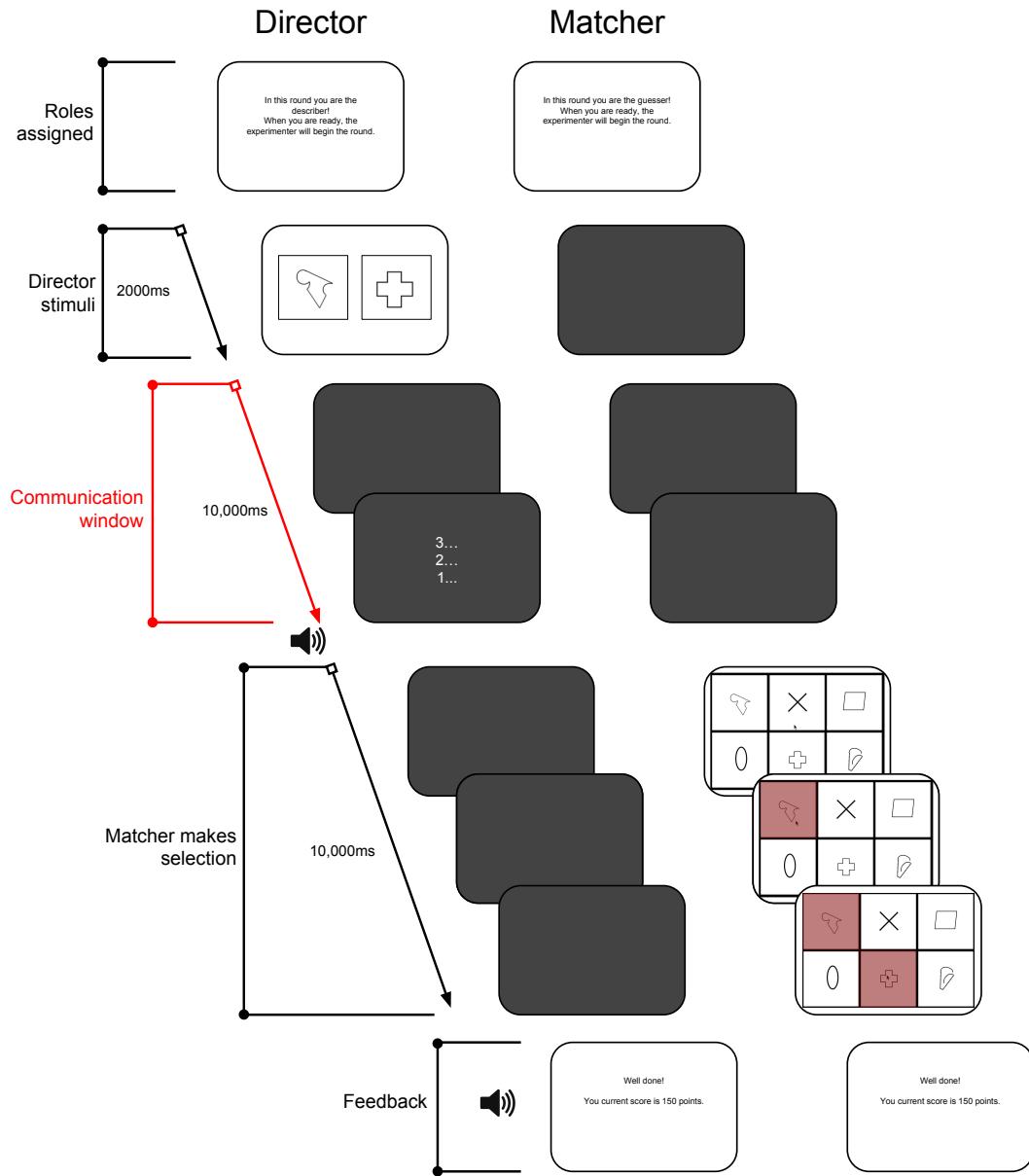


Figure 3.3: Procedure of a given trial in Experiment 3.1

included asking whether it had occurred to them during the experiment that the researchers might be studying their use of gesture. For positive responses to this question, a follow-up question asked them to rate how much they felt that this affected their behaviour during the experiment (1 = ‘Not at all’ to 7 = ‘A lot’).

3.2.3 Coding

Audiovisual data for each pair of participants was coded using a three stage process: Audio-only and video-only stages were used to code for speech and gesture respectively, with the third stage (both audio and video) used to confirm the annotations resulting from the previous stages. As each trial consisted of describing two shapes, there was potential for descriptions in both modalities to be interleaved. Special care was therefore taken in the third stage to ensure that utterances and gestures were assigned to the correct referents.

Speech

Utterance duration was coded in the audio-only stage. Only the first mention of each shape was used (if a director described each shape, then continued to describe the first shape again, this second description was excluded). Utterance duration (ms) was coded from the onset of the relevant referring noun-phrase up until either a) speech-offset, b) a new description (i.e., of the other shape) or c) a valid interruption from the listener in either modality (gestural interruptions were established in the audiovisual stage of coding). Listeners' use of the collateral channel (for instance: “yep”, “mmhm”, [nods head]) were not considered valid interruptions.

Gesture

Gesturing was identified in the video-only stage of the coding process, and was identified by the onset of any movement from the fingers up to the shoulder in either arm. To avoid gestures not related to the target shape, only movements which at least partially overlapped an identified utterance period were included,

and were assigned to the utterance which they primarily overlapped. This pairing was then confirmed in the audiovisual stage of the coding process.

Once identified, it was first established whether or not the movement constituted iconic gesturing. Any gesturing which was considered to be an attempt to represent any feature of the target shape was coded as iconic gesturing. For these gestures, the duration of iconic gesturing was measured analogously to the measure of utterance duration: Beginning at the onset of the first stroke or hold phase (excluding the initial preparation phase) up until either a) the retraction phase, b) iconic gesturing referring to a different shape (i.e., the other shape in the trial) or non-iconic gesturing, or c) a valid interruption. End-of-gesture hangs (uninformative hangs immediately prior to a retraction phase) were not included. We discerned here between end-of-gesture *hangs* (finger left hanging after tracing a shape) and end-of-gesture *holds* (hand in a specific position left hanging) which continued to convey some representational content, and were therefore included as part of gesture duration. This measure of gesture duration included any false starts, hangs, or preparation which occurred mid-gesturing, just as utterance duration included utterance-medial pauses and disfluencies.

The third stage of the coding process (audio and video) confirmed the annotations from the audio- and video- only stages, specifically categorisation of gesturing and pairing of gesturing with referents. It remained unclear as to whether some movements constituted iconic gesturing even after this third stage (in critical trials: 81 movements referencing easy shapes, and 19 referencing difficult shapes). These movements were considered to be imprecise/lax attempts at representing the shapes in space (appearing at first glance to be merely a rhythmic beat gesture emphasising spoken content), and were coded as iconic gesturing.² Additionally,

²This imbalance is in the opposite direction to the hypothesis (more gesturing for difficult-to-name shapes), meaning that the decision to code these movements as iconic gesturing is a conservative one.

this stage was used to code whether or not the utterance referred explicitly to the gesturing produced (e.g., “like this”, “like that”, “a bit here”, etc.)

3.2.4 Results

Forty-four participants in 22 pairs took part in the experiment. In the post-test questionnaire, 33 (75%) participants indicated that it had occurred to them that the researchers might be measuring their use of gesture. However, in these participants’ ratings of how much they felt that this had influenced their behaviour during the experiment (1 = ‘Not at all’ to 7 = ‘A lot’), 26 gave a rating of ≤ 3 , and only 1 gave a rating ≥ 6 .

Speakers’ familiarity with the shape in a given description was coded as falling into one of five categories: New to the experiment; Heard Previously (speaker has had this shape described to them before but has not described it themselves); Described Previously (speaker has described this shape before but not had it described to them); Heard and Described Previously (speaker has had to both describe this shape before, and has also previously had it described to them); and Heard in Previous Trial (speaker has had this shape described to them in the trial immediately preceding this one—the manipulation in Block 2). Table 3.1 shows a breakdown of these categories by both referent nameability and experimental block.

Only the critical trials were included in the analysis, numbering 880 (out of 1760) trials, or 1760 individual descriptions of shapes (2 per trial). Twelve (0.7%) descriptions were coded as missing—either due to participants running out of time, forgetting to describe a shape, or appearing to forget what shape they had seen—and were excluded from all analyses. Of the 1748 descriptions analysed, 1301

Table 3.1: Speaker familiarity with shapes across experimental blocks in Experiment 3.1. Numbers represent total numbers of observations

	Easy-to-name	Difficult-to-name
Block 1		
New	220	220
Heard Previously	99	106
Described Previously	118	110
Heard & Described Previously	0	0
Heard in Previous Trial	0	0
Missing	4	3
Block 2		
New	0	0
Heard Previously	117	53
Described Previously	62	64
Heard & Described Previously	244	118
Heard in Previous Trial	11	206
Missing	0	5

(74%) were accompanied by iconic gesturing (an attempt to gesturally represent the target shape).

Second rater

To assess the reliability of the annotating procedure, a subset of 20% (176) of critical trials was coded by an second annotator who was blind to the experimental aims (e.g., blind to the focus on the ratio between speech and gesture durations). The second annotator coded the presence and duration of iconic gesturing and the duration of the utterance. There was agreement in 93.4% of descriptions for the presence of iconic gesturing (Cohen's $k = 0.82$), and duration of iconic gesturing

and utterance duration both showed a high intraclass correlation (0.97 and 0.97 respectively).

Analysis

Analysis was carried out in R version 3.5.2 (R Core Team, 2018), using the lme4 package version 1.1-17 (Bates, Mächler, Bolker, & Walker, 2015). The duration of iconic gesturing (Z-scored) was modelled using linear mixed effects regression bootstrapped ($B = 1000$ bootstrap samples) to account for non-constant variance in the error term. Fixed effects included utterance duration (Z-scored), referent nameability (Easy-to-name vs. Difficult-to-name, dummy coded with easy-to-name as the reference level), and familiarity (New; Heard; Described; Heard & Described; Heard in the previous trial, dummy coded with New as the reference level), and all interactions. By-participant random intercepts and random slopes for utterance duration, referent nameability and familiarity were included, along with by-shape random intercepts.

Durations of speech and gesture

Figure 3.4 shows the relationship between durations of speech and gesture by referent nameability, and Figure 3.5 shows the relationship split by the speaker’s familiarity with the shape. Durations of speech and gesture co-varied: For descriptions of easy-to-name shapes which were new to the speaker, as the duration of speech increased, so did the duration of iconic gesturing ($\beta = 0.59$, (Bootstrapped 1000 samples) 95% CI [0.47, 0.72]). Importantly, a greater increase of gesture relative to speech was found for descriptions of difficult-to-name shapes (as indicated by the interaction between referent nameability and utterance duration; $\beta = 0.26$, (Bootstrapped 1000 samples) 95% CI [0.13, 0.38]).

Relative to the first appearance of shapes in the experiment, only having both heard *and* described an easy shape earlier in the experiment was associated with a reduction in the duration of iconic gesturing relative to that of speech ($\beta = -0.26$, (Bootstrapped 1000 samples) 95% CI [-0.46,-0.07]). This was modulated by the nameability of the shape ($\beta = 0.24$, (Bootstrapped 1000 samples) 95% CI [0.06, 0.49]), with previously heard and described difficult-to-name shapes resulting in less reduction of gesturing (relative to new shapes) than previously heard and described easy shapes. Full results are shown in Table 3.2.

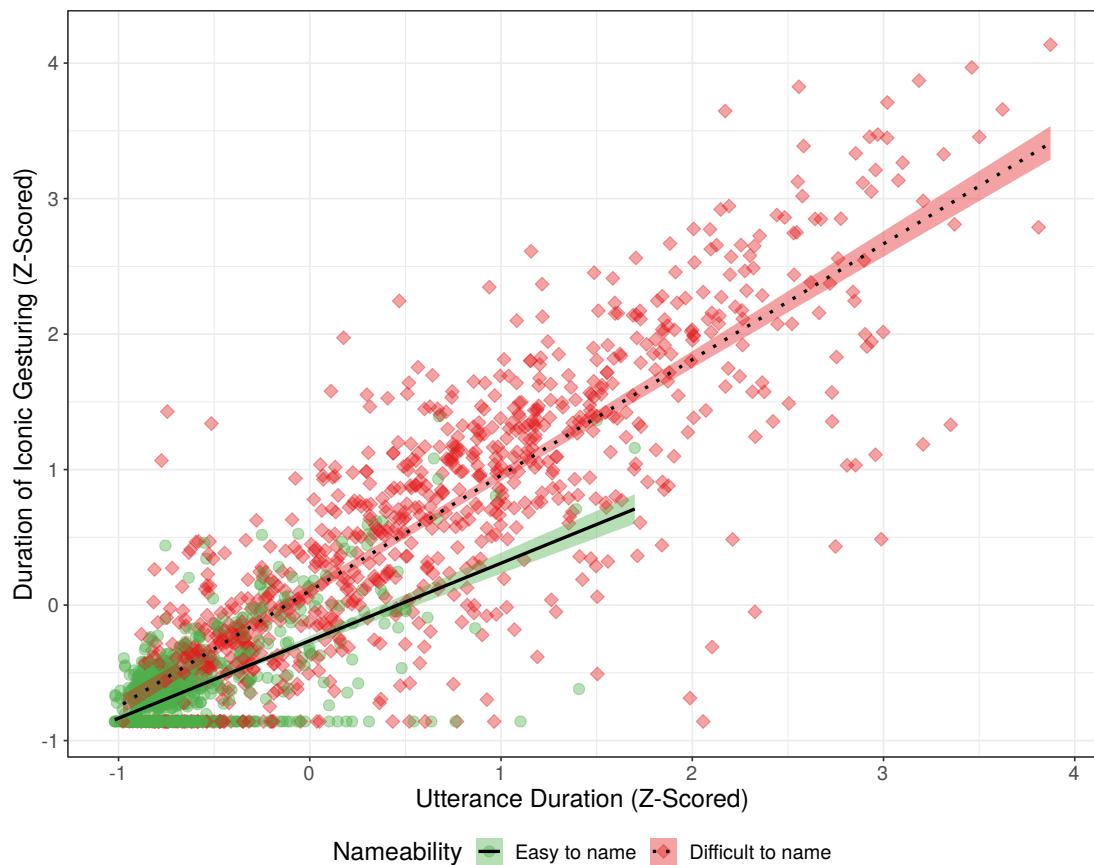


Figure 3.4: Relative durations of speech and iconic gesturing in Experiment 3.1

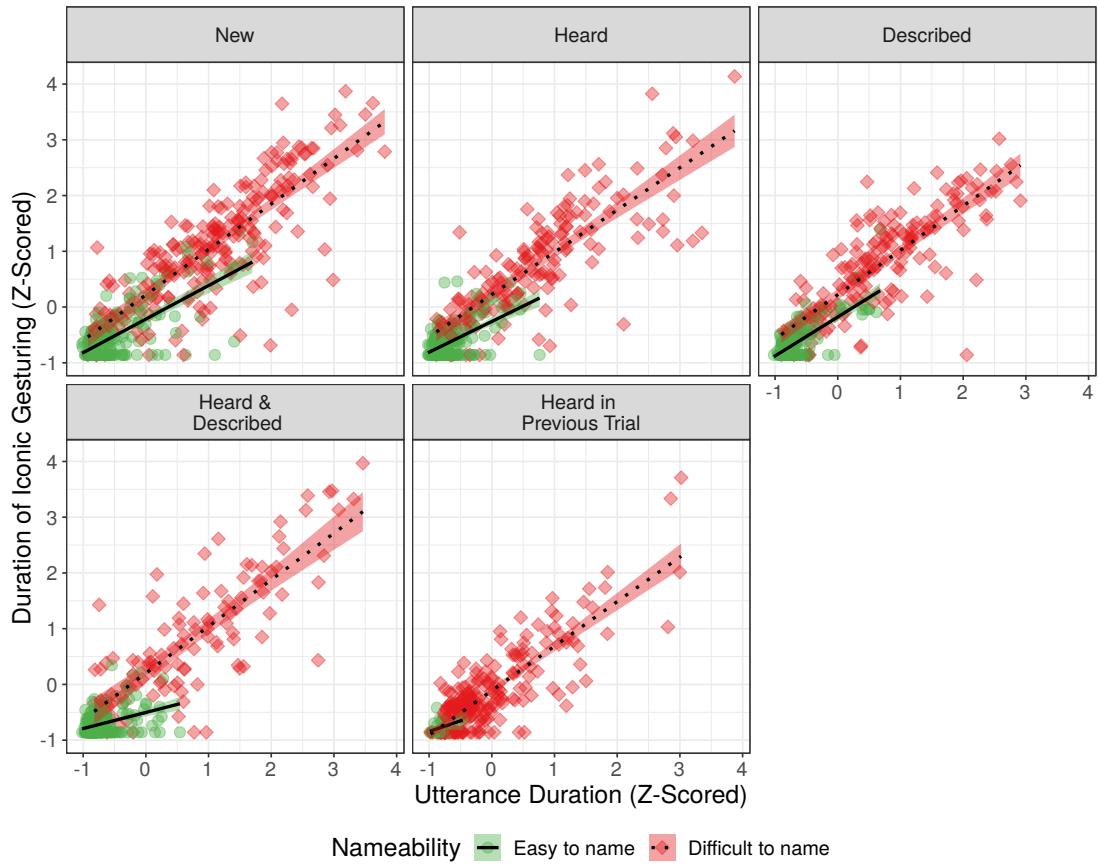


Figure 3.5: Relative durations of speech and iconic gesturing by familiarity of shape in Experiment 3.1

3.2.5 Discussion

The experiment presented here investigated the speech-gesture relationship by directly measuring the relative durations of either modality. In a collaborative shape-matching game, participants took turns to describe and match shapes which were either easy-to-name or difficult-to-name. Shapes were repeated across the experiment, and for a given description a speaker could be referring to a shape which had not been seen before, which had been described to them, which they had described previously, or both.

Results indicate that speech and gesture go ‘hand-in-hand’: The more speech

Table 3.2: Analysis of the duration of iconic gesturing (Z-scored) in Experiment 3.1. Model coefficients are reported alongside bootstrapped (1000 samples) 95% Confidence Intervals and average estimates

	β	Mean estimate	95% CI
(Intercept)	-0.22	-0.22	[-0.33, -0.11]
Utterance Duration (Z-scored)	0.59	0.59	[0.47, 0.72]
Difficulty to Name	0.40	0.40	[0.26, 0.54]
Heard Previously	-0.01	-0.01	[-0.15, 0.14]
Described Previously	-0.05	-0.04	[-0.2, 0.11]
Heard & Described Previously	-0.25	-0.25	[-0.42, -0.08]
Heard in Previous Trial	-0.33	-0.32	[-1.54, 0.96]
Utterance Duration \times Difficulty to Name	0.26	0.25	[0.13, 0.38]
Utterance Duration \times Heard Previously	0.00	-0.01	[-0.2, 0.2]
Utterance Duration \times Described Previously	-0.01	-0.01	[-0.21, 0.2]
Utterance Duration \times Heard & Described Previously	-0.26	-0.27	[-0.46, -0.07]
Utterance Duration \times Heard in Previous Trial	-0.34	-0.34	[-1.79, 1.13]
Difficulty to Name \times Heard Previously	0.01	0.02	[-0.16, 0.19]
Difficulty to Name \times Described Previously	0.11	0.11	[-0.07, 0.29]
Difficulty to Name \times Heard & Described Previously	0.21	0.22	[0.01, 0.41]
Utterance Duration \times Difficulty to Name \times Heard Previously	0.03	0.02	[-1.26, 1.25]
Utterance Duration \times Difficulty to Name \times Described Previously	0.00	0.00	[-0.21, 0.22]
Utterance Duration \times Difficulty to Name \times Heard & Described Previously	-0.03	-0.03	[-0.25, 0.18]
Utterance Duration \times Difficulty to Name \times Heard & Described Previously in Previous Trial	0.28	0.29	[0.06, 0.49]
Utterance Duration \times Difficulty to Name \times Heard & Described Previously in Previous Trial	0.24	0.23	[-1.27, 1.69]

participants produced, the longer they spent gesturing. This directly contrasts with a trade-off account, which predicts that gesture ‘takes over’ from speech (i.e., greater durations of gesturing should be associated with shorter utterances). Moreover, our findings indicate that not only do speech and gesture co-vary, but their relationship depends on conceptual load (contrasting with findings of De Ruiter et al. 2012, and supporting those of Hostetter et al. 2007b). In descriptions of shapes for which the conceptual planning of an utterance was comparatively

difficult (shapes which were difficult-to-name), descriptions resulted in greater durations of iconic gesturing relative to speech. This increase in gesturing could be taken as support for the idea that gestures facilitate speech planning. However, there is an alternative explanation: Participants may have gestured more for these shapes in order to communicate more effectively, perhaps due to less confidence in their verbal descriptions of these shapes. Such an account is fundamentally a more nuanced form of the trade-off theory: Speakers may put more effort into gesturing for less describable shapes not because it helps them to formulate verbal descriptions but to compensate for lack of specificity in the accompanying speech.

This explanation is supported by the finding that speakers' familiarity with shapes had little effect on their use of speech and gesture: Only after having previously both heard and described a shape did participants tend to reduce the amount of gesturing produced (and with this reduction being smaller for shapes which were more difficult to describe). Had participants' gestures been in aid of planning speech, we might expect this reduction to emerge when describing any shape not novel to the experiment. Instead, this result may indicate that only once interlocutors have engaged in grounding (see Clark, 1996) do they put less effort into producing multi-modal expressions.

At present, we cannot say for sure whether the higher rates of gesturing relative to speech associated with more difficult-to-name shapes is due to gesture facilitating speech production (as in Hostetter et al., 2007b) or gesture being used to compensate for poor verbal descriptions (Melinguer & Levelt, 2004). To distinguish between the two accounts requires investigating more than just rates or durations of gestures relative to the amount of speech but also the relative distribution of information in either modality.

The study presented here offers a step towards a more naturalistic investigation of speech and gesture, with two dialogically involved interlocutors in a face to face

setting. We explored how speakers distribute effort into speech and gesture (as captured in their relative durations) when referring to shapes which are either easy or hard to describe. Findings patterned with previous research suggesting that the two modalities go hand-in-hand: More speech is associated with more gesture. Additionally, we found gesturing to increase with conceptual load at a higher rate than speech. However, we highlight that to discern between whether gestures are produced to facilitate the production of speech—or whether they are produced with the intention to communicate and compensate for underspecification in speech—requires more than investigating the counts or durations of words and gestures.

3.3 Additional exploratory research

The audiovisual data captured from Experiment 3.1 offers many other possibilities for investigating how interlocutors produce speech and gesture when faced with referents differing in their relative nameability. Due to the low number of incorrect responses made by matchers—38 (4.5%) of descriptions of difficult-to-name shapes and 24 (2.8%) of descriptions of easy-to-name shapes—it was not possible to investigate how directors' use of speech and gesture influenced matchers' comprehension. However, in addition to measuring durations of speech and gesture, the annotation process also captured other aspects of production in both modalities. This section presents a selection of descriptive statistics and exploratory analyses, including an analysis of the production of different types of gestures, and the relative fluency and synchrony of speech and gesture. It is important to note that many of the variables included in Section 3.3 have not been second-coded, nor were they the intended focus of the experimental design.

First, we describe some of the features of iconic gestures used by participants in

Experiment 3.1 (specifically the number of hands used, and whether the gesture comprised static holds, dynamic strokes, or both). Secondly, we explore how speakers used other types of gestures when producing descriptions of easy- and difficult-to-name shapes. Although the majority of previous research has tended to focus on how iconic gesturing varies with conceptual demands, some studies have also measured the rates of rhythmic beat gestures and pointing gestures. For instance, Hostetter et al. (2007b) found no influence of conceptual demand on speakers' rates of beat gestures, and participants in De Ruiter et al.'s (2012) study did not differ in their production of pointing gestures dependent on how easy to verbally encode a shape was, nor whether a shape was novel or repeated. Along with beat gestures and pointing gestures, we explore how participants' use of adaptors (self-adaptive touching movements), and other gestures vary with conceptual demand (categories of gestures are defined below).

Thirdly, we investigate how the use of gesture relates to disfluency in speech. Previously, researchers have tended to study this relationship in one of two ways. The first approach has involved restricting participants' ability to gesture: Higher rates of disfluency have been found when participants are unable to gesture compared to when they are free to move (see, e.g., Finlayson et al., 2003; Rauscher et al., 1996). An alternative approach has been to measure the relative occurrences of gestures and disfluencies in naturally occurring speech: For example, Christenfeld, Schachter, and Bilous (1991) recorded (in real-time at the back of a lecture theatre!) the numbers of gestures and filled pauses produced by 31 speakers, finding an inverse relationship between the two (more gesturing being associated with less frequent filled pauses). Although these studies suggest that more gesturing is associated with less speech disfluency, other studies, (for instance, Hoetjes et al., 2014), have found no evidence that gesture has an effect on either fluency (or monotony) of speech. We assess whether the fluency of participants'

verbal descriptions of easy- and difficult-to-name shapes is influenced by their use of gesture.

Following this, we discuss how gestures can also exhibit disfluency, and investigate how this patterns with the fluency of accompanying speech. Compared to the phenomenon of disfluency in speech, gestural disfluencies have received little attention, and the taxonomies of types of disfluency in gesture are less well defined. A small number of studies have found evidence suggesting that pauses in gesture align with pauses in speech (see e.g., Esposito et al., 2001; Mayberry & Jaques, 2000), and that speech repairs which alter content (e.g., “You can [carry them both on]_{reparandum} - [tow them both on]_{repair} the same engine”) on are correlated with similar patterns of modification in gestures (Chen, Harper, & Quek, 2002). We identify possible forms of gestural disfluency, and ask whether they pattern with the occurrence of disfluency in speech.

Finally, we turn to the synchrony of speech and gesture. Research suggests that the duration by which the initiation of a gesture precedes the onset of its lexical affiliate (word or phrase related to and accompanying a given gesture) is inversely related to the familiarity of the lexical affiliate (Morrel-Samuels & Krauss, 1992). Many studies have investigated the synchrony of production between modalities (for an overview, see Wagner, Malisz, & Kopp, 2014), finding greater asynchrony between speech and gesture when, for instance, gestures convey unique information (relative to redundant gestures, see Bergmann, Aksu, & Kopp, 2011), and when movements involve actions on objects (as opposed to pantomimed gestures, see Church, Kelly, & Holcombe, 2014). In the present study, measuring the durations of utterances and gestures means that their relative onsets at which speakers begin iconic gestures relative to the verbal descriptions can be easily calculated. We ask whether, in keeping with previous research (Morrel-Samuels & Krauss, 1992), gestures precede accompanying speech by durations which are proportional to the ease with which a referent can be named.

Iconic gestures: Holds & traces

The third (audiovisual) stage of the annotation procedure also captured two features of the iconic gestures produced by participants: Whether they used one or both hands to describe an object, and whether the representational part of the gesture was conveyed dynamically (strokes and tracing movements), statically (holds), or as a combination of both. Table 3.3 shows the breakdown of number of hands used and the gestural forms produced in iconic gestures by whether shapes were easy or difficult-to-name. We note that the set of difficult-to-name shapes contained more un-closed shapes with a definite start and end point (lending themselves to being traced in the air, often requiring only one hand), and easy shapes predominantly involving closed shapes (lending themselves to more static holds involving two hands)—see Figure 3.6.

Table 3.3: Breakdown of form (static, dynamic, a combination) and number of hands used in the production of iconic gestures in Experiment 3.1, split by nameability of shape

	Easy to name	Difficult to name
Number of hands		
1	235 (52.2%)	498 (38.4%)
2	257 (47.8%)	311 (61.6%)
Gestural Form		
Static Holds	172 (35.0%)	41 (5.1%)
Dynamic Strokes	282 (57.3%)	522 (64.5%)
Combination	38 (7.7%)	246 (30.4%)

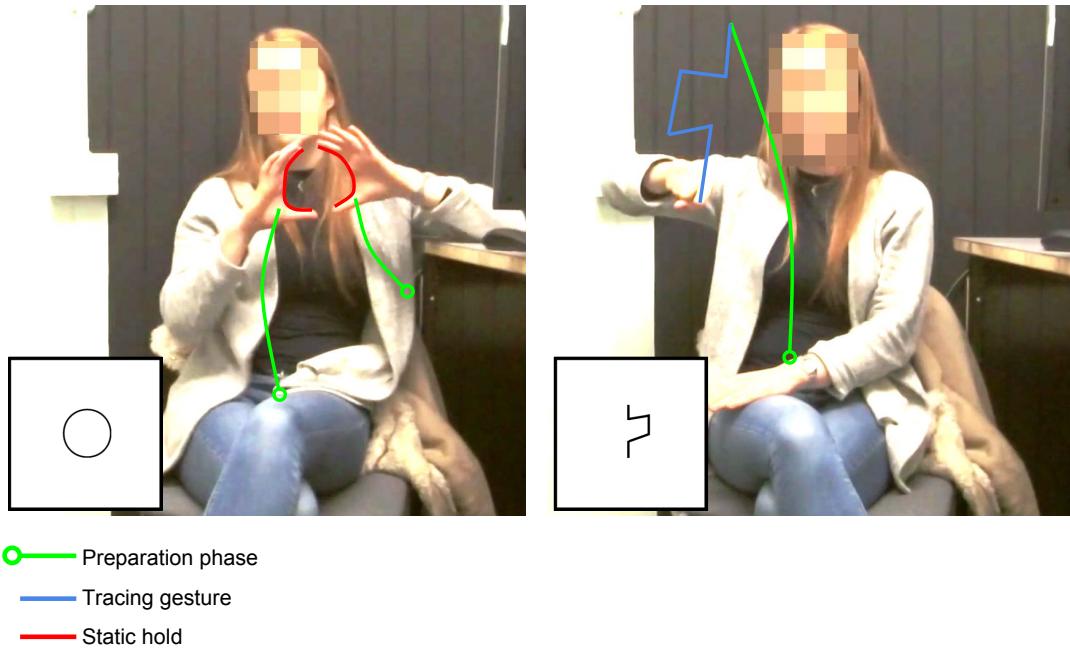


Figure 3.6: Example gestures from Experiment 3.1: A static hold representing an easy to name shape (left) and a tracing movement representing a difficult to name one (right)

Non-Iconic gestures

During the third (audiovisual) stage of the annotation process, any movements not coded as iconic gesturing were categorised as one of: Beats; Points; Adaptors; and Others. All of these required movement of either arm from the fingers up to the shoulder which partially overlapped an identified utterance period. Beat gestures were identified as movements which rhythmically matched prosody in speech but which *did not* represent any feature of the target shape. Extensions of the index finger or hand used to refer deictically to either present objects or people, as well as to previous parts of the discourse, were coded as point gestures. Movements and touching behaviours directed towards the self or objects (e.g., scratching, stroking, manipulating clothing) were categorised as adaptors. All other movements fell in to the category of other miscellaneous gesticulating. We

note that this category may be broad, and including, for instance, some of the movements which have previously been termed *interactive* (or *discourse-related*) gestures (see Bavelas et al., 1992)—movements such as shrugs which may refer to issues between interlocutors—which may be more prevalent in descriptions of difficult-to-name shapes than easy-to-name ones. In any cases where gesturing was ambiguous as to which utterance it accompanied (i.e., non-representational gesturing which overlapped both utterance periods), it was assigned to both utterances. This was the case for 22 gestures (20 Adaptor gestures, 1 Beat gesture, 1 Other gesture).

Table 3.4 shows the proportion of descriptions of easy-to-name and difficult-to-name shapes in which each type of gesture occurred. The occurrence of each type of gesturing in a trial in each trial was modelled using mixed effects logistic regression with referent nameability (easy-to-name vs. difficult-to-name, deviation coded) and utterance duration (Z-scored) as fixed effects and by-participant random intercepts. We note that these equate to a set of non-independent tests, inasmuch as the occurrence of one type of gesturing decreases the likelihood of other types, and so findings should be taken with caution.

Results (see Table 3.5) revealed that when describing shapes which were more difficult to name, speakers were more likely to produce point gestures ($\beta = 1.65$, $SE = 0.43$, $p < .001$) and less likely to produce beat gestures ($\beta = -0.63$, $SE = 0.23$, $p = .007$). We reasoned that the association between referent nameability and use of pointing gestures may have been driven by the increased opportunities to refer to recent discourse when describing difficult-to-name shapes in Block 2 (in which difficult-to-name shapes were repeated in consecutive trials). This was supported by the association disappearing when the subset of 245 (14%) descriptions of shapes seen in the previous trial were removed ($\beta = 0.32$, $SE = 0.67$, $p = .63$). The production of other types of gesture (adaptor gestures

and miscellaneous other movements) were not found to be associated with the nameability of the shape being described.

It is important to note that types of gesture are not always clear-cut: The numbers of beat and iconic gestures between descriptions of easy-to-name and difficult-to-name shapes in the present study may reflect the gradation between types of gesturing—i.e., as spoken descriptions become more clear (in that they have a common name), speakers' movements may decrease in representational specificity in favour of adding emphasis to speech. This patterns with the movements identified during the annotation procedure which were eventually coded as imprecise iconic gestures (81 referencing easy-to-name, and 19 difficult-to-name).

Table 3.4: Proportion of descriptions of easy-to-name and difficult-to-name shapes in which iconic, point, beat, adaptor or other gestures occurred in Experiment 3.1

	Easy-to-name	Difficult-to-name
Iconic Gestures	492 (56.2%)	809 (92.4%)
Point Gestures	10 (1.1%)	37 (4.2%)
Beat Gestures	103 (11.8%)	60 (6.8%)
Adaptors	62 (7.1%)	57 (6.5%)
Other Gestures	19 (2.2%)	35 (4.0%)

Gesturing and fluency of speech

All utterances were transcribed from the onset of the noun phrase up until offset of the description of that shape (or listener interruption—i.e., the same period used for the measure of utterance duration). Disfluencies within this period were identified as falling into one of six categories: filled pauses; insertions; deletions; substitutions; articulation errors; and repetitions (see Shriberg, 1996). The total number of words (excluding disfluencies) used in this period provided a measure of utterance length. Because of the nature of the experimental task, care was taken to discern between speech that was disfluent and speech which was intentionally

Table 3.5: Model results for numbers of different types of gestures, Experiment 3.1

	β	SE	p
Beat gestures			
(Intercept)	-2.52	(0.15)	<.001
Difficulty to Name	-0.63	(0.23)	.007
Utterance Duration	0.01	(0.12)	.99
Var(1—Participant)	0.55		
Point gestures			
(Intercept)	-4.79	(0.45)	<.001
Difficulty to Name	1.65	(0.43)	<.001
Utterance Duration	-0.21	(0.19)	.29
Var(1—Participant)	2.40		
Adaptor gestures			
Intercept)	-3.30	(0.27)	<.001
Difficulty to Name	-0.35	(0.27)	.19
Utterance Duration	0.19	(0.13)	.15
Var(1—Participant)	1.74		
Other gestures			
(Intercept)	-4.04	(0.29)	<.001
Difficulty to Name	-0.29	(0.40)	.47
Utterance Duration	0.62	(0.16)	<.001
Var(1—Participant)	1.03		

repetitive speech (e.g., “bits here, here, here and here”). Several participants tended to make noises accompanying strokes of gestures (e.g., “a shape like this: dun [stroke], dun [stroke], dun [stroke]”). These verbalisations were not coded as disfluencies, nor was their repetition considered to be disfluency (each one aligning with a different, meaningful part of a gesture).

Fluency of speech (fluent vs. disfluent) was modelled using logistic mixed effects regression with fixed effects of utterance length (Z-scored), Iconic gesturing (yes vs. no, deviation coded), referent nameability, (easy-to-name vs. difficult-to-name, deviation coded) and all interactions, and by-participant and by-shape random intercepts. Results revealed a main effect of utterance length ($\beta = 1.93$, SE = 0.31, $p < .001$): Longer verbal descriptions tended to be more disfluent. An interaction between iconic gesturing and utterance length ($\beta = -1.41$, SE = 0.61, $p = .02$) suggests that this increased likelihood for longer utterances to be disfluent is

reduced when speakers produce iconic gesturing. Full model results are shown in Table 3.6.

Table 3.6: Analysis of disfluency of spoken descriptions produced in Experiment 3.1

	β	SE	p
(Intercept)	-1.99	(0.23)	<.001
Difficulty to Name	-0.05	(0.40)	.90
Iconic Gesturing	0.69	(0.39)	.07
Number of Words (Z-Scored)	1.93	(0.31)	<.001
Difficulty to Name × Iconic Gesturing	0.62	(0.77)	.42
Difficulty to Name × Number of Words	0.10	(0.62)	.87
Iconic Gesturing × Number of Words	-1.41	(0.61)	.02
Difficulty to Name × Iconic Gesturing × Number of Words	-2.20	(1.23)	.07
Var(1—Participant)	0.51		
Var(1—Shape)	0.06		
Total	1748		
Participant	44		
Shape	20		

Fluency of gesture and fluency of speech

In the present study, two aspects of gestural fluency were recorded in iconic gestures during the second (video-only) stage of the annotation process: *content repetitions* and *false starts*. Content repetitions were identified as any repetition of gestural content—either the repetition or reversal of a gestural stroke, or static holds in which the hands maintain a given shape (e.g. a circle), but rhythmically beat. These repetitions often appeared to be deliberate and effortful movements which were likely intended to communicate, and in the third stage of the annotation process it became clear that they often coincided with apparently deliberate repetitions in speech (e.g., “a curve, and then a sharp bit and then a smaller bit [traces outline of shape], so curve, sharp, small [repeats tracing gesture]”).

Gestural false starts involved the trajectory of a gesture being immediately repeated

with some form of content modification or extension (e.g., a stroke which is restarted and subsequently lengthened, or produced at a different angle). These were often identifiable by a sudden change in velocity mid-gesturing as the hand returned to either the starting point or the last vertex, after which the gesture was completed (see Figure 3.7).

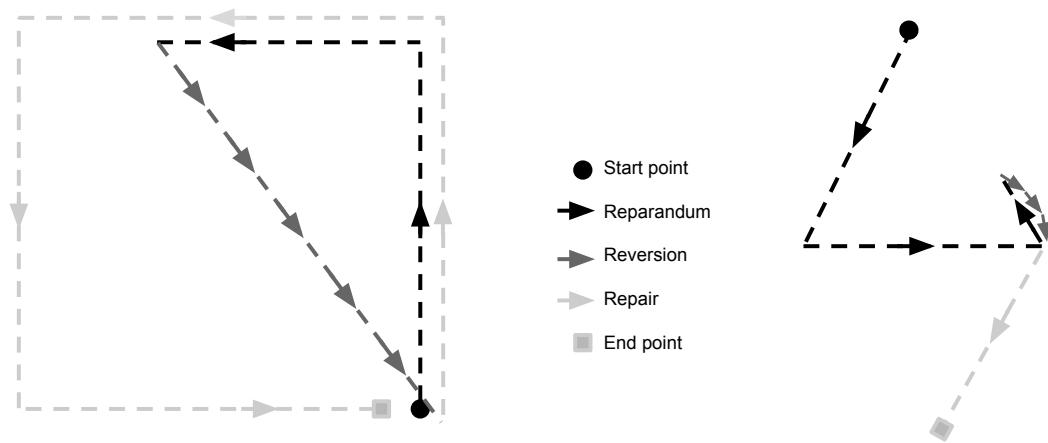


Figure 3.7: Trajectory of a gestural false-start. Frequency of arrows indicates relative velocity of gestural strokes

Trials were coded for the presence or absence of each of these gestural disfluencies. Table 3.7 shows the numbers of gestures in which content repetitions and false-starts were recorded, split by the nameability of the shape being described. Because of the high number of apparently intentional repetitions, we did not investigate the relationship between gestural repetitions and speech fluency, instead focussing on false starts in gesture. However, we note that it is possible that an increased likelihood of intentionally repeating gestural content for difficult-to-name shapes which is driving the differences in relative durations of gesturing and speech found in Section 3.2.4.

The presence of false starts in gestures was modelled using logistic regression with fixed effects of fluency of speech (fluent vs. disfluent, deviation coded), iconic

Table 3.7: Proportions of gestures of easy-to-name and difficult-to-name shapes which included false starts and repetitions

	Easy to name	Difficult to name
False starts	5 (0.6%)	95 (10.8%)
Repetitions	47 (5.4%)	299 (34.1%)

gesture duration (Z-scored), referent nameability (easy-to-name vs. difficult-to-name, deviation coded) and all interactions, and by-participant and by-shape random intercepts. Results revealed that false starts in iconic gestures were associated with disfluent utterances ($\beta = 2.19$, SE = 0.79, $p = .006$).

Table 3.8: Analysis of false starts in gestures produced in Experiment 3.1

	β	SE	p
(Intercept)	-3.37	(0.40)	<.001
Iconic Gesture Duration (Z-Scored)	0.14	(0.72)	.85
Disfluency	2.19	(0.79)	.006
Difficulty to Name	1.38	(0.80)	.08
Iconic Gesture Duration \times Disfluency	-0.93	(1.43)	.52
Iconic Gesture Duration \times Difficulty to Name	1.10	(1.43)	.44
Disfluency \times Difficulty to Name	-0.51	(1.59)	.75
Iconic Gesture Duration \times Disfluency \times Difficulty to Name	1.02	(2.87)	.72
Var(1—Participant)	0.00		
Var(1—Shape)	0.03		
Total	1301		
Participant	44		
Shape	20		

Synchrony of speech and gesture

Previous research into the synchrony of speech and gesture has involved measuring the onset of gesture as the onset of movement, taken to indicate the inception of an idea (McNeill, 1992), and thereby including the preparation phase. In the present study, such an approach would not be appropriate due to each trial involving

descriptions of a pair of shapes, between which speakers often did not return their hands to a resting position. We therefore capture the temporal synchrony of speech and gesture by measuring the time between the relative onsets of the noun phrase of a verbal description and of the first stroke or hold phase of iconic gesturing. Any of these durations (from onsets of verbal descriptions to onsets of gestural strokes/holds) falling more than three standard deviations from the mean were excluded from the analysis (31 of the 1301 descriptions involving iconic gesturing). The remaining 1270 durations were modelled using linear mixed effects regression with fixed effects of utterance duration (Z-scored), referent nameability (Easy-to-name vs. Difficult-to-name, deviation coded), and familiarity (New; Heard; Described; Heard & Described; Heard in the previous trial, dummy coded with New as the reference level), and all interactions. By-participant random intercepts and random slopes for utterance duration and referent nameability were included, along with by-shape random intercepts.

The onset of gesture tended to follow onset of the noun phrase in speech (as indicated by a positive intercept: $\beta = 293.16$, $SE = 48.14$, $t = 6.09$), with this duration increasing for longer utterances ($\beta = 216.08$, $SE = 48.14$, $t = 4.49$). The naming difficulty of a shape influenced the synchrony of speech and gesture, with gestures for difficult-to-name shapes beginning earlier (in relation to noun phrase onset in speech) than gestures of easy-to-name shapes ($\beta = -215.70$, $SE = 89.55$, $t = -2.41$), with this effect greater in longer utterances ($\beta = -223.73$, $SE = 91.45$, $t = -2.45$). Full model results are shown in Table 3.9.

These findings contrast with previous research suggesting that gestures tend to precede or coincide (but rarely follow) the words they are affiliated with (see e.g., Chui, 2005; Morrel-Samuels & Krauss, 1992). The details of how we measured both onset of speech and onset of gesture offers some explanation. Firstly, the duration by which gesture precedes speech may be due to the preparation phase of gestures, which was not taken into account in our measure of gesture onset.

Table 3.9: Experiment 3.1: Analysis of durations from onset of noun-phrase in speech to onset of first stroke or hold phase in iconic gesture (Z-Scored)

	β	SE	t
(Intercept)	293.16	(48.14)	6.09
Utterance Duration	216.08	(48.14)	4.49
Difficulty to Name	-215.70	(89.55)	-2.41
Heard Previously	-68.07	(64.89)	-1.05
Described Previously	11.22	(62.54)	0.18
Heard & Described Previously	60.23	(73.26)	0.82
Heard in Previous Trial	148.86	(516.88)	0.29
Utterance Duration \times Difficulty to Name	-223.73	(91.45)	-2.45
Utterance Duration \times Heard Previously	-58.74	(83.10)	-0.71
Utterance Duration \times Described Previously	-13.94	(80.18)	-0.17
Utterance Duration \times Heard & Described Previously	110.06	(91.19)	1.21
Utterance Duration \times Heard in Previous Trial	175.19	(643.74)	0.27
Difficulty to Name \times Heard Previously	75.23	(129.19)	0.58
Difficulty to Name \times Described Previously	133.49	(124.67)	1.07
Difficulty to Name \times Heard & Described Previously	-74.67	(146.94)	-0.51
Utterance Duration \times Difficulty to Name \times Heard Previously	198.89	(1033.81)	0.19
Utterance Duration \times Difficulty to Name \times Heard in Previous Trial	125.59	(166.99)	0.75
Utterance Duration \times Difficulty to Name \times Heard & Described Previously	84.85	(161.37)	0.53
Utterance Duration \times Difficulty to Name \times Heard in Previous Trial	-166.65	(182.18)	-0.92
Utterance Duration \times Difficulty to Name \times Heard in Previous Trial	179.81	(1287.26)	0.14
Var(residual)	222263.20		
Var(1—Participant)	23233.57		
Var(Utterance Duration—Participant)	11587.52		
Var(Difficulty to Name—Participant)	39739.71		
Var(1—Shape)	6514.93		
Total	1270		
Participant	44		
Shape	20		

Secondly, we measured gesture onset relative to onset of the noun phrase in speech, not the specific lexical affiliate of each gesture. It is possible that in many descriptions the part of speech corresponding to the gestural content appeared

later on (e.g., “a square with a [line going up here]_{lexical affiliate}”). This may also explain why durations from onset of noun phrase to onset of gesture increased with utterance duration: Utterances may vary in their length prior to the lexical affiliate of gestures.

Taking these considerations into account, care must be taken in how we interpret the finding that referent nameability influences the synchrony of speech and gesture. At face value, this result patterns with previous findings in the literature that less familiar words are associated with earlier onsets of corresponding gestures (Morrel-Samuels & Krauss, 1992)—the present study found less easily named shapes were associated with earlier onsets of gestures. However, this result would also be present if the lexical affiliates of gestures simply occur earlier in descriptions of difficult-to-name shapes than they do in descriptions of easy-to-name ones. For example, in our measure, the description “A [line with a square bit coming out]_{lexical affiliate}” will record an earlier onset of gesture than the description “A triangle which is [pointing up]_{lexical affiliate}”, even if the timing of gesture relative to lexical affiliate is the same in both. This may also account for why the influence of utterance duration in delaying gesture onset was reduced for difficult-to-name shapes.

3.4 Chapter discussion

The present chapter has investigated some of the ways in which speakers vary their production of gestures according to the conceptual demands of formulating spoken descriptions of shapes. In a collaborative matching game, participants took it in turns to describe and match shapes which could be either easy-to-name or difficult-to-name. The durations of both gestural and spoken components of participants’ descriptions were measured, with results patterning with previous

research suggesting that the two modalities go hand-in-hand: The more we speak, the more we gesture. Crucially, descriptions of difficult-to-name shapes resulted in greater durations of (iconic) gesturing relative to speech, adding to the body of evidence suggesting that the relationship between speech and gesture depends on conceptual load (Hostetter et al., 2007b, with more gesture relative to speech when conceptual demands are greater, as in). However whether this increase is due to the facilitatory effects of gesturing for speech production processes, or to compensation for underspecification in speech, remains unclear. Further research on how information is distributed between modalities is required to answer this question.

Along with the relative durations of speech and gesture, additional analyses point towards their relative onsets varying depending upon how easily named a shape was: Patterning with previous findings (Morrel-Samuels & Krauss, 1992), shapes which were more difficult-to-name elicited gestures which occurred earlier relative to the onset of the noun phrase in spoken descriptions. We noted that this may reflect descriptions of difficult-to-name shapes eliciting gestures which occur earlier relative to their lexical affiliates, but it may equally indicate that difficult-to-name shapes simply elicit gestures (and their related speech) at an earlier point in descriptions. This latter explanation is distinct from the relative synchrony of gestures with their lexical affiliates studied previously (e.g., Chui, 2005; Morrel-Samuels & Krauss, 1992). However, the finding remains an interesting one, suggesting that speakers enlist the use of gestures at an earlier point when referents are less easily encoded verbally.

Additional analyses relating to the fluency of both speech and gesture in Experiment 3.1 suggested that there was an association between disfluent speech and false-starts in gestures, in keeping with previous work on other suggested forms of gestural disfluency (Chen et al., 2002; Esposito et al., 2001; Mayberry & Jaques, 2000), although we note once more that these results are exploratory

and should be taken with caution. The longer an utterance was, the more likely it was to be disfluent. However, the production of iconic gesturing weakened this association, suggesting that the act of gesturing may aid speech planning, in keeping with Finlayson et al. (2003) and Rauscher et al. (1996).

A further finding from Experiment 3.1 was that descriptions of easy-to-name shapes resulted in more beat gesturing than those of difficult-to-name shapes. We suggest that this may have been due to an increase in beat gestures as speakers become more certain of their verbal descriptions, but we note that many of the iconic gestures produced in reference to difficult-to-name shapes also included rhythmic beats. Future work could investigate how speakers' production of these types of gestures change as referring expressions become grounded in dialogue—i.e., is there a move from effortful iconic gesturing and descriptive verbal descriptions to the use of beat gestures as conceptual pacts in speech increase?

In the broader context of this thesis, the present chapter supports the idea that the production of gesturing (specifically iconic gesturing) is dependent upon how easy a referent is to refer to in speech. The durations, and possibly onsets, of certain types of gesturing may therefore signal information about the difficulties incurred in producing speech. Turning to comprehension, this suggests that signals of speech planning difficulty may be available in the visual modality as well as the spoken one. Previous research has shown that speakers produce more speech disfluencies under greater cognitive load (e.g., Arnold et al., 2000; Barr, 2001; Beattie, 1979), and that listeners can draw on the presence of disfluency in speech to inform anticipations of the conceptual demands of upcoming referents (see Arnold et al., 2007, 2004; Barr & Seyfeddinipur, 2010; Corley et al., 2007). We suggest that the same may be true of non-linguistic behaviours presented in the visual modality.

Chapter 4

Gesture as a signal of conceptual demand

In the previous chapter, we investigated how the production of gestures relative to speech varies according to the conceptual demands required to describe an object. By measuring the occurrence and durations of gestures relative to speech, we established that when describing shapes for which the conceptual planning of an utterance was comparatively difficult (in that they did not have a familiar name), speakers tended to produce more and longer iconic gestures (relative to spoken descriptions) than they did when referring to familiar, easily named shapes. Furthermore, gestures representing more difficult-to-name shapes were found to precede the noun phrase of verbal descriptions by a greater duration than gestures of easily named shapes. The flip-side of these findings is that the occurrence and duration of certain types of gesturing—as well as the temporal asynchrony of these gestures with accompanying speech—may provide a listener with signals of upcoming message content.

We draw parallels her to Arnold et al.'s (2007) study, in which the presence of

disfluency in speech (“click on [the]/[thee - uh -] red …”) was found to bias listeners to predict that the speaker was about to mention a less familiar object (e.g., a squiggle as opposed to an ice-cream cone). Arnold et al. (2007) established that this *disfluency~unfamiliarity bias* (the bias toward anticipating unfamiliar referents following disfluency) was evident in the expectations which listeners held during the moment-to-moment processing of speech (evidenced by their eye and mouse movements alongside the unfolding utterance). Given speakers’ tendencies to produce more gesturing for referents which are more difficult-to-name (Chapter 3) or which require greater conceptual demands, it is not infeasible that listeners may treat the presence of iconic gesturing in a similar way. For instance, an utterance of “click on the ...” when accompanied by iconic gesturing may lead listeners to expect a referent which is more difficult-to-name. This possibility is the subject of the next two chapters. Chapter 4 investigates whether the presence of different types of gesturing influences listeners’ explicit predictions about upcoming message content. Following this, Chapter 5 asks the same of the expectations which listeners hold during the moment-to-moment processing of speech and gesture.

In the current chapter reports two experiments designed with the aim of exploring how the presence of different types of gesturing influences listeners’ predictions of upcoming referents. Experiments 4.1 and 4.2 present participants with a visual display comprising two objects and a video of a speaker. Each trial presents fragments of multi-modal (audio and video) instructions to click on one of the two objects, and tasks participants with clicking on the object they believe the speaker is about to refer to. In Experiment 4.1, we manipulate whether the video component of the instruction shows the speaker producing an iconic gesture or shows them sitting motionless. In critical trials, the two objects in the display comprise an easy-to-name shape (e.g., a letter, number or geometrical shape) and a difficult-to-name one (e.g., a squiggly shape), and the instructions are truncated

before the presentation of any information in either speech or gesture (when present) which distinguish between the objects in the display. This ensured that in making predictions about upcoming referents, participants would be responding to the occurrence of gesturing, and not its content (i.e., the full shape it represented). To investigate whether listeners are sensitive to the type of gesturing, and not the simple occurrence of movement, Experiment 4.2 included videos of the speaker either motionless or producing an adaptor gesture such as fidgeting or tapping.

In addition to capturing mouse clicks to objects in the display, participants' eye and mouse movements were recorded. The aim of this was to investigate the time course of any predictions which participants might make based on the presence of gesturing, with the possibility of developing the paradigm to study whether such predictions occur during the real-time processing of language.

Results revealed that following iconic gesturing—but not adaptor gesturing—participants were more likely to click on the less easily named shape compared to instructions presented with no gesturing. This suggests that listeners associate this specific representational form of gesturing with reference to objects which are more difficult to verbally encode. Furthermore, listeners' tendencies to predict mention of the more difficult-to-name shape (as indicated by mouse clicks to these objects) were influenced by the duration of iconic gesturing relative to speech: Longer (and therefore earlier) gestures resulted in more predictions of conceptually demanding content.

4.1 Signals of speech planning difficulty

Some things are easier to describe in words than others. The effort required to produce descriptions may vary in accordance with an object's complexity,

familiarity, or whether there is a clear framework for conceptualisation e.g., a geometric configuration for a series of otherwise unconnected dots (see, Hostetter et al., 2007b). As a consequence, the descriptions which speakers produce vary systematically in how they are delivered, both in speech and in gesture: Speakers tend to hesitate and produce more filled pauses such as “um” and “uh” when experiencing increased cognitive load (e.g., Beattie, 1979; Bortfeld, Leon, Bloom, Schober, & Brennan, 2001; Cook et al., 2009), and use more gestures when producing descriptions which involve greater conceptual demand (Hostetter et al. 2007b, Chapter 3), or producing less preferred syntactic structures (Cook et al., 2009).

Despite (and perhaps because of) speakers’ messages varying with respect to manner of spoken and non-verbal delivery, listeners are able in everyday communication to navigate and decode meaning from this complex input. Research has highlighted the efficiency of language comprehension, with evidence that listeners’ evaluation of referring expressions emerges within 200 ms of the auditory onset of a target word (see Allopenna et al., 1998). What is more, listeners appear able to use variations in the speech stream to their advantage, with features specific to spoken language such as emphasis (Dahan, Tanenhaus, & Chambers, 2002) and fluency (Bailey & Ferreira, 2003; Barr, 2001; Corley et al., 2007) influencing comprehension at the early stages of the comprehension process. One example of this is the perception of disfluency as a signal of upcoming referents. In two studies from Arnold et al., participants showed an initial tendency to fixate objects which are unfamiliar and discourse-new (rather than familiar or previously mentioned objects) more following disfluent speech than following fluent speech (Arnold et al., 2007, 2004). This result is reflective of the fact that disfluencies tend to occur before unpredictable, less familiar, and low frequency words (see e.g., Barr, 2001; Beattie, 1979; Schnadt & Corley, 2006). Listeners’ biases towards new information following a disfluency have since been replicated in a mouse-tracking

study, additionally finding the effect to be speaker-specific, and dependent upon what was new for a particular speaker, and not just what was new for the listener (Barr & Seyfeddinipur, 2010). The bias towards more difficult referents following disfluent speech has even been shown to occur for objects in an artificial lexicon, suggesting that listeners spontaneously infer what is difficult-to-name in a given situation (Heller, Arnold, Klein, & Tanenhaus, 2015).

In establishing the disfluency~unfamiliarity bias, Arnold et al. (2007) conducted both a gating experiment and an eye-tracking experiment. In the gating task, participants saw a familiar object and unfamiliar object each presented in two colours (e.g., a red ice-cream cone, a black ice-cream cone, a red squiggle and a black squiggle). Participants were presented with fragments of utterances which were truncated at various points and contained no information which disambiguated between the familiar and unfamiliar objects. Crucially, these utterances were either fluent or disfluent (e.g., “Click on [the]/[thee uh] red”). Participants were tasked with guessing which object a spoken instruction was about to refer to. Results revealed that participants tended to choose the less familiar objects following disfluent utterance fragments, and the more familiar objects following fluent ones, with this difference increasing with the length of fragment heard. Subsequently, Arnold et al.’s eye-tracking task, in which participants heard the full instructions (e.g., “Click on [the]/[thee uh] red <referent>”), found that listeners’ anticipatory fixations to objects were influenced by the fluency of the utterance: Following disfluency, from the onset of the color word participants showed a preference to fixate the unfamiliar color-matched object over the familiar one.

The common explanation of the mechanism underlying listeners’ disfluency-based expectations is that “um” and “uh” are treated as signals which indicate information about the speaker’s cognitive processes. Arnold et al. (2007, 2004) claim that the biases towards discourse-new and unfamiliar objects following filled pauses reflect listeners’ inferences that disfluency is caused by an increased

cognitive load due to naming of difficult-to-name objects. If speakers' production of gestures varies systematically with the conceptual demand required to formulate verbal descriptions of objects, it stands to reason that listeners may draw similar inferences about the causes of variations in the visual channel. That is to say, listeners may associate the occurrence of gesturing with less familiar/easily-named objects in a similar way to the occurrence of speech disfluency.

Relative to manner of spoken delivery, however, research into how listeners' real-time comprehension is influenced by information in the visual channel is limited. Studies of multi-modal language comprehension have shown that gesturing facilitates listeners' understanding (for an overview see Hostetter, 2011), but these have tended to use after-the-fact measures (e.g., message recall, or effective use of information from the message in a subsequent task)

A small number of studies, however, have investigated how information in the two modalities is integrated during comprehension. For instance, by measuring Event Related Potentials (ERP), Kelly et al. (2004) found an N400 effect when speech was semantically incongruent with a preceding gesture (e.g., "short" following a gesture representing tallness). Similarly, Özyürek et al. (2007) found that when information which is incongruent with sentential context is presented to a listener, the corresponding brain responses are similar when the mismatching information is presented in gesture as to when it is presented in speech.

These studies point towards an account of the content of gesture (i.e., what a gesture represents) being integrated with the content of speech at the early stages of language processing. However, no studies (to our knowledge) have investigated whether listeners exploit the mere occurrence of gesturing to inform comprehension, for instance as signals similar to filled pauses about the difficulty of upcoming referents.

This chapter presents a preliminary exploration of this possibility, asking whether listeners explicitly associate the occurrence of different types of gesturing with less easily named referents. We initially focus on *iconic* gestures in Experiment 4.1, subsequently extending the investigation to *adaptor* gestures (self- or object-directed movements) in Experiment 4.2.

A growing body of work has investigated speakers' use of iconic gesturing under conditions which vary with respect to the effort required to produce verbal descriptions. Iconic gestures are those which represent the spatial and kinetic properties of physical, concrete items (McNeill, 1992). Speakers tend to produce more iconic gestures when describing from memory (as opposed to describing a referent which is visually present, see De Ruiter, 1998b; Wesp et al., 2001); when describing objects that are especially taxing on spatial working memory (Morsella & Krauss, 2004); and when describing objects which require generating geometric conceptualisations (in comparison to when those conceptualisations are given to them, see Hostetter et al., 2007b). Experiment 3.1, presented in the previous chapter, patterns with these studies: Participants' descriptions of less nameable shapes were found to include more occurrences, greater durations, and earlier onsets (see also Morrel-Samuels & Krauss, 1992) of iconic gesturing relative to speech, supporting an account which associates this type of gesturing with conceptual demand.

Providing a parallel to studies which suggest that speech disfluency biases listeners to predict the less familiar of two shapes (gating task, Experiment 1, Arnold et al., 2007), we manipulated the presence of gesturing in a visual-world paradigm. Participants viewed a display comprising two objects and a video of a speaker. In critical trials, the two objects displayed consisted of an easy-to-name shape (a letter, number, punctuation symbol, or simple geometric shape) and a difficult-to-name variant. Fragments of speech and gesture were presented to participants, which in critical trials contained no information which disambiguated between the

two objects. Analogously to the gating task in Arnold et al. (2007), participants were tasked with guessing which object the instruction was about to refer to. Crucially, we manipulated whether the videos showed a fragment of gesturing, or showed the speaker sitting motionless. If listeners associate the occurrence of gesturing with speech planning difficulty, then we would expect a higher proportion of clicks to the more difficult-to-name shape following trials in which the speaker was seen gesturing. Because fragments of gestures varied in length, we were also able to investigate whether listeners were sensitive to the relative durations of gesture and speech.

Additionally to recording which shapes participants clicked in each trial, we also recorded participants' eye and mouse movements. Using these measures, we also explored the potential of adapting the visual world paradigm to include a video display of the speaker to capture the time course of multi-modal comprehension. This was predicated on evidence which suggests that most gestures are perceived through peripheral vision (see e.g., Gullberg & Kita, 2009). The influence of gesturing on participants' eye and mouse movements towards either object in the display will indicate when (if at all) any association between gesturing and less nameable objects emerges. As participants are tasked with making explicit predictions about upcoming referents, the time course of their eye and mouse movements is used largely to establish the potential viability of developing further experiments in which participants are presented with full instructions (as in Arnold et al., 2007, 's eye-tracking task) to investigate whether listeners associate gestures with less easily named objects in the moment-to-moment processing of speech.

4.2 Experiment 4.1

In Experiment 4.1, participants viewed a visual world comprising an easily named object (letter, number, geometric shape), a more difficult-to-name object (squiggle) and a video of a speaker. Fragments of instructions to click on an object were presented in audio and video, which in critical trials was truncated immediately prior to the point at which speech and gesture (when present) would disambiguate between the two objects. Participants were tasked with clicking on the object which they guessed the speaker was referring to. We manipulated whether the videos showed the speaker producing iconic gesturing (the shape of which was ambiguous as to which object it represented, explained in Section 4.2.1 below), or sitting motionless. The same audio recording of a fragment of speech was used across all critical trials, and the durations of gesture fragments varied. This made it possible to also investigate whether the relative durations of speech and gesture influenced participants' predictions of which shape the speaker was about to refer to.

Twenty self-reported native English speaking participants took part in the experiment, recruited from the University of Edinburgh community, in return for a payment of £4. Care was taken to ensure that participants had not taken part in any of the other experiments presented in this thesis. All participants were right-handed mouse users with normal or corrected-to-normal vision. Consent was obtained in accordance with the University of Edinburgh's Psychology Research Ethics Committee guidelines (reference number: 228-1617/1). The experiment was pre-registered at <https://osf.io/t68be/>

4.2.1 Method

Images

A set of 120 symbols and shapes were seen in pairs in each trial. This set comprised 60 easily-named shapes (letters, numbers, keyboard symbols and geometrical shapes), and 60 difficult-to-name variants. Difficult-to-name variants were created by drawing part of an easily named shape and extending it in a novel direction/manner (see Figure 4.1) The aim of this was to make it possible that a pair of shapes (easy-to-name and its difficult-to-name variant) could share the same initial shape or trajectory when gestured in space. 40 of these shapes (20 easy shapes and their difficult counterparts) were used in 20 critical trials, in which easy-to-name shapes were displayed alongside their corresponding difficult-to-name variant. The remaining 80 shapes were presented in 40 filler trials, which displayed either an easy-to-name shape and an unrelated difficult-to-name shape (20 trials); two easy-to-name shapes (10 trials); or two difficult-to-name shapes (10 trials). Pairings of the shapes in filler trials were randomly selected.

Audio

For each of the critical shapes, recordings were constructed of a speaker producing utterances of “The one you should click on is the <referent>”. To control for manner of spoken delivery, the same recording (duration = 1534 ms) of the initial part of the sentence (truncated at onset of the referent) was used in all critical trials. Ten of the filler trials used different recordings of this same initial sentence fragment, and the remaining 30 contained various disfluencies and discourse manipulations (see Table 4.1) to make it more believable that the stimuli were not scripted.

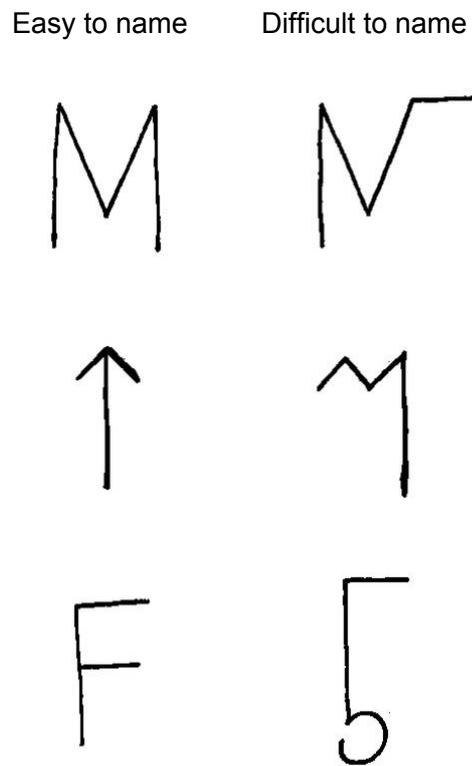


Figure 4.1: Examples of easy-to-name shapes and their difficult-to-name variants in Experiment 4.1

Video

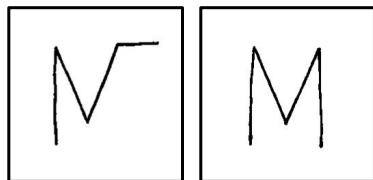
In each trial, the two shapes were displayed alongside a video which purported to show the speaker in the audio recording. Videos presented the speaker either producing a fragment of an iconic gesture or sitting motionless. 120 video clips were recorded of a volunteer drawing the symbols in space. For videos referring to the 40 critical shapes, the volunteer was instructed to make the initial part of the gesture as similar as possible for both shapes in each easy- and difficult-to-name pair. Gestures of these shapes were either tracing motions with an index finger of one hand (e.g., tracing a circle); static holds in which the holds in either hand were staggered (e.g., index finger and thumb of one hand held in a half triangle shape, followed by the other hand raising to do the same and make a full triangle),

Table 4.1: Disfluencies and discourse manipulations in filler items in Experiment 4.1

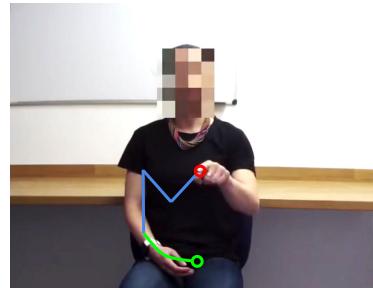
No. of Utterances	Example
10	The one you should click on is the <referent>.
3	Okay, click on the one which is like a <referent>.
3	So you should click on the <referent>.
2	Err, the you one you should click on is the <referent>.
2	Okay, the one you should click on is <i>thee - um</i> , <referent>.
2	All right, so you should click on the <referent>.
2	Okay, so click on the <referent>.
2	Okay, so the one you should click on is the <referent>.
2	Okay, the one you should click on is the <referent>.
2	You should click on the <referent>.
2	You should click on the one that's like a <referent>.
2	You should click on the one which is a <referent>.
1	The one you should click on is the <i>er</i> , <referent>.
1	You should click on <i>thee- er</i> , the <referent>.
1	Click on the one which is like a <referent>.
1	Okay, you should click on the <referent>.
1	So this one is the <referent>.
1	The one you should click ... is the <referent>.

or a combination of the two (e.g., one hand holds the vertical line of a letter K, the other hand traces the diagonal lines). Videos were flipped horizontally so that the gesture fragment mapped directly to the shapes seen by the participants. For each video, the point-of-disambiguation was identified as the frame of the video in which the gesture disambiguated between the easy shape and its difficult-to-name variant (see Figure 4.2).

Possible referents



Ambiguous gesture fragment



- Onset of movement
- Preparation phase
- Ambiguous gesturing
- Point of disambiguation
- Unambiguous gesturing
- Retraction phase
- End of movement

Full gestures

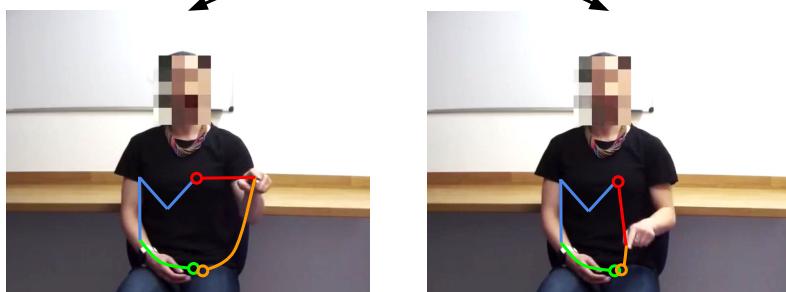


Figure 4.2: Example gestures to each of an easy-to-name shape and its difficult-to-name variant in Experiment 4.1

For the remaining 80 videos in which the gesture represented a filler shape, the volunteer was asked simply to draw each shape in space. In all videos, the volunteer's face was pixelated, so that when presented simultaneously with the audio recordings it appeared as if the spoken utterances were produced by the person in the video. Participants were informed that this pixelating was to maintain anonymity. In all recordings of gestures the volunteer was asked to

simultaneously say “The one you should click on is the thing”, ensuring that, when edited, the pixelated region displayed some movement across time.

A further 20 video clips of 10 seconds were recorded of the speaker sitting motionless in a neutral posture, which were subsequently sampled randomly (with replacement) in any trial (critical or filler) presenting no gesturing.

Timing and truncation of stimuli

In critical trials, both audio and video were truncated so that no disambiguating information was presented in either modality, and were timed such that both streams stopped simultaneously. In other words, had they continued, disambiguation in speech would have occurred simultaneously with disambiguation in gesture. One result of this was that the durations of gesture fragments varied across shapes. For critical trials, the duration of gesturing (including preparation phase) ranged from 500 ms to 1750 ms (Mean = 1280 ms, SD = 345 ms). In trials in which the video displayed the speaker sitting motionless (i.e., in the No Gesture condition), the same number of frames of video were presented as in the equivalent trial showing that shape in the Gesture condition. Duration of gesturing was included as a predictor in the analysis to establish whether listeners are sensitive to the relative durations of speech and gesture in making their predictions (i.e., perceive longer gestures to more strongly signal difficulty in speech, as suggested in Chapter 3).

Filler trials were truncated at a later point, presenting participants with up to 200 ms (randomly determined on each trial) of speech and gesture which referred to one of the two shapes. This was to encourage participants to pay attention to the stimuli throughout the experiment. Because filler trials displayed shapes which did not share the same trajectory when gestured in space, the point of disambiguation in speech and gesture varied in these trials (i.e., for filler trials in

which the video showed a gesture, it was possible to disambiguate which shape the gesture referred to prior to point of disambiguation in speech). In filler trials, the exact timings of gesture relative to speech were individually determined for each video recording, based on what appeared most believable to the researchers. As in critical trials, filler trials in which the video showed no gesturing were presented for the same number of frames as in the video of the speaker producing a gesture for that shape.

Lists

The 20 critical pairs of shapes were counterbalanced across four lists, each containing 10 gesture videos (five showing the initial ambiguous fragment of a gesture of an easy-to-name shape, five showing a the ambiguous fragment of a gesture of a difficult-to-name shape), and 10 showing videos without gesturing. Because videos in critical trials were never presented beyond the point at which gestures disambiguated between shapes, two of these lists present stimuli which are theoretically indiscernible from the other two lists. However, despite the initial fragments of gestures of a nameable shape and its difficult-to-name counterpart being, to all intents and purposes, ambiguous between the two shapes, they may differ in, for example, velocity, size, or hesitancy during motion. We therefore controlled for any sensitivity listeners may have to such differences in gesturing style by following this two (Gesture vs. No gesture) by two (Easy-to-name vs. Difficult-to-name shape) design—the latter manipulation being simply a precautionary stimulus check which could be included in the analysis.

In the 40 filler trials, 20 included a video showing the speaker gesturing, and the remaining 20 showed the speaker sitting motionless. In each set of 20, 10 presented the speaker referring to (in gesture when present, and in up to 200 ms of speech post-truncation) an easily named shape and 10 to a difficult-to-name shape. In

each set of 10, five displayed two shapes of the same nameability, and five displayed shapes of differing nameabilities. Shapes in filler trials were randomly assigned to these conditions on each run of the experiment.

Across the experiment, each participant saw a total of 30 trials in which the speaker gestured and 30 in which they did not. Alongside the critical trials (in which audio and video were ambiguous between shapes), filler trials ensured that participants saw an equal number of references to easy shapes as difficult shapes, with an equal number of gestures to easy and difficult shapes.

Cover story

A key aspect of the study was that participants believed that speech and gesture had been produced naturally and concurrently. Participants were told that the recordings were the result of a previous experiment, in which speakers were presented with the same pairs of shapes as in the present study. To ensure that we excluded from the analysis any participants who did not believe this deception, a post-test questionnaire assessed whether participants noticed anything strange about the audio and video. After participants were debriefed about the true nature of the experiment, it was explained to them that the speech and gesture had been artificially constructed (and not even produced by the same people) and they were asked again verbally whether it had occurred to them during the experiment that the recordings might not be real. Participants who indicated in either the post-test questionnaire or during verbal questioning that they did not believe the supposed origins of the stimuli were subsequently removed from the analysis.

Procedure

The experiment was presented using OpenSesame version 3.1 (Mathôt, Schreij, & Theeuwes, 2012). Stimuli were displayed on a 21 in. CRT monitor with a resolution of 1024×768 , placed 850 mm from an Eyelink 1000 Tower-mounted eye-tracker which tracked eye movements at 500 Hz (right eye only). Audio was sampled at 44100 Hz and presented in stereo from speakers on either side of the monitor. Videos were played at 20 frames per second, and mouse coordinates were sampled at every frame (every 50 ms).

Participants were told they would see a series of pairs of shapes and symbols, some familiar, and some made-up. They were told that they would also see fragments of videos of a speaker who had seen the same shapes with one of them highlighted, and had been tasked with “providing an instruction along the lines of ‘the one you should click on is the’ and then describing the highlighted shape”. They were told that the lengths of fragments varied, and instructed to click on the shape that they thought the speaker was describing or was about to describe. Once participants had read the instructions, the eye-tracker was calibrated. Recalibration occurred between trials where necessary.

Figure 4.3 shows the procedure of a given trial. Each trial began with a manual drift correction using a central fixation point, that changed from grey to red upon successful fixation. Following the red fixation point (500 ms), two shapes measuring 250×250 pixels were presented horizontally to the left and right of the midpoint of the screen, such that the center of each shape was located 15% of the screen-width inside from either edge of the display. Easy and difficult shapes were presented equally often on each side. Filler trials, in which the speaker’s description disambiguated between shapes both in their gesture and in up to 200 ms of speech after point-of-truncation, displayed the referred to shape equally often on each side.

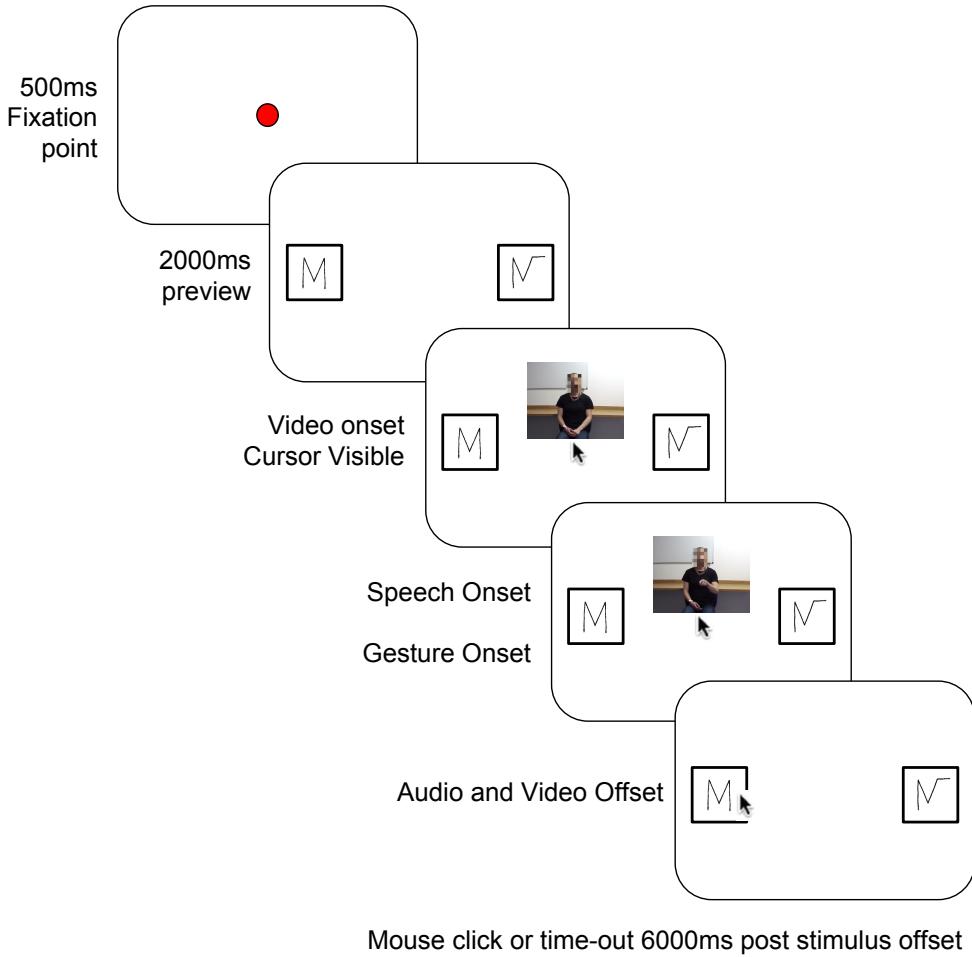


Figure 4.3: Procedure for a trial in Experiment 4.1

After 2000 ms, the video appeared, measuring 373×293 pixels, and centred horizontally, and with the bottom edge at the vertical midpoint of the display. Simultaneously, a mouse pointer was made visible and centred. Playback of the utterance began at the assigned frame of the video. Video and audio were presented for the length of fragment (up to the point of disambiguation for critical trials, and from 0 to 200 ms post-speech disambiguation in filler trials). After this, both audio and video were truncated and the video disappeared. Participants used the mouse to click on one of the two shapes. Trials timed out 6000 ms after truncation. Once either participants had clicked the mouse or the trial had timed

out, the stimuli disappeared and were replaced by a grey fixation dot, signifying the beginning of the next trial.

Participants completed five practice trials prior to the main experiment, including two trials showing a video of the speaker gesturing. One of these trials showed a gesture which represented one of two difficult shapes, and one showed a gesture which represented one of two easy shapes. The remaining three practice trials presented the speaker sitting motionless and with partial spoken descriptions of 1) one of two easy shapes 2) one of two difficult shapes and 3) an easy shape in an easy/difficult pair.

Along with mouse-clicks to objects in the display, participants' eye movements and mouse coordinates were recorded throughout each trial.

4.2.2 Results

No participants indicated either in the post-test questionnaire or in response to verbal questioning that they did not believe the proposed origins of the audiovisual stimuli.

Analysis

Analysis was carried out in R version 3.5.1 (R Core Team, 2018), using the lme4 package version 1.1-17 (Bates et al., 2015). Out of 400 critical trials, two trials in which participants did not click on either shape were excluded from all analyses. Mouse-movements beyond the outer edge of either shape were considered to be 'overshooting' and were not included in calculations (0.5% of samples).

Mouse-clicks (on either the easy-to-name or difficult-to-name shape) were modelled

using mixed effects logistic regression, with fixed effects of presence of gesture (No Gesture vs. Gesture, deviation coded). A main effect of the gesture duration (Z-scored) was included, along with its interaction with presence of gesture. This allowed us to test whether any effect of gesture on participants' decisions to click on either object was moderated by the duration and onset of gesturing relative to speech. A main effect of the source of the gesture fragment (whether the fragment came from a recording of a gesture of the easy- or the difficult-to-name shape, deviation coded) and its interaction with presence of gesture were also included. This controlled for any sensitivity that participants may have to subtle differences in the way our volunteer produced initial parts of gestures in the videos depending on whether they were gesturing an easy or difficult-to-name shape. These measures (gesture duration and source of fragment) only apply to videos of gestures, but values are matched by videos of no gestures for each item. Random intercepts and slopes for presence of gesture were included both by-participant and by-item (pair of shapes).

Reaction times to click on a shape (measured from the point at which audiovisual stimuli stopped) were log transformed and analysed using mixed effects linear regression with the fixed effects of presence of gesture (No Gesture vs. Gesture, deviation coded), gesture duration (Z scored), gesture source (Easy-to-name vs. Difficult-to-name, deviation coded) and Shape clicked (Easy-to-name vs. Difficult-to-name, deviation coded). Two-way interactions between presence of gesture and all other fixed effects were also included, and random intercepts and slopes of presence of gesture were included by-participant and by-item.

Analyses for both eye and mouse movements were conducted over the time window beginning at 600 ms prior to the truncation of speech and gesture and extending for 1200 ms, by which point the display had consisted of only the two shapes for 600 ms. If participants formulated hypotheses about upcoming referents during the presentation of speech and gesture stimuli, we would expect a bias in fixations on

(and mouse movements to) one object over the other to have emerged within the first half of this time window (i.e., prior to the point at which speech and gesture would have disambiguated had they continued). To account for the possibility that the video component of the display delayed any fixations and movements to objects (and so prevented any biases from being detected), we included the 600 ms after audio and video were truncated. Due to the scarcity of previous eye-tracking studies investigating gestures, the entire time-course of eye and mouse movements from onset of speech to the mean click time will also be discussed.

Eye fixation data was averaged into 20 ms bins (of 10 samples) prior to analysis. For each bin, we calculated the proportion of time spent fixating the easy shape or the difficult shape, resulting in a measure of the proportions of fixations on either shape over time.

The position of the mouse was sampled every 50 ms, corresponding to 2.5 bins of eye-tracking data. Using the X coordinates only, we calculated the number of screen pixels moved and the direction of movement (towards the easy or the difficult shape) from the onset of speech. The cumulative distance travelled towards each shape was calculated for each bin, and divided by the cumulative distance moved in any direction. The resulting measure was the proportion of cumulative distance travelled towards either shape from speech onset.

The proportions of fixations and mouse movements to either object (easy-to-name shape and difficult-to-name shape) were empirical logit transformed (Barr, 2008). The resulting difference between difficult-to-name shape and easy-to-name shape yielded measures for which a value of zero in either measure indicates no bias towards either shape, and positive and negative values indicate a bias towards the difficult shape and easy shape respectively.

Empirical logit transformed fixation bias was modelled using linear mixed effects

models, including fixed effects of gesture (gesture vs. no gesture, deviation coded) as well as orthogonal linear and quadratic effects of time and their interaction with gesture (see Mirman, Dixon, & Magnuson, 2008). Higher-order time terms were not included as their inclusion was not found to improve model fit as indicated by likelihood ratio test and Bayesian information criterion (with a decrease of ≥ 10 considered improvement, following Raftery 1995).

Random intercepts and slopes for gesture and both degrees of time were included by-participant and by-item. The mouse-movement bias was modelled analogously with only linear effects of time (no higher order polynomials resulted in improved model fit). Following Baayen (2008), we considered effects in these models to be significant where $|t| > 2$.

Object clicks and response times

Across the critical items in the experiment, participants clicked on the difficult-to-name shape in 58% of critical trials and the easy-to-name shape in 42%. Table 4.2 shows the numbers of clicks on each type of shape by presence of gesture. Model results (see Table 4.3) revealed an overall tendency to predict that the speaker was about to mention the more difficult-to-name shape ($\beta = 0.43$, $SE = 0.19$, $p = .026$). The presence of gesturing was found to influence listeners' guesses about which shape the speaker was about to refer to: Participants were more likely to click on the more difficult-to-name shape following videos showing gesturing than following videos of a speaker sitting motionless ($\beta = 1.07$, $SE = 0.42$, $p = .011$). This likelihood to click on the difficult-to-name shape was greater following longer gesture fragments, as indicated by both the main effect of gesture duration ($\beta = 0.29$, $SE = 0.14$, $p = .034$) and the significant interaction between presence of gesture and gesture duration ($\beta = 0.63$, $SE = 0.26$, $p = .016$). The source of the gesture fragment (whether it was the initial part of a gesture or an

easy-to-name shape or the initial part of a gesture of a difficult-to-name shape) was not found to influence which object participants clicked on. Analysis of time to click (measured from the offset of stimuli) revealed that participants were quicker to click on a shape in trials showing videos of gesturing than in those showing videos of no gesturing ($\beta = -0.15$, SE = 0.07, $t = -2.23$). Full results of the model are shown in Table 4.4.

Table 4.2: Breakdown of mouse clicks recorded on each shape (easy or difficult) by condition in critical trials in Experiment 4.1

	No Gesture	Ambiguous Iconic Gesture
Clicks to Easy-to-name	104 (52.5%)	62 (31.0%)
Shape		
Clicks to Difficult-to-name	94 (47.5%)	138 (69.0%)
Shape		

Table 4.3: Model results for mouse clicks to difficult-to-name shapes over easy-to-name ones in Experiment 4.1

	β	SE	p
(Intercept)	0.43	(0.19)	.026
Gesture	1.07	(0.42)	.011
Gesture duration (Z-scored)	0.29	(0.14)	.034
Difficulty-to-name of gesture source	0.20	(0.24)	.42
Gesture \times Gesture duration (Z-scored)	0.63	(0.26)	.016
Gesture \times Difficulty-to-name of gesture source	-0.57	(0.48)	.24
Var(1—Participant)	0.42		
Var(Gesture—Participant)	2.35		
Var(1—Item)	0.04		
Var(Gesture—Item)	0.00		
Total	398		
Participant	20		
Item	20		

Table 4.4: Model results for times taken to click the mouse in Experiment 4.1

	β	SE	t
(Intercept)	7.59	(0.06)	116.90
Gesture	-0.15	(0.07)	-2.23
Gesture duration (Z-scored)	0.01	(0.02)	0.37
Difficulty-to-name of gesture source	-0.02	(0.03)	-0.58
Difficulty-to-name of clicked shape	0.04	(0.04)	1.11
Gesture \times Gesture duration (Z-scored)	-0.03	(0.05)	-0.56
Gesture \times Difficulty-to-name of gesture source	0.02	(0.07)	0.35
Gesture \times Difficulty-to-name of clicked shape	-0.08	(0.08)	-1.04
Var(1—Participant)	0.07		
Var(Gesture—Participant)	0.05		
Var(1—Item)	0.00		
Var(Gesture—Item)	0.02		
Total	398		
Participant	20		
Item	20		

Eye movements

Figure 4.4 shows the time course of fixations to all items in the display (easy and difficult shapes and video of the speaker) for the 1200 ms centered on stimulus offset, split by the presence of iconic gesturing.

Analysis of the time window beginning at 600 ms preceding stimulus offset and extending for 1200 ms revealed a significant intercept term, indicating an overall bias to fixate the difficult-to-name shape across this window ($\beta = 0.69$, $SE = 0.14$, $t = 5.01$). This bias to more difficult-to-name shapes increased over the course of the window (as indicated by a linear effect of time $\beta = 3.62$, $SE = 0.91$, $t = 3.99$), with a significant curvature (the increase in the difficult-to-name bias becoming steeper as the window progressed $\beta = 1.59$, $SE = 0.69$, $t = 2.31$). While the linear increase in fixations to difficult-to-name shapes over the easy-to-name ones was not influenced by the presence of gesturing in the video, the quadratic term was—indicating a steadier increase in the difficult-shape bias—found in

trials showing the video gesturing as opposed to sitting motionless ($\beta = -2.51$, SE = 0.33, $t = -7.72$). Figure 4.5 shows the empirical logit transformed bias towards difficult-to-name shapes over easy-to-name ones in the relevant window of analysis, along with the fitted values from the model, and Table 4.5 shows the full model results.

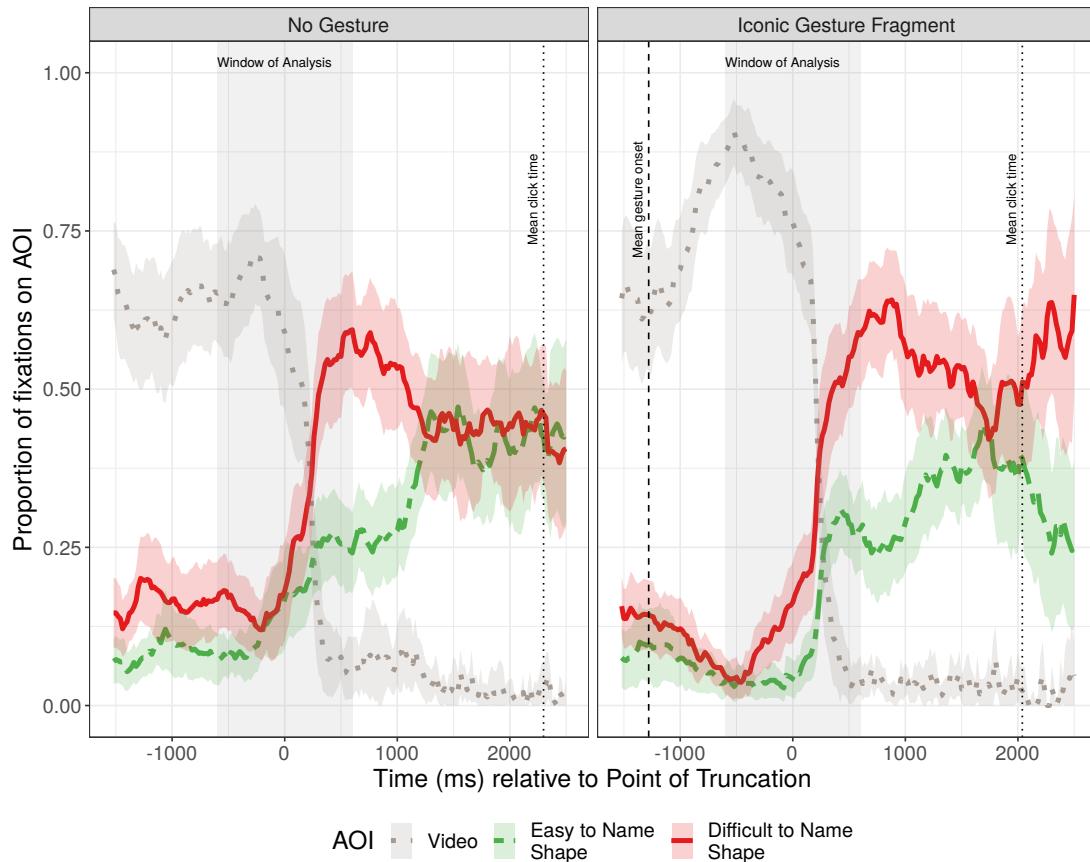


Figure 4.4: Eye-tracking results for critical trials in Experiment 4.1: Proportion of fixations to each object (easy or difficult-to-name shape) and the video, from speech onset to 2000 ms post stimulus offset, calculated out of the total sum of fixations for each 20 ms time bin. Shaded areas represent 95% confidence intervals derived via bootstrapping subject data (R=1000).

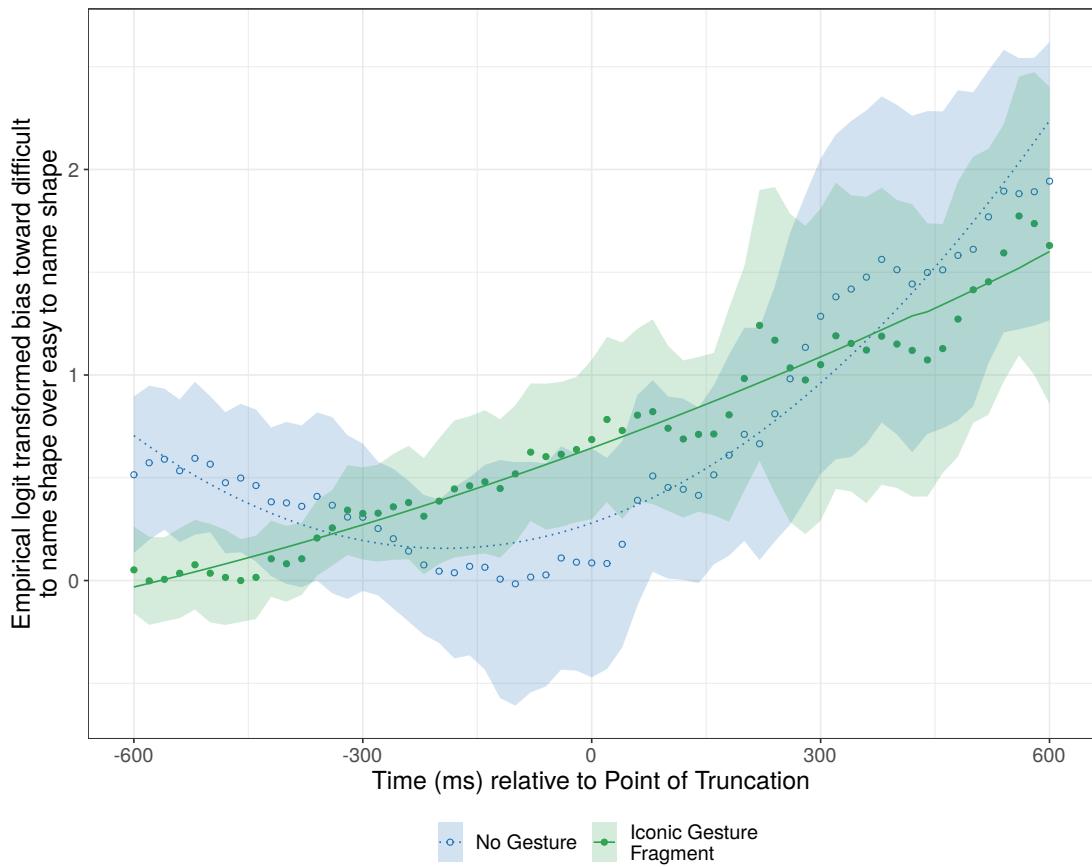


Figure 4.5: Empirical logit transformed fixation bias towards difficult-to-name shapes over easy-to-name shapes in critical trials in Experiment 4.1 for the 1200 ms window centered on stimulus offset. Lines represent fitted values of the model.

Mouse movements

Figure 4.6 shows the time course of proportions of cumulative mouse movements towards the easy-to-name and difficult-to-name shapes over the 1200 ms centered on stimulus offset, split by the presence of iconic gesturing. Analysis of the 600 ms either side of the offset of stimuli indicated that participants were no more likely to move towards one shape over the other throughout this period, and this was not influenced by whether or not the video showed the speaker gesturing (full model results are shown in Table 4.5).

Table 4.5: Model results for eye- and mouse-tracking analysis over the 1200 ms window centered on point-of-truncation in Experiment 4.1

	Fixations			Mouse Movements		
	β	SE	t	β	SE	t
(Intercept)	0.69	(0.14)	5.01	-0.06	(0.15)	-0.41
Gesture	-0.00	(0.25)	-0.01	-0.17	(0.34)	-0.50
Time	3.62	(0.91)	4.00	0.03	(0.15)	0.22
Time ²	1.59	(0.69)	2.31			
Gesture × Time	0.25	(0.33)	0.77	0.02	(0.21)	0.09
Gesture × Time ²	-2.51	(0.33)	-7.72			
Var(residual)	10.49			4.40		
Var(1—Participant)	0.24			0.28		
Var(Gesture—Participant)	0.81			1.25		
Var(Time—Participant)	9.52			0.18		
Var(Time ² —Participant)	4.14					
Var(1—Item)	0.14			0.16		
Var(Gesture—Item)	0.44			0.96		
Var(Time—Item)	6.37			0.07		
Var(Time ² —Item)	4.89					
Total	24249			9188		
Participant	20			20		
Item	20			20		

4.2.3 Discussion

Experiment 4.1 aimed to establish whether listeners' predictions about which of two shapes a speaker is about to refer to are modulated by the presence of iconic gesturing. In a visual world paradigm in which participants were presented with pairs of shapes (one easy-to-name, one difficult-to-name) and initial, ambiguous fragments of multi-modal instructions to click on one of the shapes, we manipulated whether the instructions showed the speaker producing an iconic gesture or sitting motionless. Crucially, in critical trials, participants were presented with no disambiguating information in either modality: Speech was truncated immediately prior to the referring expression, and gestures were truncated at the point at which the trajectory of gestures disambiguated between either shape, with the

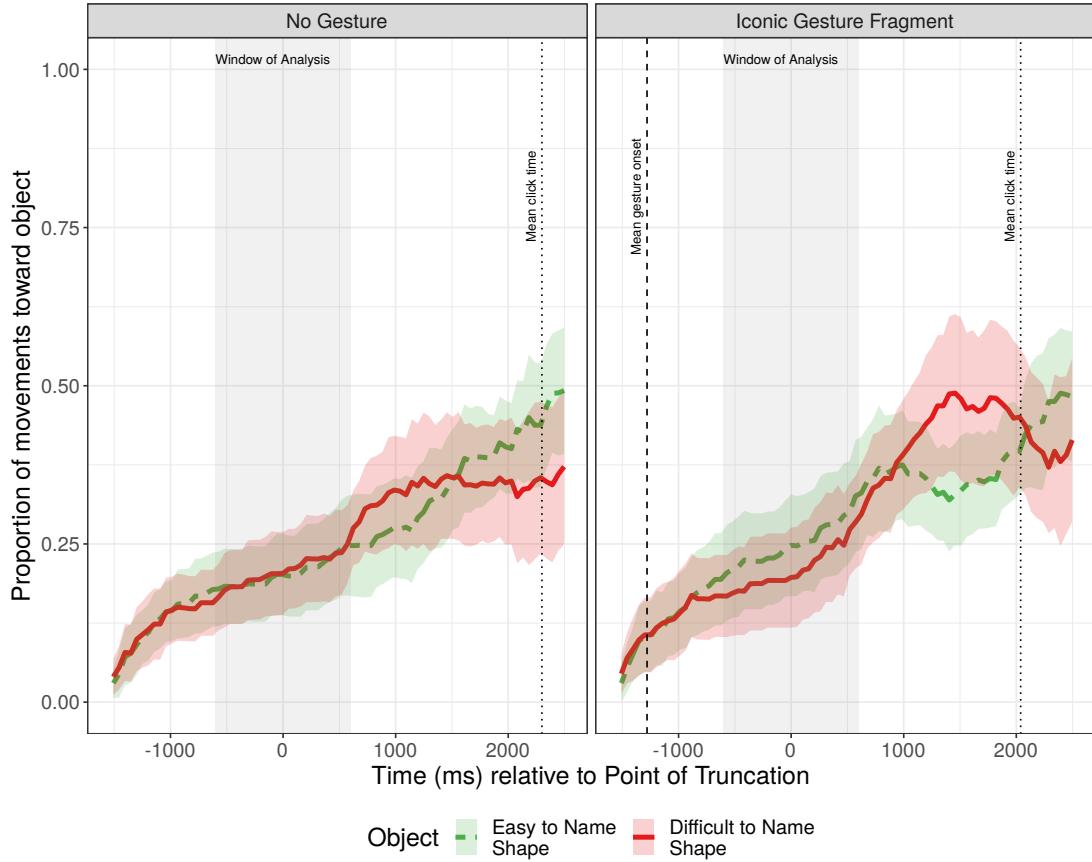


Figure 4.6: Mouse-tracking results for critical trials in Experiment 4.1: Proportion of cumulative distance travelled toward each object from speech onset to 2000 ms post stimulus offset. Proportions were calculated from the total cumulative distance participants moved the mouse until that time bin. Shaded areas represent 95% confidence intervals derived via bootstrapping subject data ($R=1000$).

presentation of audio and video timed such that they both ended at the same point. Results suggest that listeners' explicit predictions about the conceptual difficulty of upcoming referents is influenced by whether or not the speaker produces iconic gestures alongside speech: Participants were more likely to click on the difficult-to-name shape following speech with iconic gesturing than speech without. Listeners associated iconic gesturing with difficult-to-name shapes despite the gesture fragments remaining ambiguous between the two options (easy-to-name and difficult-to-name shapes) in all critical trials, suggesting that the effect is

driven by the act of gesturing as a signal. Additionally, participants were quicker to click on an object following the speech which was accompanied by gesture than speech which was not, suggesting that the presence of iconic gesturing, although ambiguous, aided the formulation of participants' guesses as to which object the speaker was about to refer to.

Due to the varying durations of gesture fragments used, we were also able to investigate whether the effects of gesture on participants' predictions of upcoming referents was influenced by the relative durations of gesture and speech (which was the same across all critical trials). Greater durations of gesturing relative to speech resulted in an increased tendency to click on the difficult-to-name shape over the easy-to-name one, possibly contributing to the impression that the speaker is having difficulty in forming a verbal description. This may be due to these fragments being more visually salient: Participants might simply have missed more of the shorter gestures, although this is unlikely given that the shortest gesture fragments lasted 500 ms. Alternatively, it may reflect that participants are sensitive to the durations of speech and gesture, and adjust predictions of upcoming content accordingly to the perceived effort which the speaker puts into either modality. This account has an attractive symmetry with the findings from Experiment 3.1, in which the production of referring expressions to more difficult-to-name shapes (relative to those referring to familiar shapes) were associated with greater durations of iconic gesturing relative to speech.

Lastly, it may be that listeners are sensitive to the temporal asynchrony of the relative onsets of speech and gesture. Because presentation of both speech and gesture were timed to end simultaneously, variation in the duration of gesture fragments necessarily entailed variation in the relative timing of gesture onset with speech onset. Previous research suggests that there is an optimal synchrony for the integration of speech and gesture: Habets, Kita, Shao, Özyürek, and Hagoort (2011) found that N400 effects elicited by mismatched speech and gesture

disappeared after a certain amount of delay between modalities. In relation to how speech-gesture asynchrony varies in language production, Experiment 3.1 found that descriptions of more difficult-to-name shapes resulting in greater temporal asynchrony between onset of gestures and onset of the noun phrase in speech (see also Morrel-Samuels & Krauss, 1992, for further support that gesture onset precedes speech onset by a magnitude inversely proportional to a referent's familiarity). Any sensitivity listeners may have to asynchrony in the relative onsets of speech and gesture therefore would appear to be warranted by their potential validity as a signal of the language production process. Due to the nature of the experimental stimuli (speech and gesture ending simultaneously, with gesture duration varying backwards from this point), it is impossible to discern whether participants were responding to the relative durations of speech and gesture, or the relative onset timings.

Eye-tracking analyses of the 1200 ms centred on the offset of audiovisual stimuli revealed that, following both speech with gesturing and speech without, participants tended to fixate more on the difficult-to-name shape than the easy-to-name shape over time. A significant quadratic effect of time indicated that the increase in this bias towards the difficult-to-name shape became steeper over the course of the window. This curvature is perhaps to be expected here: This window is centered at the point at which audio and video stop, meaning that fixations to other objects in the display are likely to increase at a greater rate in the second half of the window (i.e., after the video has disappeared from the display). Although this curvature was influenced by the presence of gesturing in the video, we suggest that this is likely because videos of gestures are more visually salient than those of a static speaker, resulting in delayed fixations to other objects in the display. This explanation is supported by visual inspection of the time-course of fixations prior to stimulus offset which shows a greater proportion of fixations to the video when it includes gesturing (see Figure 4.4).

Interestingly, mouse-tracking analysis over this time window did not pattern with the eye-tracking results: Participants' tendencies to fixate the difficult-to-name shape over the easy-to-name shape were not borne out in their mouse movements, in which they showed no preference for either shape. Visual inspection of the time-course of mouse movements over the longer window (up until mean time to click, see Figure 4.6) offers some insight. The tendency to eventually click on the difficult-to-name shape following speech accompanied by iconic gesturing appears to pattern with more movements of the mouse towards these shapes, but this only emerges comparatively late on (about 600 ms post stimulus offset). A similar pattern in this later window is perhaps interpretable in the time course of fixations (Figure 4.4), with a greater attenuation in the fixation bias to the difficult-to-name shape at the mean time of mouse click following videos showing no gesturing.

It is likely that these differences are due to the different sensitivities of eye-tracking and mouse-tracking respectively, with the visual salience of difficult-to-name shapes relative to familiar shapes influencing participants' fixations more than their mouse movements (see also Tavakoli, Ahmed, Borji, & Laaksonen, 2017, for a discussion of mouse data for models of visual saliency). This can also be seen earlier on in the time course, with a greater proportion of fixations towards the difficult shape than the easy shape during the presentation of speech and gesture (see Figure 4.4). The difference between measures may be exaggerated in a paradigm such as the one presented here, in which mouse-movements reflect participants' commitment to complete a task within a limited time (to click on an object), as opposed to studies in which participants freely view (and move the cursor around) visual scenes.

The results of Experiment 4.1 indicate that, when tasked with making explicit predictions about upcoming referents, the presence of iconic gesturing biases listeners towards expecting more difficult-to-name shapes. Furthermore, this bias is greater following longer gesture fragments, reflecting a possible sensitivity either

to the relative durations or the synchrony of onsets of speech and gesture. Eye- and mouse-tracking measures suggest that these predictions may have emerged relatively late on, appearing only 600 ms after the speech and gesture stimuli had stopped, and likely only detectable in mouse movements. Patterns of fixations appeared to be confounded by the greater visual salience of difficult-to-name shapes relative to easy-to-name ones, as well as participants attending more to videos of gesturing relative to videos of a motionless speaker.

Listeners may similarly associate different types of gesturing with speakers experiencing difficulty in naming objects, and for some of these types of gestures it may be less important for listeners to attend to the exact trajectory or shape of the gestures. Adaptor gestures are movements or touching behaviours directed towards the self, objects, or others. These are often considered to “indicate internal states typically related to arousal or anxiety” (Hans & Hans, 2015, p.47), and are comparable to speech disfluencies in that they contain no representational or semantic content, serving only as potentially informative signals about a speaker’s meta-cognitive states (in contrast to iconic gestures which could be perceived as intentionally communicative efforts). Unlike iconic gestures, the production of adaptor gestures has not been so widely studied in relation to situations with increased cognitive demand or speech planning difficulty (and in Experiment 3.1 was not found to differ for easy- or difficult-to-name shapes), although one study suggests that finger-tapping may facilitate the retrieval of rare words based on their definitions (see Ravizza, 2003). It is possible, however, that listeners may associate adaptor gesturing with difficulty in speech planning regardless of the validity of such an association. Experiment 4.2 explores the possibility that adaptor gestures (specifically finger-tapping, fidgeting, and adjustments to clothing) might lead listeners to expect more difficult-to-name objects.

4.3 Experiment 4.2

4.3.1 Method

Figure 4.7 shows the procedure of a given trial in Experiment 4.2. The procedure was the same as that of Experiment 4.1 but instead of displaying iconic gesture fragments, videos showed a speaker making an adaptor gesture (fidgeting, scratching, tapping etc.). Gestures in these videos were not representational (were not fragments of gestures which were produced in reference to either shape), meaning that the 20 pairs of shapes used in critical trials were counterbalanced over only two lists rather than four. Each pair which was seen with a video of the speaker producing an adaptor gesture in one list was seen with a video of the speaker sitting motionless in the other. Because there was no point at which gestures disambiguated between shapes, the relative timings of gesture and speech were less crucial. As in Experiment 4.1, filler trials included up to 200 ms of audio and video after the point-of-truncation, and were balanced such that each participant saw an equal number of trials in which the speaker partially described an easy-to-name shape as a difficult-to-name one, within which there was an even split of presence of adaptor gesturing.

Twenty-two self-reported native speakers of English took part in the experiment in return for £4. Participants were recruited from the University of Edinburgh community, with the constraint that they had not previously taken part in other experiments presented in this thesis. All participants were right-handed mouse users with normal or corrected-to-normal vision. Consent was obtained in accordance with the University of Edinburgh's Psychology Research Ethics Committee guidelines (reference number: 99-1718/1)

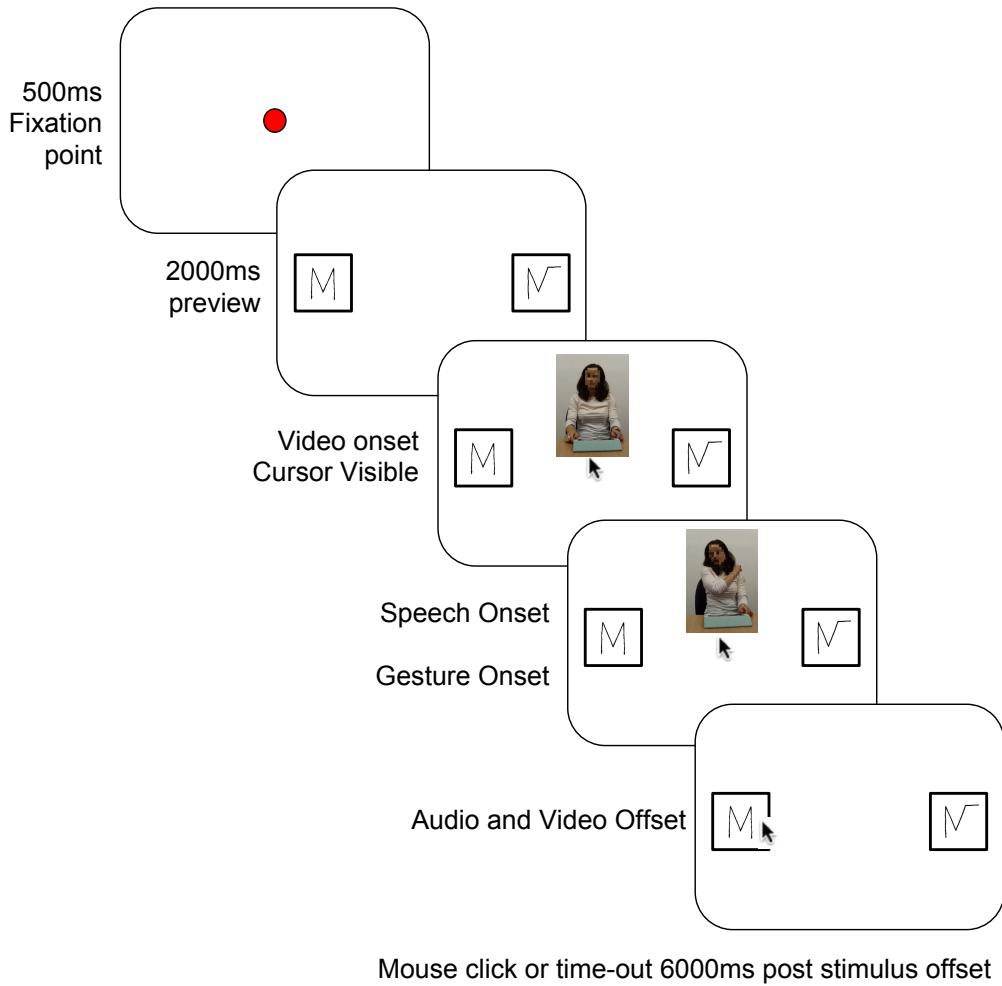


Figure 4.7: Procedure for a trial in Experiment 4.2

4.3.2 Results

Data from two participants who indicated suspicions about the origins of the stimuli were removed from the analysis. Out of the remaining 400 trials, all trials resulted in participants clicking on one of the objects. Mouse coordinate data (but not mouse clicks or eye-tracking data) from the first participant was missing due to an error in the experiment script which was fixed for subsequent participants. Mouse-tracking and time-to-click analysis was therefore run on the data from 19 participants (one fewer than analysis of eye movements of objects clicked). Of the

mouse coordinate data from these 19 participants, 0.5% of samples were excluded from analysis due to the X-coordinates being beyond the outer edge of either shape.

Analysis

As in Experiment 4.1, Mouse-clicks to shapes (easy-to-name vs. difficult-to-name) were modelled using mixed effects logistic regression, with fixed effects of gesture (No Gesture vs. Gesture, deviation coded), and random intercepts and slopes for gesture both by-participant and by-item (pair of shapes) were included. Reaction times to click on a shape (measured from the point at which audiovisual stimuli stopped) were log transformed and analysed using a mixed effects linear regression, with fixed effects of presence of gesture, object clicked, and their interaction, and random intercepts and slopes of presence of gesture by-participant and by-item. Analysis of eye- and mouse-tracking data followed the same procedure as for Experiment 4.1, above (as in Experiment 4.1). As in Experiment 4.1, only the inclusion of a quadratic term for time was found to improve model fit for the eye-tracking analysis, and only a linear term for time in the mouse-tracking analysis.

Object clicks and response times

Across the experiment, participants clicked on the difficult shape in 43.5% of critical trials and the easy shape in 56.5%. Table 4.6 shows the numbers of clicks to each type of shape split the presence of adaptor gesturing. Results revealed that participants did not show the same overall tendency as was present in Experiment 4.1 to click on the difficult-to-name shape. Participants were no more likely to click on either the easy-to-name shape or the difficult-to-name

one depending upon whether the video showed the speaker producing an adaptor gesture or sitting motionless. Analysis of time to click (measured from the offset of stimuli) revealed no effect of gesturing, but found that participants were overall slower to click on difficult-to-name shapes than easy-to-name ones ($\beta = 0.11$, SE = 0.04, $t = 2.49$) Tables 4.7 and 4.8 show full results of analyses for objects clicked and times taken to click respectively.

Table 4.6: Breakdown of mouse clicks recorded on each shape (easy-to-name or difficult-to-name) by gesture condition in critical trials in Experiment 4.2

	No Gesture	Adaptor Gesture
Clicks to Easy-to-name	120 (60.0%)	106 (53.0%)
Shape		
Clicks to Difficult-to-name	80 (40.0%)	94 (47.0%)
Shape		

Table 4.7: Model results for mouse clicks to difficult-to-name shapes over easy-to-name ones in Experiment 4.2

	β	SE	p
(Intercept)	-0.33	(0.26)	.20
Gesture	0.40	(0.26)	.13
Var(1—Participant)	0.83		
Var(Gesture—Participant)	0.04		
Var(1—Item)	0.25		
Var(Gesture—Item)	0.33		
Total	400		
Participant	20		
Item	20		

Eye movements

Figure 4.8 shows the time course of fixations to all objects in the display (video, easy-to-name shape, difficult-to-name shape) for Experiment 4.2 split by the

Table 4.8: Model results for times taken to click the mouse in Experiment 4.2

	β	SE	t
(Intercept)	7.52	(0.06)	126.27
Gesture	0.03	(0.04)	0.70
Difficulty-to-name of clicked shape	0.11	(0.04)	2.49
Gesture × Difficulty-to-name of clicked shape	-0.09	(0.08)	-1.19
Var(1—Item)	0.01		
Var(Gesture—Item)	0.00		
Var(1—Participant)	0.05		
Var(Gesture—Participant)	0.00		
Total	380		
Item	20		
Participant	19		

presence of adaptor gesturing. Analysis of the time window beginning at 600 ms preceding stimulus offset and extending for 1200 ms revealed no overall bias towards either object over this window, nor any linear or quadratic effects of time. Following videos of the speaker producing an adaptor gesture, a significant quadratic effect of time ($\beta = -0.85$, SE = 0.38, $t = -2.26$) indicated a tendency to fixate the difficult object over the easy one in the middle of this window. Figure 4.9 shows the fitted values from the model and the empirical logit bias towards difficult-to-name objects over easy-to-name ones over the relevant window of analysis, and Table 4.9 shows the model results.

Mouse movements

Figure 4.10 shows the time course of mouse movements towards easy-to-name and difficult-to-name shapes in Experiment 4.2 split by the presence of adaptor gesturing. Analysis of the 1200 ms period centered on the offset of stimuli revealed no overall bias to move more towards either easy or difficult-to-name objects, nor any effect of time, presence of gesture, nor their interaction (full model results are shown in Table 4.5).

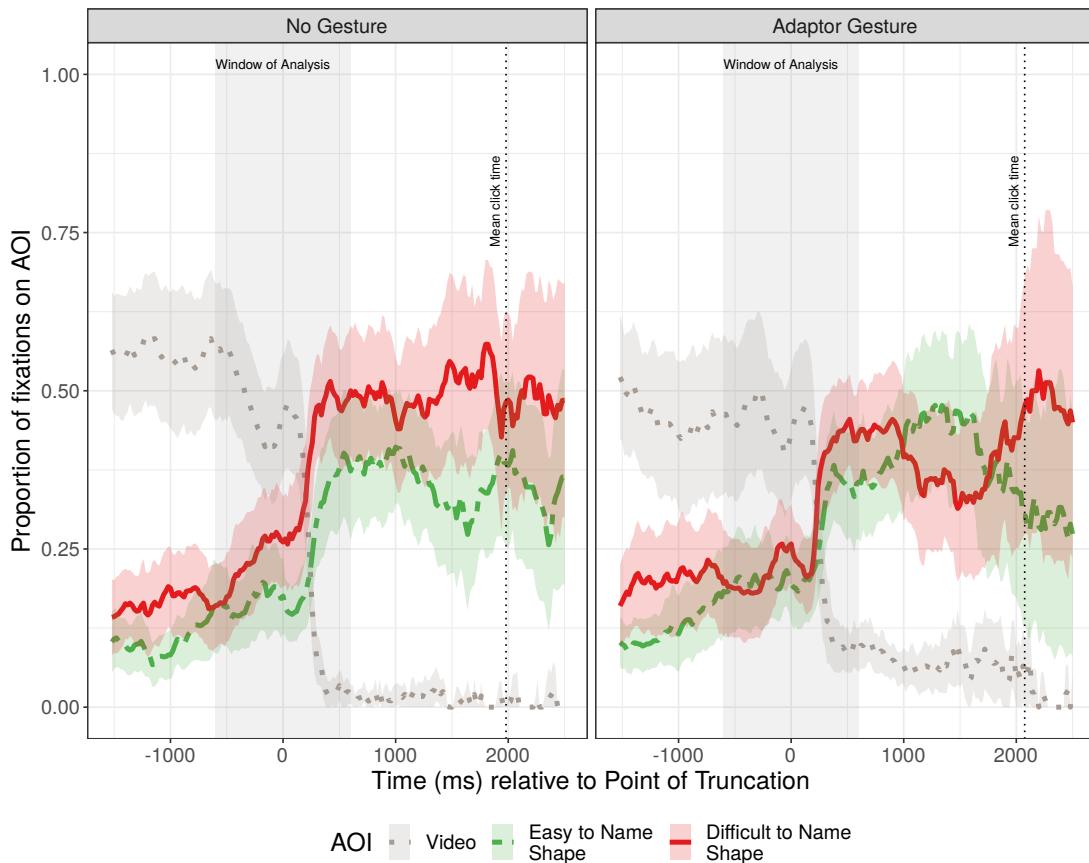


Figure 4.8: Eye-tracking results for critical trials in Experiment 4.2: Proportion of fixations to each object (easy or difficult-to-name shape) and the video, from speech onset to 2000 ms post stimulus offset, calculated out of the total sum of fixations for each 20 ms time bin. Shaded areas represent 95% confidence intervals derived via bootstrapping subject data ($R=1000$).

4.4 Discussion

Experiments 4.1 and 4.2 investigated whether speakers' non-verbal behaviours influence listeners' explicit predictions about the relative conceptual difficulty of the object they are about to refer to, focussing on iconic and adaptor gestures respectively. Presenting participants with the initial, ambiguous fragments of audio and video of a speaker providing an instruction to click on an object, we



Figure 4.9: Empirical logit transformed proportion bias towards difficult-to-name shapes over easy-to-name shapes in critical trials in Experiment 4.2 for the 1200 ms window centered on stimulus offset. Lines represent fitted values of the model.

tasked participants with deciding (via mouse click) which of an easy-to-name shape and a difficult-to-name one they thought the speaker was about to mention. We manipulated whether the video stimuli showed the speaker producing a gesture or sitting motionless, and ensured that critical trials contained no information in either speech or gesture (when present) which disambiguated between the two objects.

Listeners' predictions of upcoming referents were found to be influenced by the presence of iconic gesturing (Experiment 4.1) but not adaptor gesturing

Table 4.9: Model results for eye- and mouse-tracking analysis over the 1200 ms window centered on point-of-truncation in Experiment 4.2

	Fixations			Mouse Movements		
	β	SE	t	β	SE	t
(Intercept)	0.36	(0.18)	1.93	-0.03	(0.14)	-0.21
Gesture	0.33	(0.29)	1.11	0.06	(0.25)	0.24
Time	1.52	(1.01)	1.51	-0.28	(0.22)	-1.28
Time ²	-0.33	(0.86)	-0.39			
Gesture × Time	0.35	(0.38)	0.92	-0.15	(0.22)	-0.67
Gesture × Time ²	-0.85	(0.38)	-2.26			
Var(residual)	14.01			4.45		
Var(1—Participant)	0.46			0.16		
Var(Gesture—Participant)	0.89			0.49		
Var(Time—Participant)	12.61			0.34		
Var(Time ² —Participant)	6.49					
Var(1—Item)	0.21			0.20		
Var(Gesture—Item)	0.78			0.73		
Var(Time—Item)	7.07			0.35		
Var(Time ² —Item)	7.58					
Total	24187			9499		
Participant	20			19		
Item	20			20		

(Experiment 4.2). Participants appeared to associate the presence of iconic gesturing with descriptions of more conceptually difficult objects, as indicated by their tendency to click on these objects when gesturing was present in the video. This finding suggests that, in certain situations, listeners interpret a speaker's production of iconic gesturing as a sign that they are experiencing difficulty in producing a verbal description, much like they do with the production of speech disfluency (Arnold et al., 2007). Interestingly, the tendency to predict the speaker to be about to refer to the more difficult-to-name shape was found to increase with greater durations of gesturing, patterning with previous research which points towards the relative durations and onsets of speech and gesture varying in accordance with conceptual difficulty or lexical familiarity (see Experiment 3.1 and Morrel-Samuels and Krauss 1992). Moreover, participants were faster to click

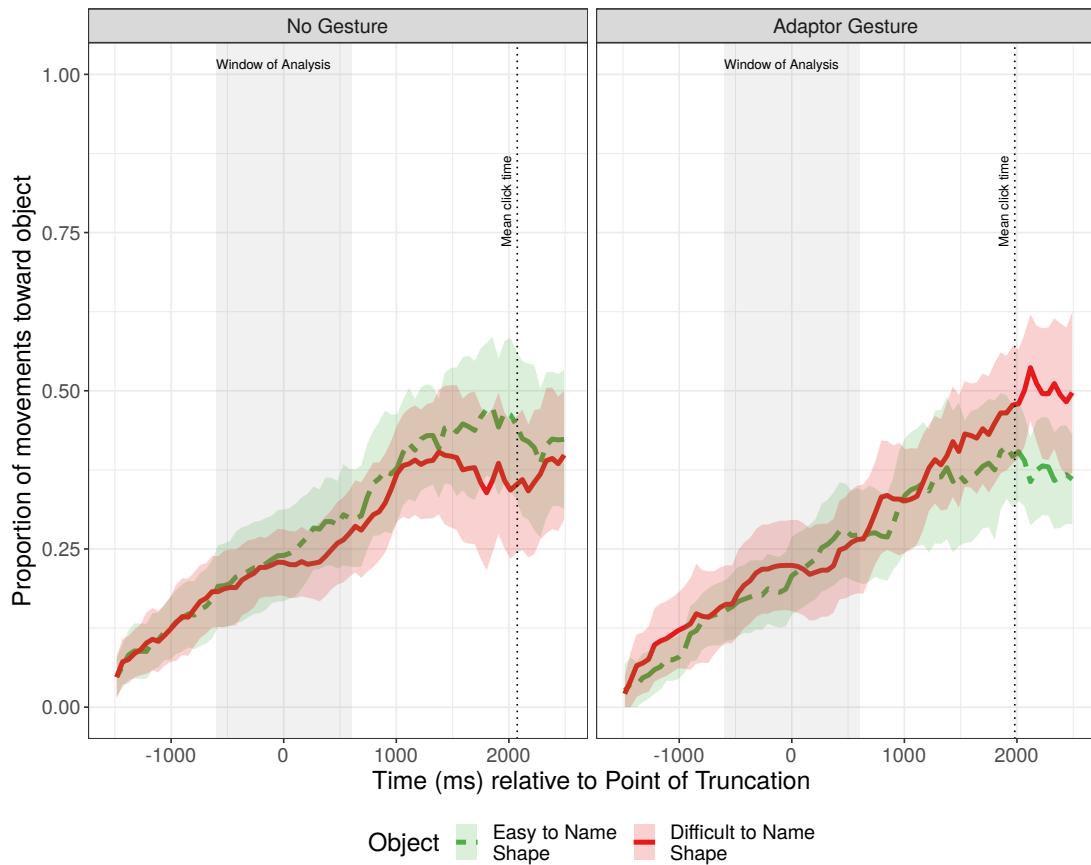


Figure 4.10: Mouse-tracking results for critical trials in Experiment 4.2: Proportion of cumulative distance travelled toward each object from speech onset to 2000 ms post stimulus offset. Proportions were calculated from the total cumulative distance participants moved the mouse until that time bin. Shaded areas represent 95% confidence intervals derived via bootstrapping subject data ($R=1000$).

on an object following an utterance fragment presented with iconic gesturing, suggesting that the occurrence of gesture facilitated the decision making processes, in keeping with research which points found that questions with gestures result in faster responses (Holler, Kendrick, & Levinson, 2017).

The fact that listeners' predictions about upcoming referents were not influenced by whether or not the speaker was seen to produce an adaptor gesture indicates that in forming these predictions listeners are sensitive to the type of gesture produced,

and not simply associating any movement on the speaker's part as indicative of speech planning difficulty. The lack of any association drawn by listeners between adaptor gesturing and more difficult-to-name referents may reflect the fact that the production of this type of gesture is not so strongly associated with such objects in natural communication. Experiment 3.1, for example, found no difference in the use of adaptor gesturing depending upon whether speakers were describing easy-to-name or difficult-to-name shapes. Alternatively, the range of possible speaker-internal states (emotional and cognitive) signalled by adaptor gesturing may result in a weakening of any tendency to attribute such gestures to one specific cause.

To explore the possibility of developing a paradigm in which it is possible to detect if the presence of gesturing influences listeners to formulate expectations about upcoming referents in real-time (alongside the presentation of the stimulus, as in the eye-tracking task in Arnold et al. 2007), we also measured participants' eye and mouse movements in both experiments presented here. Over a time-window of 1200 ms centered around the point at which audio and video stopped, the presence of gesture in the video had little effect, only influencing the curvature of an increasing bias to fixate the difficult-to-name shape over time which was present in both conditions (Gesture vs. No Gesture). Notably, this effect in Experiment 4.2 was in the opposite direction to that found in Experiment 4.1: Whereas iconic gesturing was associated with a steadier increase (relative to no gesture) in the bias to fixate the difficult-to-name shape, Experiment 4.2 found adaptor gesturing to result in a steeper initial increase in this bias relative to no gesture trials. It is hard to reconcile this difference with our initial explanation that the relative visual salience of gestures in Experiment 4.1 delayed participants from fixating to either shape.

Mouse-tracking analyses revealed that across the same time-window (600 ms either side of the offset of speech and gesture) there was no evidence of effects of the

presence of either type of gesturing on participants' mouse movements. Visual inspection of the wider pattern of mouse movements suggest that the tendency to click on difficult-to-name shapes following speech with gesture in Experiment 4.1 may be present in the time course of mouse-trajectories, but only emerging late on. This may in part be due to the lack of time pressure to make the decision to click on an object in the display: Participants were given 6 seconds from the offset of audio and video to click on an object. It is possible that by framing the task as a reaction time game, in which participants are tasked with clicking on the object named by the speaker as quickly as possible, it would be possible to provide a parallel study to Arnold et al.'s eye-tracking task and establish whether the presence of gesturing, like the presence of speech disfluency, guides listeners' real-time anticipations of upcoming referents. The next chapter aims to investigate more thoroughly whether the association between iconic gesturing and more difficult referents guides listeners' anticipations about upcoming referents alongside the moment-to-moment processing of the multi-modal input by introducing time pressure into the experimental paradigm. It also presents participants with full audiovisual utterances (as opposed to being truncated at the point of disambiguation as in the current chapter), to investigate whether this association holds in a more naturalistic setup (i.e., in a situation where they are not explicitly tasked with forming such predictions).

Chapter 5

Gesture modulating listener expectations in real time

In everyday communication people often display an ability to predict what someone else will say, for instance in that interlocutors often finish one another's sentences (Clark & Wilkes-Gibbs, 1986). The ability to anticipate upcoming message content has been studied at various levels: Predictive accounts of comprehension of language hold that the preactivation of upcoming items or features (e.g., shape, see Rommers, Meyer, Praamstra, & Huettig, 2013) is constrained by aspects of language ranging from syntactic structure and sentential context (e.g., Altmann & Kamide, 1999; Kamide et al., 2003) to common ground (Keysar, Barr, Balin, & Brauner, 2000). As well as the words themselves, there are many other aspects of communication which may help this predictive process. Speakers may, for instance, vary their non-linguistic behaviours during production of speech, and in doing so signal information about their message.

Drawing on research which has shown that listeners are sensitive to some of these non-linguistic behaviours in the audio-channel (specifically speech disfluencies),

the previous chapter explored the possibility that this sensitivity extends to the non-verbal delivery of an utterance. Specifically, the experiments presented in Chapter 4 investigated whether certain types of gesturing could be interpreted as signals that the speaker was about to refer to something which was more difficult to verbally encode.

Both Experiments (4.1 and 4.2) were gated-tasks, presenting participants with truncated instructions to click on one of two objects in a display (one easy-to-name shape, one difficult-to-name shape). Instructions were multi-modal, presented in audio and video. We manipulated whether the video component showed the speaker sitting motionless or producing a ambiguous fragment of iconic gesturing (Experiment 4.1) or adaptor gesturing (Experiment 4.2). Relative to speech without gesture, speech with iconic gesturing resulted in more predictions that the speaker was about to refer to the more difficult-to-name of two shapes. The presence of adaptor gesturing was not found to have an effect of participants' guesses about upcoming referents. Parallel to research suggesting that listeners associate speech disfluency with less familiar objects (see Arnold et al., 2007), the results of Experiment 4.1 suggest that, in some situations, listeners perceive the act of producing an iconic gesture as a signal that they are experiencing difficulty in planning and producing speech.

The extent to which listeners might have formulated hypotheses based on the occurrence of iconic gesturing in an on-line manner—i.e., holding probabilistic expectations about upcoming content alongside the immediate processing of the linguistic and gestural input—was less clear. The increased likelihood to predict a speaker to be referring to a difficult-to-name shape following iconic gesturing (Experiment 4.1) appeared to be reflected later on in listeners' mouse-movements (see Figure 4.6). In Experiment 5.1, we extend Experiment 4.1 to include the full (non-truncated) utterances and videos, and introduce an element of time pressure,

the lack of which in Experiment 4.1 we suggested may have resulted in the late emergence of an effect of gesturing.

In Experiment 5.1, participants are presented with full instructions to click on an object, and are tasked with clicking on the shape described by the speaker as quickly as possible. As in Experiment 4.1, we manipulate whether the video component of the instruction presents an iconic gesture or presents no gesture. Instructions are temporarily ambiguous: Before a certain point (the point where stimuli were truncated in Experiment 4.1), both speech and gesture (when present) contain no information which unambiguously refers to either shape. By studying participants' eye and mouse movements during the period prior to this point, we aim to establish whether an associations between iconic gesturing and difficult-to-name shapes is reflected in listeners' on-line predictions about the unfolding expression. Results are inconclusive, but offer insight into some of the problems involved in studying gestures which redundantly co-express the same content as speech.

5.1 Signals guiding on-line expectations

A growing body of work is showing that the way in which speech is delivered can have an important influence in guiding comprehension. For example, speech disfluencies (which tend to occur before low frequency words, see Beattie 1979) have been shown to prepare listeners for less predictable words (as evidenced by a reduced N400 effect, see Corley et al., 2007) and objects which are unfamiliar or new to the discourse (Arnold et al., 2007, 2004). Similarly, the prosody of speech has been shown to influence the syntactic structure which listeners initially assign to an utterance (see Kjelgaard & Speer, 1999).

Comparatively, the effects of manner of non-verbal delivery (i.e., the accompaniment of gestures, facial expressions and other movements) on listeners' on-line comprehension has received little attention. A small number of studies have shown that semantic content contained in speakers' gestures has an immediate effect on listeners' comprehension, pointing towards integration of information in the two modalities occurring at the early stages of the comprehension process (see Kelly et al., 2004; Özyürek et al., 2007; Wu & Coulson, 2005). One such study conducted by Holle and Gunter (2007) shows how a speaker's gestures can influence listeners' activations of less frequent word meanings, thus facilitating subsequent disambiguation. In Holle and Gunter's (2007) ERP study, participants heard utterances which contained ambiguous homonyms, and saw videos of the speaker producing a gesture which supported either the dominant or subordinate meaning (for example, the utterances in 1 were presented alongside 'ball' with gestures representing either serving a ball or ballroom dancing).

- (1a) **Dominant:** Sie kontrollierte den Ball_{amb} was sich im [spiel beim aufschlag deutlich zeigte]_{disambiguation}
- (1b) Translation: She controlled the ball_{amb} which during the [game at the serve clearly showed]_{disambiguation}
- (1c) **Subordinate:** Sie kontrollierte den Ball_{amb} was sich im [tanz mit brautigam deutlich zeigte]_{disambiguation}
- (1d) Translation: She controlled the ball_{amb} which during the [dance with the bridegroom clearly showed]_{disambiguation}

Holle and Gunter (2007) found that the content of the gesture influenced the size of the N400 effect at the point at which speech subsequently disambiguated between the two possible meanings ("game/dance"), suggesting that listeners' expectations of upcoming speech are based on their interpretations of a speaker's previous gestures.

Studies such as Holle and Gunter (2007) show that the information represented in a gesture and not in speech—i.e., the semantic content of a non-redundant (at the point in the utterance at which it is presented) gesture—is integrated along a similar time frame to the content of speech. However, with the exception of the experiments in Chapter 4, to our knowledge no studies have investigated whether the occurrence of gesturing (and not just its content) guides listeners’ predictions in an on-line manner. The mechanism by which this might be possible has been widely studied in relation to speech disfluency, with both eye- and mouse-tracking studies showing that listeners display biases to predict discourse new (Arnold et al., 2004; Barr & Seyfeddinipur, 2010) and difficult to describe (Arnold et al., 2007) objects which emerge alongside the perceptual input of speech. Results from the previous chapter indicate that listeners do associate the presence of iconic gesturing with more difficult-to-name shapes when tasked with forming these predictions explicitly (by clicking on the object they thought the speaker was about to refer to), which may indicate a perception of gesturing as a signal that the speaker is encountering difficulty in planning or producing speech.

As previously discussed in Chapter 4, in investigating the disfluency~difficulty bias, Arnold et al. (2007) conducted both a gating task—in which participants heard ambiguous fragments of speech which were either fluent or disfluent—and an eye-tracking task, in which they heard full instructions. In the gating task, participants made explicit predictions about upcoming content by clicking on the object they believed the speaker to be about to refer to. In the eye-tracking task, listeners simply had to click on the object named by the speaker, with results revealing a tendency to fixate on the less familiar in the period following speech disfluency but prior to the referent-noun onset.

Experiments 4.1 and 4.2 presented gesture-based parallels to Arnold et al.’s gating task in that they presented listeners with ambiguous fragments of speech with and without ambiguous fragments of gesturing, tasking them with click on the object

they believed the speaker to be about to refer to. We also recorded participants' eye and mouse movements throughout this experiment, with a possible late effect of iconic gesturing being evident (visual inspection of Figures 4.6 and 4.4). However, these measures reflect how participants behaved when tasked with forming explicit predictions, meaning that comparisons to the literature on the on-line attribution of speech disfluency (Arnold et al., 2007, 2004; Barr, 2001, e.g.,) are limited.

To provide a complement to Arnold et al.'s (2007) eye-tracking task, Experiment 5.1 aims to establish whether the association between iconic gesturing and difficult-to-name shapes found in Experiment 4.1 is present in the predictions listeners make during the moment-to-moment processing of full utterances and iconic gestures rather than only the initial ambiguous fragments. To increase the likelihood of any anticipatory fixation or mouse-movement biases being detectable alongside the unfolding speech and gesture, we introduce a time pressure to the task.

Framed as a reaction time game, Experiment 5.1 tasks participants with clicking on the object which is described by the speaker as quickly and accurately as possible. As before, the point-of-disambiguation occurs simultaneously in both speech and gesture, meaning that up until this point the linguistic and gestural content of a critical trial is (temporarily) ambiguous between the easy-to-name and difficult-to-name shape. By studying participants' fixations and movements of the mouse towards these shapes in a pre-disambiguation window, we aim to investigate whether listeners form on-line expectations about upcoming content based on the presence of iconic gesturing, even when the content of speech (and gesture, when present) will subsequently disambiguate.

Additionally, by studying the time-course of these measures in the period following disambiguation, as well as participants' response times and error rates (as indicated by mouse clicks on objects), we aim to explore how gesture influences reference comprehension, for instance by resulting in faster responses. The task is framed

as a reaction time game with the aim of encouraging participants to fixate on the shapes rather than the video during in the critical periods of interest. Results found no evidence to suggest that listeners' on-line predictions of upcoming content were influenced by the presence of temporarily ambiguous iconic gesturing, with these measures possibly confounded by participants' task strategies during the experiment. Results from the period following disambiguation suggest that while the presence of gesturing in the video may have delayed participants from fixating on (and moving the mouse towards) the shapes, it ultimately had a facilitatory effect on their response times.

5.2 Experiment 5.1

Experiment 5.1 presented participants with instructions to click on one of two objects (out of an easy-to-name shape or a difficult-to-name shape). The experiment was framed as a reaction time game, and participants were tasked with clicking on the object named by the speaker as quickly as possible. As in Experiment 4.1, critical trials presented with the point-of-disambiguation simultaneously in speech and gesture (when present), meaning that prior to this point there was no information in either modality which unambiguously referred to one of the objects in the display—i.e., linguistic and gestural input was temporarily ambiguous. In contrast to Experiment 4.1, where audio and video were truncated at point-of-disambiguation, Experiment 5.1 presented participants with the subsequent disambiguation information in speech and gesture (when present). Eye and mouse coordinates were tracked over the course of the experiment, along with mouse-click responses (shape clicked) and response times. If the presence of iconic gesturing influences listeners' on-line predictions of upcoming content, we would expect more fixations and mouse movements towards the difficult-to-name shapes than the easy-to-name ones following videos of gesturing (as opposed to

those showing no gesture). Additionally, if gesturing facilitates comprehension, we would expect faster response times and fewer errors following videos of gesturing.

5.2.1 Method

Forty participants were recruited from the University of Edinburgh community, in return for a payment of £4 or university credit, for a desired sample size of 24 participants who were both self-reported native speakers of English and believed the cover story about the experimental stimuli.¹ No participants had taken part in any of the other experiments presented in this thesis. All participants were right-handed mouse users with normal or corrected-to-normal vision. Consent was obtained in accordance with the University of Edinburgh's Psychology Research Ethics Committee guidelines (ref number: 227-1617/1) The experiment was pre-registered at <https://osf.io/t68be/>

Materials

The materials in Experiment 5.1 were taken from the same video recordings as those used in Experiment 4.1. Whereas in Experiment 4.1 the stimuli were truncated at the point at which speech and iconic gesture disambiguated between objects on the screen, Experiment 5.1 presented participants with the full instructions (in both audio and video) to click on one of the objects. Experiment 5.1 followed the same two (Gesture vs. No Gesture) by two (Easy vs. Difficult shape) design as Experiment 4.1. However, this latter manipulation was not (as it was in Experiment 4.1) simply a precautionary stimulus check for sensitivity to how gesture fragments might differ depending on whether they were taken from gestures

¹The large number of excluded participants is due to the recruitment procedure not discriminating between native and non-native speakers.

to easy or difficult shapes: In the current experiment, stimuli were presented beyond the point-of-disambiguation, at which point linguistic and gestural content varied across the Easy vs. Difficult manipulation.

Critical trials used the same initial utterance fragment of “The one you should click on is the” followed by a spoken description of the referent. Verbal descriptions of easy shapes were either of the form “letter …”, “number …” or named the geometrical shape or symbol (e.g., “rectangle”, “ampersand”). Verbal descriptions of difficult shapes described the relative orientation of lines which made up the shape, and care was taken to ensure that descriptions of easy and difficult-to-name pairs did not share the same onset.

One notable change to the stimulus presentation from Experiment 4.1 was that video and speech began simultaneously in all trials. This meant that, for some items, the video appeared with the speaker mid-way through the preparation phase of the gesture.² This was done to introduce more consistency across trials as to the beginning of the speaker’s instructions, to avoid the possibility that any variation in the durations from video onset to speech onset could be interpreted as the time taken for the speaker to initiate speech (and so linked to speech planning difficulty).

As in Experiment 4.1, 20 critical trials were counterbalanced across four lists, each including 10 Gesture trials (five referring to the easy-to-name shape, five referring to the difficult-to-name one) and 10 No-Gesture trials (five easy, five difficult). Filler trials remained the same as in Experiment 4.1 apart from being extended to present the entire utterance and gesture.

²Care was taken to ensure that this was always before the first stroke or hold phase.

Procedure

Experiment 5.1 was presented using OpenSesame version 3.1 (Mathôt et al., 2012), displayed on a 21 in. CRT monitor with a resolution of 1024×768 , placed 850 mm from an Eyelink 1000 Tower-mounted eye-tracker which tracked eye movements at 500 Hz (right eye only). As before, audio was sampled at 44100 Hz, video at 20 fps and mouse coordinates every frame of the video (every 50 ms).

Participants were told the same cover story as for Experiment 4.1, and the same post-test questioning was conducted. As before, data from any participant who indicated suspicion that the stimuli were artificially constructed and that speech and gesture had not been produced simultaneously were excluded from analysis. The experiment was framed as a reaction time game - participants were tasked with clicking on the object described by the speaker as quickly as possible, and were informed that they would receive a score for how fast they responded on each trial. As an incentive, participants were shown a scoreboard containing fake scores at the start of the experiment, and were told that they would be able to add their final score to the scoreboard at the end of the experiment.

Figure 5.1 shows the procedure of a trial in Experiment 5.1. As in Experiment 4.1, trials began with a manual drift correction, after which the two objects were displayed. The duration of this display was extended from 2000 to 3000 ms, with the aim of giving participants more time to study and become familiar with the shapes, hopefully reducing any effect of the visual salience of difficult objects. After 3000 ms the video appeared, audio began playing, and the cursor was made visible and centered. Speech and gesture were presented in full. Video clips stopped at the end of the gesture, at which point the last frame of the video (which showed the speaker having completed the gesture and with her hands returned to a resting position on her legs) was presented until either participants clicked on an object or the trial timed out (6000 ms after point-of-disambiguation).

A substantial change from the procedure of Experiment 4.1 was that participants received feedback on their response times between trials. The aim of this was to motivate participants to respond quickly (and thereby fixate on the shapes rather than the video). Reaction times were split into four grades, each corresponding to an animal associated with varying degrees of speed (Snail; Tortoise; Monkey; Cat). After each trial, participants were told which category their response fell into, along with their overall score. The first time a response equated to a given category, they saw a video of that animal along with their score. Subsequent feedback presented only a cartoon picture of the relevant animal. If participants clicked on the wrong object, their response was categorised as a panda and they received feedback accordingly. If participants clicked on the correct object before point-of-disambiguation, their response fell into the grade of *wizard*.

Eye movements, mouse coordinates and object clicked (Easy vs. Difficult object, or alternatively, Referent vs. Competitor) were recorded for each trial. After the experiment, participants answered a short questionnaire, including one question aimed at establishing whether they suspected the proposed origins of the stimuli. A further question asked participants if they had found anything in the stimuli which they felt helped them to respond quickly.

5.2.2 Results

Analysis

Analysis was carried out in R version 3.5.1 (R Core Team, 2018), using the lme4 package version 1.1-17 (Bates et al., 2015). Data from 15 non-native English speaking participants was removed from all analyses, leaving data from 25 participants. In all 500 analysed trials, participants clicked on one of the objects within the time limit. Mouse movements beyond the outer edge of either object

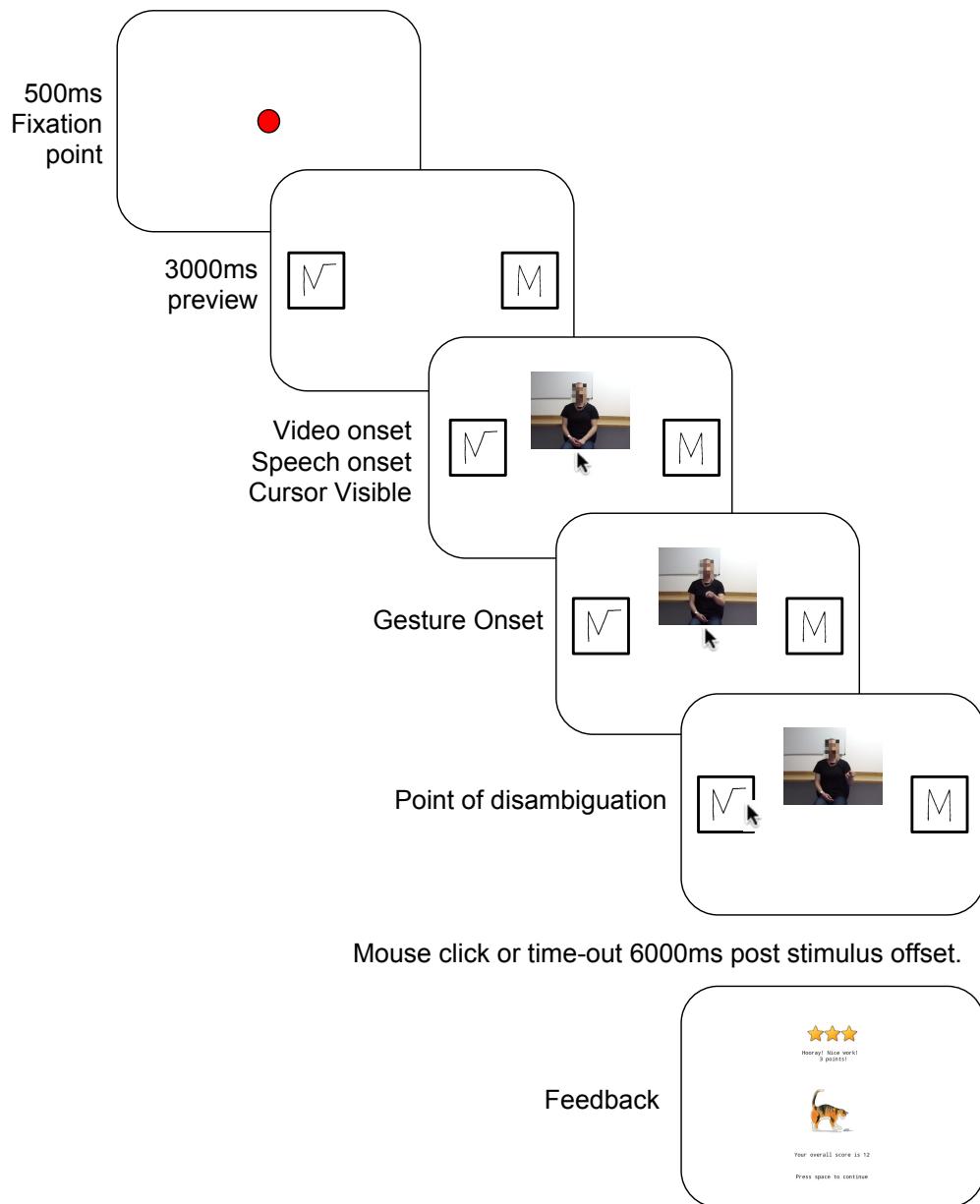


Figure 5.1: Procedure for a trial in Experiment 5.1

made up 0.8% of samples, and were considered overshooting and therefore excluded from analysis.

Analysis of fixations and mouse movements was conducted on two time windows: From 600 ms prior to point-of-disambiguation up until 200 ms post-disambiguation,

and from 200 ms to 1000 ms post-disambiguation. The first window covers the region during which any predictions listeners might formulate about upcoming message content based on ambiguous input might emerge. The second window covers how participants respond to the unfolding disambiguation in both speech and gesture (when present), and continues to just beyond the average response time (920 ms after disambiguation).

For analysis of the first time window, the proportions of fixations and mouse movements towards either object were calculated and subsequently transformed using the empirical logit transformation as in Experiment 4.1, reflecting biases towards the difficult-to-name shape over the easy-to-name one. These were then modelled analogously to Experiment 4.1, using linear mixed effects regression with fixed effects of gesture (Gesture vs. No Gesture, deviation coded), Time (Z-scored) and their interaction. For both eye-tracking and mouse-tracking models, higher order polynomials for time did not significantly improve model fit (as indicated by likelihood ratio tests and Bayesian information criterion [BIC]) and were not included. Random intercepts and slopes of gesture and time were specified both by-item (pair of shapes) and by-participant.

Analysis over the second window was conducted on the empirical logit transformed proportions of fixations and mouse movement biases towards the referent (the shape described by the speaker) over the competitor (shape not described). These fixation and mouse movement biases were modelled using linear mixed effects regression models with fixed effects of gesture (Gesture vs. No Gesture, deviation coded), nameability of the referent (Easy-to-name vs. Difficult-to-name, deviation coded), and both eye- and mouse-tracking models included orthogonal linear, quadratic and cubic terms for time. The number of higher order polynomials for time included in models was based on whether their inclusion improved model fit as indicated by both likelihood ratio tests and a decrease of and a decrease of ≥ 10 in BIC (following Raftery, 1995). Only linear and quadratic effects of

time and their interactions will be interpreted, with higher order terms reflecting less relevant effects in the tails (see Mirman et al., 2008). Three-way interactions between gesture, referent nameability and each degree of time were included as fixed effects. Random intercepts and effects of gesture, referent nameability, and linear and quadratic terms for time were included both by-participant and by-item. Full model specification can be found in the results section below.

Incorrect mouse clicks were modelled using mixed effects logistic regression, with gesture and referent nameability as fixed effects, and by-participant random effects of gesture and referent nameability, and random intercepts by-participant and by-item. Due to the possibility of participants clicking on an object before point-of-disambiguation reflecting a valid prediction of upcoming referents, we measured reaction times from speech onset (rather than from the point of truncation used in Experiment 4.1). These were log transformed and modelled using mixed effects linear regression, with fixed effects of gesture, nameability of the referent, and their interaction, and random intercepts and slopes of gesture by-participant, and random intercepts by-item. As in Experiments 4.1 and 4.2, we considered effects in these models to be significant where $|t| > 2$ (see Baayen, 2008).

5.2.3 Pre-disambiguation

Eye movements

The time course of fixations to all items in the display (easy- and difficult-to-name shapes and the video of the speaker), from onset of speech and extending to just beyond the average time taken to click the mouse, can be seen for trials referring to easy-to-name shapes and difficult-to-name shapes in Figures 5.2 and 5.3 respectively.

Analysis of the window beginning at 600 ms before point-of-disambiguation and extending for 800 ms revealed that participants tended to fixate more upon the easy object than the difficult one as the window progressed (as indicated by a main effect of time $\beta = -1.46$, SE = 0.69, $t = -2.11$). There was no evidence to suggest an effect of the presence of gesture, nor an interaction of gesture and time. Full results are shown in Table 5.1.

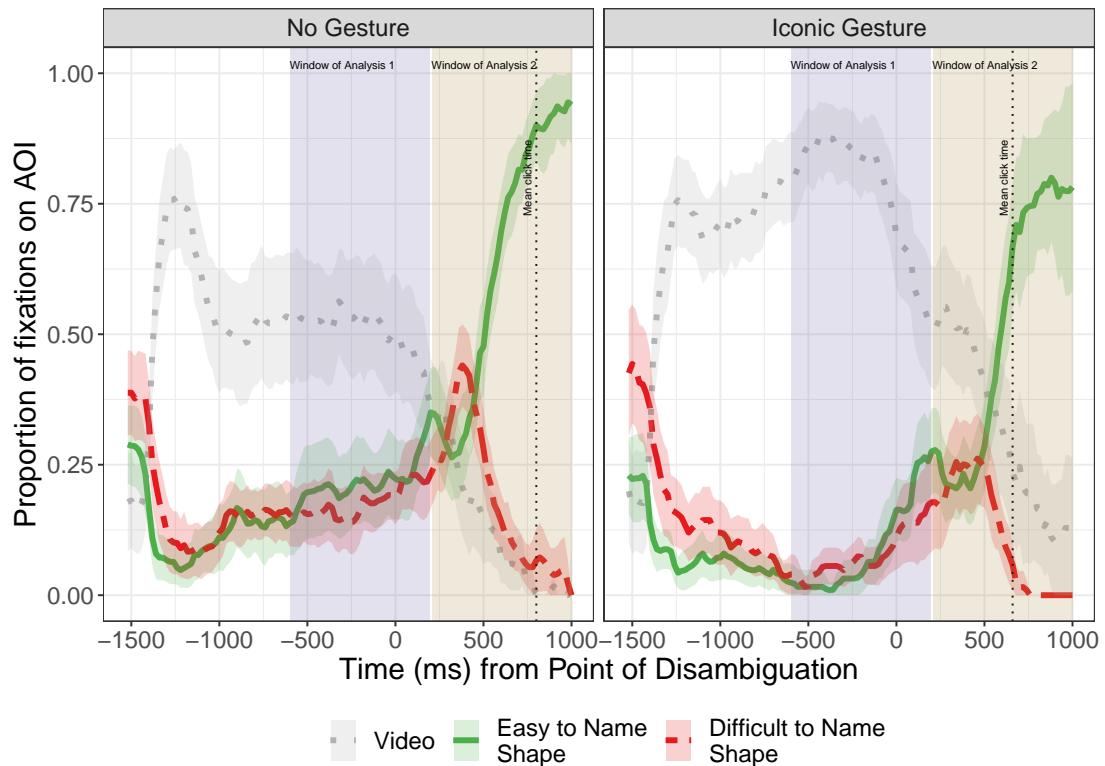


Figure 5.2: Eye-tracking results for critical trials that name an easy-to-name shape in Experiment 5.1: Proportion of fixations to each object (easy-to-name and difficult-to-name shapes) and the video, from speech onset to 1000 ms post-disambiguation, split by presence of gesturing. Proportions calculated out of the total sum of fixations for each 20 ms time bin. Shaded areas represent 95% confidence intervals derived via bootstrapping subject data ($R=1000$).

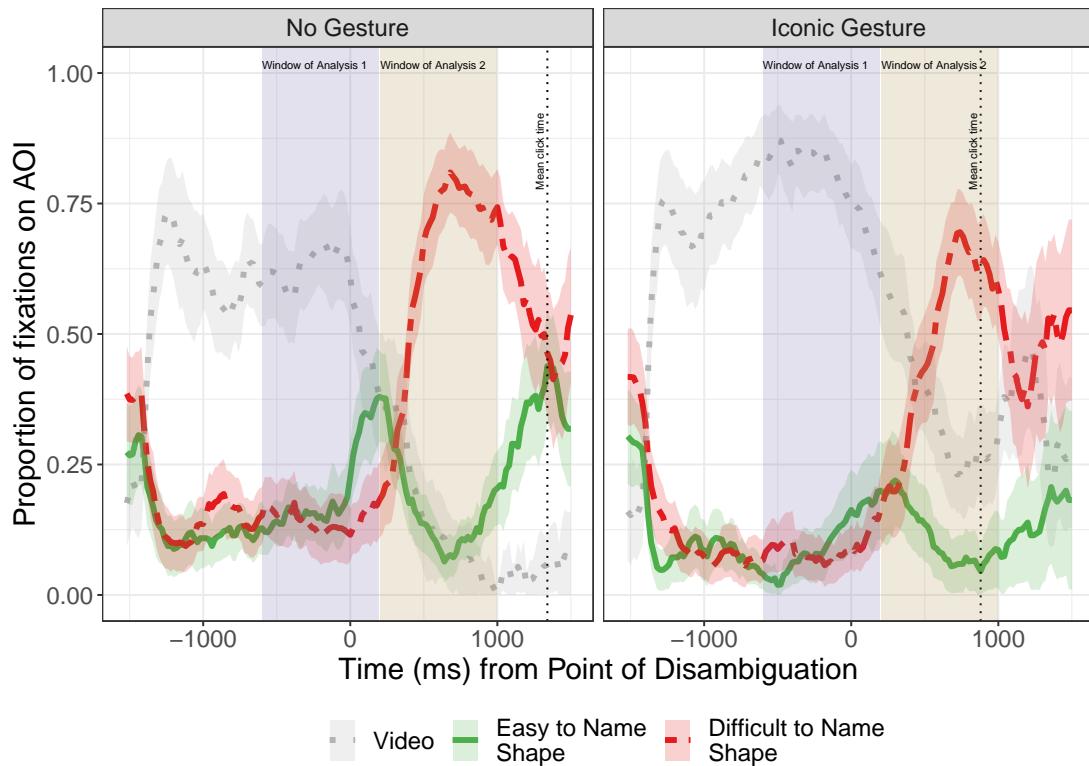


Figure 5.3: Eye-tracking results for critical trials that name a difficult-to-name shape in Experiment 5.1: Proportion of fixations to each object (easy-to-name and difficult-to-name shapes) and the video, from speech onset to 1500 ms post-disambiguation, split by presence of gesturing. Proportions calculated out of the total sum of fixations for each 20 ms time bin. Shaded areas represent 95% confidence intervals derived via bootstrapping subject data ($R=1000$).

Mouse movements

The time course of mouse movements towards easy-to-name and difficult-to-name shapes for the period from speech onset and extending to just beyond the average time taken to click the mouse can be seen for trials referring to easy-to-name shapes and difficult-to-name shapes in Figures 5.4 and 5.5 respectively.

Over the 800 ms period commencing at 600 ms prior to point-of-disambiguation, a

Table 5.1: Model results for eye- and mouse-tracking analysis in Experiment 5.1 over the period from 600 ms before point-of-disambiguation to 200 ms after.

	Fixations			Mouse Movements		
	β	SE	t	β	SE	t
(Intercept)	-0.163	(0.130)	-1.26	-0.586	(0.198)	-2.96
Gesture	0.195	(0.260)	0.75	-0.068	(0.314)	-0.22
Time	-1.459	(0.690)	-2.11	-0.518	(0.148)	-3.50
Gesture \times Time	-0.064	(0.242)	-0.27	-0.395	(0.201)	-1.96
Var(residual)	7.222			4.971		
Var(1—Participant)	0.209			0.723		
Var(Gesture—Participant)	0.954			0.934		
Var(Time—Participant)	6.337			0.222		
Var(1—Item)	0.162			0.193		
Var(Gesture—Item)	0.558			1.172		
Var(Time—Item)	4.157			0.058		
Total	20327			7407		
Participant	25			25		
Item	20			20		

significant intercept indicated that participants showed an overall tendency across this window to have moved the mouse more towards the more easily named shape ($\beta = -0.59$, SE = 0.20, $t = -2.96$). Patterning with eye movements over this window, participants showed an increasing bias to move the mouse towards the easy-to-name shape over the difficult-to-name shape as the window progressed ($\beta = -0.52$, SE = 0.15, $t = -3.50$). There was no evidence to suggest an effect of gesture nor an interaction between gesture and time. Full results are shown in Table 5.1.

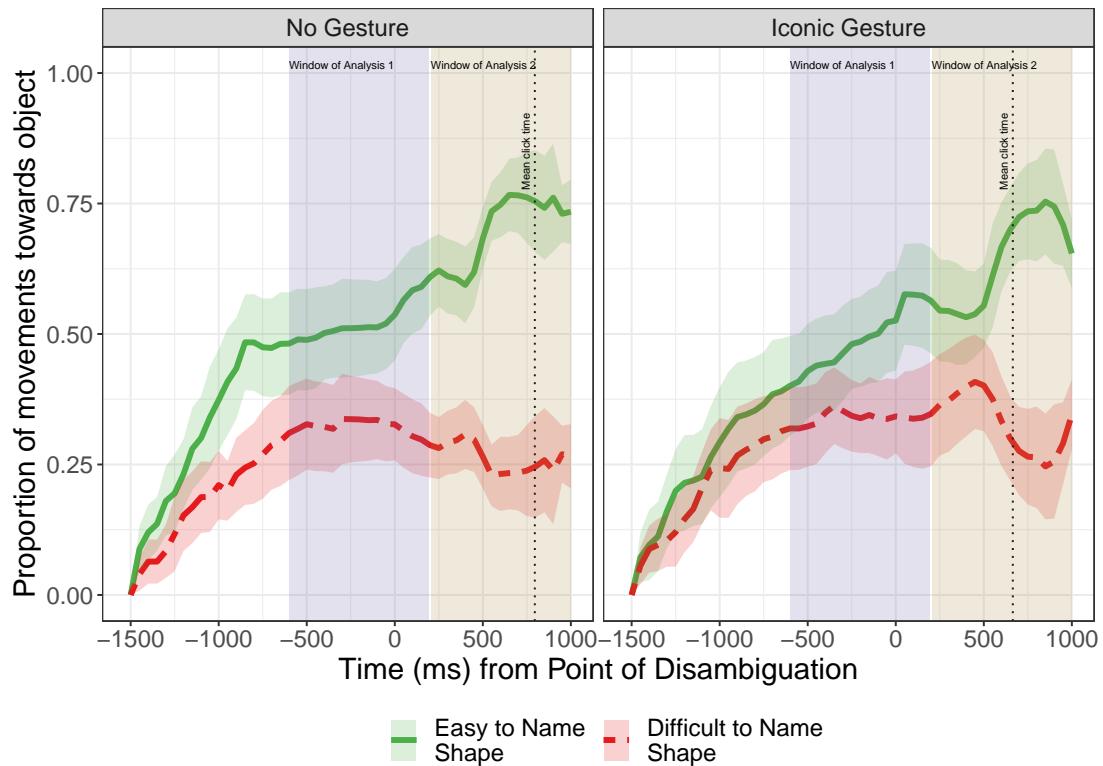


Figure 5.4: Mouse-tracking results for critical trials that name an easy-to-name shape in Experiment 5.1: Proportion of cumulative distance travelled toward each object (easy-to-name shape and difficult-to-name shape) from speech onset to 1000 ms post-disambiguation, split by presence of gesturing. Proportions were calculated from the total cumulative distance participants moved the mouse until that time bin. Shaded areas represent 95% confidence intervals derived via bootstrapping subject data ($R=1000$).

5.2.4 Post-disambiguation

Eye movements

Analysis over the window from 200 ms after point-of-disambiguation and extending for 800 ms (up until just after the average time of 920 ms to click on either object) revealed an overall tendency to fixate the referent over the competitor across this

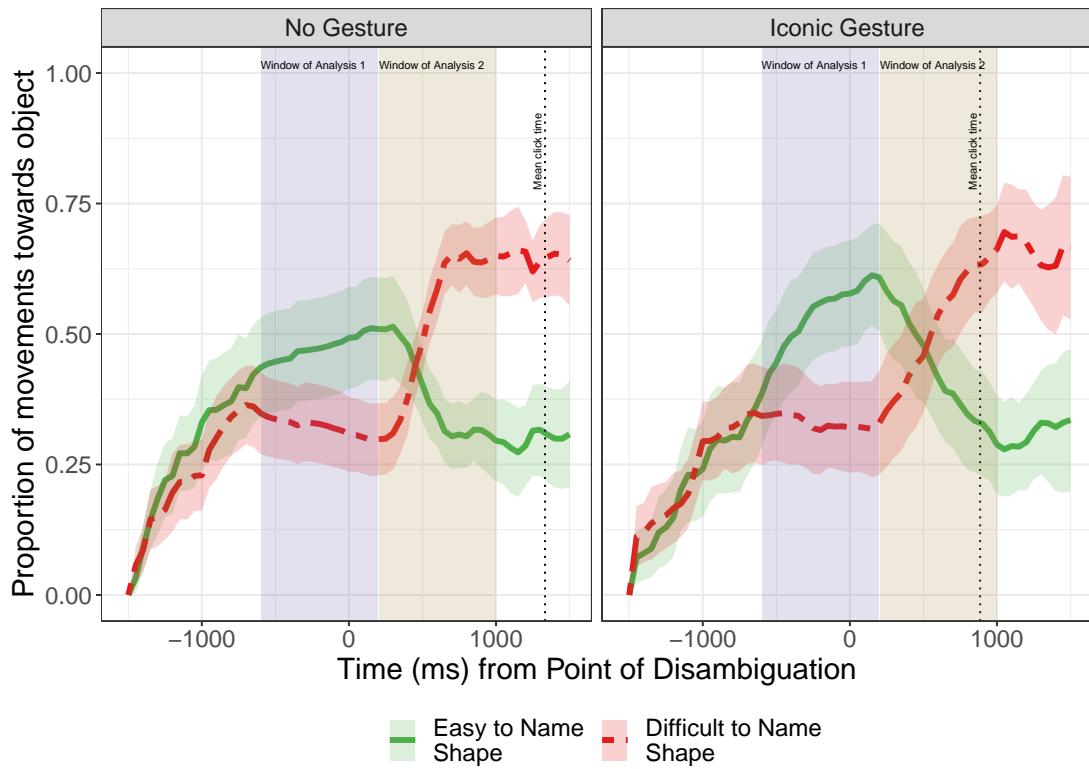


Figure 5.5: Mouse-tracking results for critical trials that name a difficult-to-name shape in Experiment 5.1: Proportion of cumulative distance travelled toward each object (easy-to-name shape and difficult-to-name shape) from speech onset to 1500 ms post-disambiguation, split by presence of gesturing. Proportions were calculated from the total cumulative distance participants moved the mouse until that time bin. Shaded areas represent 95% confidence intervals derived via bootstrapping subject data ($R=1000$).

window (a significant intercept term: $\beta = 2.34$, $SE = 0.19$, $t = 12.10$), which increased as this window progressed (linear effect of time: $\beta = 10.33$, $SE = 0.98$, $t = 10.57$). The increasing tendency to fixate the referent over the competitor was reduced when target shapes were more difficult-to-name ($\beta = -4.51$, $SE = 0.41$, $t = -10.88$), but increased more quickly (as indicated by an interaction of referent nameability and the quadratic term $\beta = -4.50$, $SE = 0.39$, $t = -11.44$).

A significant interaction between presence of gesture and quadratic effect of time indicated that when videos showed the speaker gesturing, participants' increasing tendency to fixate the referent over the competitor emerged more gradually ($\beta = 2.44$, SE = 0.39, $t = 6.19$). This effect was greater for gestures of difficult-to-name shapes (interaction between the quadratic term for time, gesture and referent nameability: $\beta = 2.40$, SE = 0.78, $t = 3.06$).

Gestures of difficult-to-name shapes also resulted in a weakening of the overall fixation bias to the referent across the window ($\beta = -0.25$, SE = 0.12, $t = -2.04$). Model results are shown in Table 5.2, and Figure 5.6 shows the empirical logit transformed bias towards the referred to object over the competitor during the relevant window of analysis, along with the fitted values from the model.

Mouse movements

In the 800 ms window beginning 200 ms after point-of-disambiguation, participants' mouse movements largely patterned with their eye movements. Across the window, participants showed a bias to have moved the mouse more towards the referent than the competitor ($\beta = 0.75$, SE = 0.15, $t = 5.08$), and this increased over the 800 ms window ($\beta = 2.08$, SE = 0.36, $t = 5.86$). Although the overall referent bias was reduced in trials referring to difficult-to-name shapes ($\beta = -0.87$, SE = 0.27, $t = -3.20$), the increase in the referent bias over time was greater ($\beta = 1.24$, SE = 0.27, $t = 4.69$), likely reflecting that at the start of this window participants tended to have already moved the mouse towards the easier-to-name object.

The presence of gesture influenced the linear increase in the referent-bias over time ($\beta = 0.68$, SE = 0.27, $t = 2.56$, but this emerged more gradually, as indicated by the interaction with gesture and the quadratic term of time ($\beta = 0.69$, SE = 0.25, $t = 2.80$)). A significant interaction between gesture and nameability ($\beta = 0.27$,

Table 5.2: Model results for eye- and mouse-tracking analyses in Experiment 5.1 over the period from 200 to 1000 ms after point-of-disambiguation

	Fixations			Mouse Movements		
	β	SE	t	β	SE	t
(Intercept)	2.34	(0.19)	12.10	0.75	(0.15)	5.08
Gesture	-0.30	(0.30)	-1.00	-0.28	(0.25)	-1.14
Difficulty to Name	-0.25	(0.40)	-0.62	-0.87	(0.27)	-3.20
Time	10.33	(0.98)	10.57	2.08	(0.36)	5.86
Time ²	-1.99	(0.72)	-2.76	-0.19	(0.16)	-1.13
Time ³	-2.81	(0.19)	-15.08	-0.58	(0.11)	-5.06
Gesture × Difficulty to Name	-0.25	(0.12)	-2.04	0.27	(0.12)	2.15
Gesture × Time	-0.71	(0.41)	-1.72	0.68	(0.27)	2.56
Gesture × Time ²	2.44	(0.39)	6.19	0.69	(0.25)	2.80
Gesture × Time ³	-0.48	(0.37)	-1.28	0.07	(0.23)	0.33
Difficulty to Name × Time	-4.51	(0.41)	-10.88	1.24	(0.27)	4.69
Difficulty to Name × Time ²	-4.50	(0.39)	-11.44	-0.33	(0.25)	-1.35
Difficulty to Name × Time ³	3.85	(0.37)	10.39	0.28	(0.23)	1.24
Gesture × Difficulty to Name × Time	0.74	(0.82)	0.90	-0.57	(0.53)	-1.09
Gesture × Difficulty to Name × Time ²	2.40	(0.78)	3.06	-0.31	(0.49)	-0.63
Gesture × Difficulty to Name × Time ³	-2.35	(0.74)	-3.18	0.43	(0.45)	0.94
Var(residual)	11.55			3.68		
Var(1—Participant)	0.34			0.26		
Var(Gesture—Participant)	1.12			0.86		
Var(Difficulty to Name—Participant)	1.35			1.17		
Var(Time—Participant)	9.92			1.39		
Var(Time ² —Participant)	4.74			0.16		
Var(1—Item)	0.45			0.20		
Var(Gesture—Item)	0.87			0.45		
Var(Difficulty to Name—Item)	2.00			0.49		
Var(Time—Item)	10.26			1.05		
Var(Time ² —Item)	5.84			0.11		
Total	16062			5669		
Participant	25			25		
Item	20			20		

$SE = 0.12, t = 2.15$) indicated that when trials presented the speaker referring to a difficult-to-name object, the videos with gesturing resulted in an increased

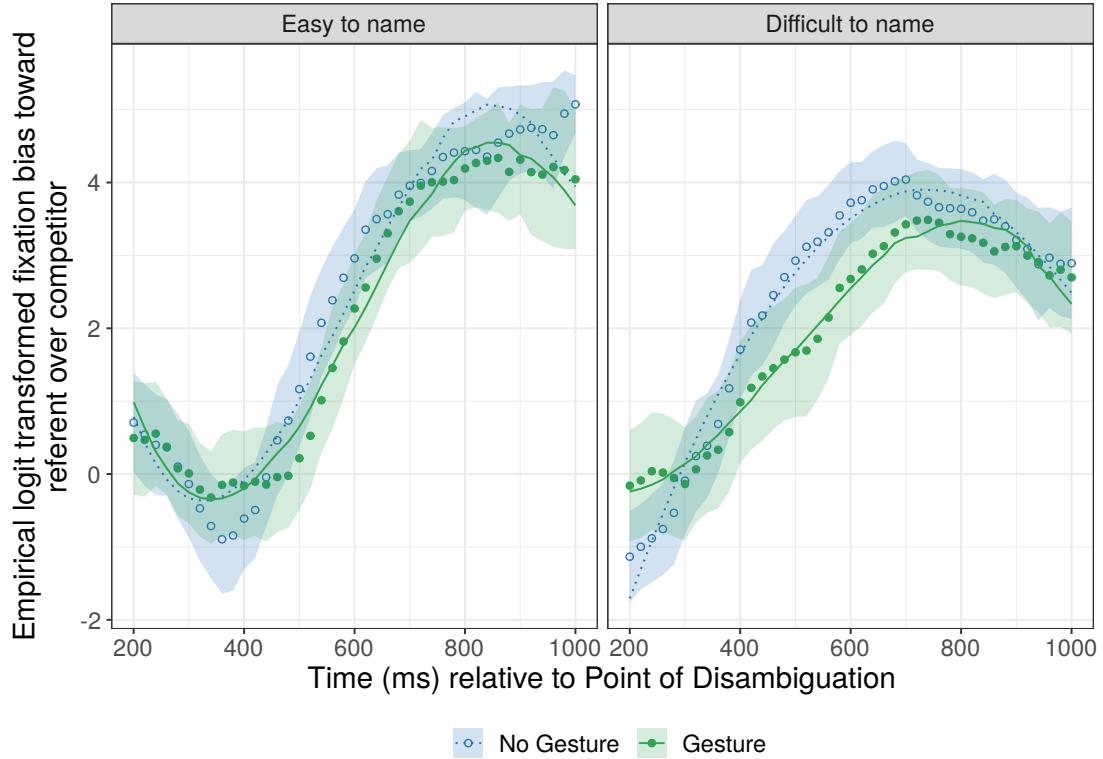


Figure 5.6: Empirical logit transformed fixation bias towards referred to shape over competitor shapes in critical trials in Experiment 5.1 from 200 ms to 1000 ms after point-of-disambiguation, by presence of gesture and split referent nameability. Lines represent fitted values of the model.

overall mouse movement bias to the referent across this window compared to videos without gesturing (whereas the opposite effect was found in the bias to fixate the referent). Full results of the mouse-tracking analysis are shown in Table 5.2, and Figure 5.7 shows the empirical logit transformed movement bias towards the referred to object over the competitor during the relevant window of analysis, along with the fitted values from the model.

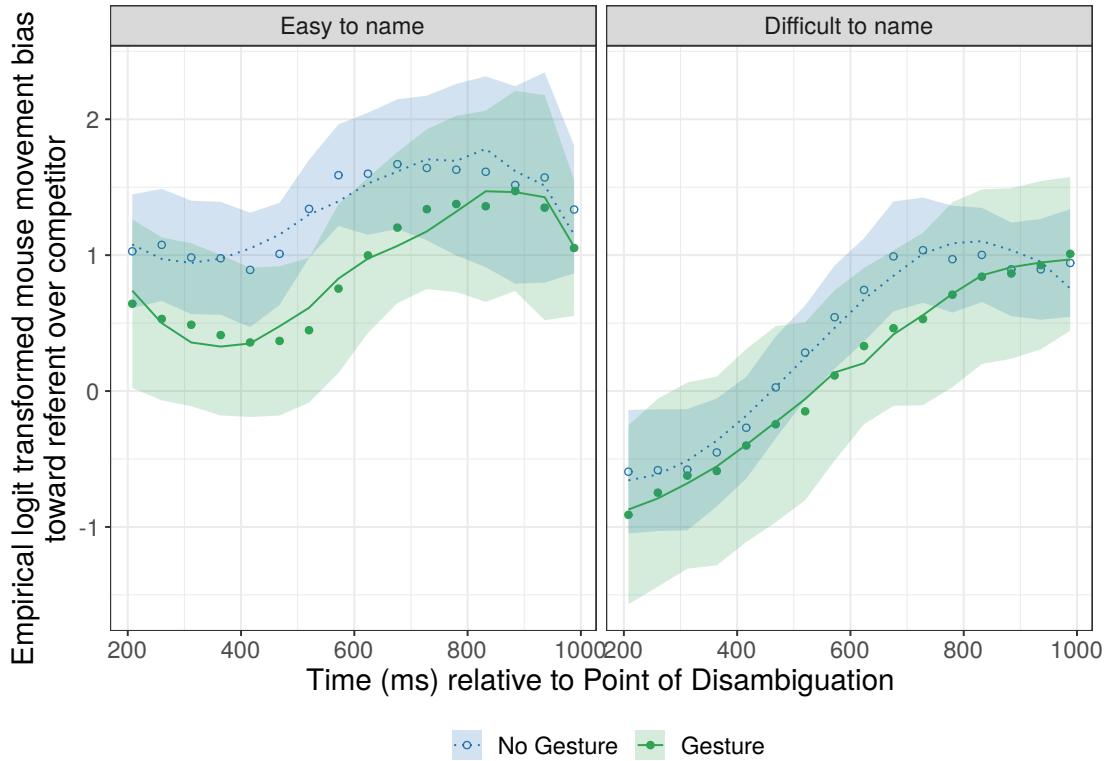


Figure 5.7: Empirical logit transformed mouse movement bias towards referred to shape over competitor shapes in critical trials in Experiment 4.1 from 200 ms to 1000 ms after point-of-disambiguation, by presence of gesture and split by referent nameability. Lines represent fitted values of the model.

Object clicks and response times

Table 5.3 shows the error rate (proportion of trials in which participants clicked on the competitor) by the nameability of the referent (Easy-to-name vs. Difficult-to-name) and the presence of gesture. Analysis of incorrect mouse clicks revealed an overall tendency to click on the correct shape ($\beta = -3.24$, SE = 0.43, $p < .001$). A greater proportion of clicks to the competitor (i.e., incorrect interpretations of which shape the speaker was referring to) was associated with both trials in which the speaker gestured ($\beta = 2.19$, SE = 0.77, $p = .004$), and trials in which they referred to the more difficult shape ($\beta = 1.02$, SE = 0.50, $p = .04$).

Participants took more time to click on a shape when the speaker referred to a difficult-to-name shape in comparison to when they referred to an easy-to-name shape ($\beta = 0.15$, SE = 0.02, $t = 9.52$). The presence of iconic gesturing was associated with faster response times ($\beta = -0.09$, SE = 0.02, $t = -4.47$), and there was a significant interaction of gesturing and codability of the referent, with a greater reduction in response times following gesturing for difficult-to-name shapes than easy-to-name ones ($\beta = -0.07$, SE = 0.03, $t = -2.36$). Tables 5.4 and 5.5 show the full results of models of incorrect mouse clicks and time-to-click respectively.

Table 5.3: Breakdown of incorrect mouse clicks (clicks to the shape not described by the speaker) in critical trials in Experiment 5.1 by presence of gesture and nameability of shape

	Easy Shape	Difficult Shape	
		No Gesture	Gesture
Easy Shape	0 (0%)		11 (8.8%)
Difficult Shape	7 (5.6%)		23 (18.4%)

Table 5.4: Model results for mouse clicks to competitor over referent in Experiment 5.1

	β	SE	p
(Intercept)	-3.24	(0.43)	<.001
Gesture	2.19	(0.77)	.004
Difficulty to Name	1.02	(0.50)	.04
Var(1—Participant)	0.18		
Var(Gesture—Participant)	1.14		
Var(Difficulty to Name—Participant)	1.27		
Var(1—Item)	0.09		
Total	500		
Participant	25		
Item	20		

Table 5.5: Model results for times taken to click the mouse in Experiment 5.1

	β	SE	t
(Intercept)	7.79	(0.01)	571.01
Gesture	-0.09	(0.02)	-4.47
Difficulty to Name	0.15	(0.02)	9.52
Gesture \times Difficulty to Name	-0.07	(0.03)	-2.36
Var(1—Participant)	0.001		
Var(Gesture—Participant)	0.004		
Var(1—Item)	0.002		
Total	459		
Participant	25		
Item	20		

5.3 Discussion

Following from Experiment 4.1, in which a gated-task established that the presence of iconic gesturing influenced listeners’ guesses that upcoming content would be more difficult-to-name, Experiment 5.1 attempted to investigate how this type of gesturing might influence listeners’ anticipations alongside the unfolding speech and gesture streams.

Experiment 5.1 presented participants with full utterances and iconic gestures, and was framed as a reaction time game in which participants were tasked with responding quickly according to the shape being described in speech and gesture (when present). Within a visual display comprising an easy-to-name shape, a difficult-to-name shape, and a video of a speaker, we investigated how participants’ preference for either object (as measured by eye and mouse movements) varied depending upon whether the speaker was seen to produce iconic gesturing along with speech or whether they did not gesture. Audio and video were carefully constructed such that both speech and gesture were temporarily ambiguous between the two objects in the display. Therefore, any bias which participant displayed towards one object over the other during the pre-disambiguation period could be

attributed to the presence of gesturing (i.e., the *fact that the speaker gestured*) rather than a specific gesture’s shape or trajectory.

5.3.1 Pre-disambiguation

Results revealed that in the pre-disambiguation window (from 600 ms before point-of-disambiguation to 200 ms after), participants tended to fixate and move the mouse towards the more easy-to-name of two objects as the window progressed. Contrary to predictions, participants’ eye and mouse movements were not influenced by whether or not speech was presented with gesture. Results showed no evidence that the presence of iconic gesturing guided predictions of upcoming content in real-time. This contrasts with Experiment 4.1 which suggested that gesturing did guide participants explicit predictions (but relatively late on).

Interestingly, participants displayed biases to fixate on (and move the mouse towards) the easy-to-name shape in this window—directly contrasting with Experiment 4.1 in which visual inspection of the equivalent time window suggests a possible preference for the difficult-to-name shapes instead (see Figure 4.4, Chapter 4). In the previous chapter, we suggested that this may be due to the relative visual saliency of difficult-to-name shapes. The differences between experiments may in part be explained by the fact that the present experiment increased the duration for which shapes were seen prior to the audio and video beginning by 1000 ms from Experiment 4.1, and so any extra looking at difficult-to-name objects had already been conducted.

A further explanation of this discrepancy may be found in participants’ responses to the post-test questionnaire, in which nine participants (36%) reported following a specific strategy in order to complete the task—holding the cursor over an easy-to-name shape and waiting for the onset of speech to disambiguate between

the two. This explanation may also account for how participants behaved after the point-of-disambiguation depending upon whether the speaker described an easy-to-name or difficult-to-name shape. For instance, the tendency to move the mouse towards the referent (shape described by the speaker) in the period from 200 to 800 ms after disambiguation was greater for trials in which the speaker described a difficult-to-name shape—likely because at this point many participants had already moved the cursor toward the easy-to-name shape.

5.3.2 Post-disambiguation

Eye- and mouse-tracking results during the post-disambiguation period suggest that gesture delayed participants from fixating on and moving the mouse towards the referent over the competitor. This is likely indicative of participants fixating more on videos when there is gesturing present, and this is supported by visual inspection of the time-course of fixations to the videos (see Figures 5.2 and 5.2). The emergence of the fixation bias to the referent was slower for gestures which represented difficult-to-name shapes (than those representing easy-to-name ones), perhaps because these gesture were less familiar, and so more visually salient.

This explanation is not consistent, however, with the analysis of the time taken to click the mouse. As predicted, participants were quicker to click the mouse following descriptions which were presented in both speech and gesture compared to speech alone, in keeping with the findings from Experiment 4.1. Interestingly, this effect was greater in descriptions of difficult-to-name shapes. Taken together with the increased rate of errors following gestures, it is possible that this reflects an association between the presence of gesturing and difficult-to-name shapes, rather than a facilitatory effect of the content (post-disambiguation) of gestures on reference comprehension. The fact that gestures of easy-to-name shapes also improved reaction times (relative to speech without gestures), however, is

inconsistent with this view, pointing instead towards listeners responding more quickly (but not necessarily more accurately) to speech with gestures than speech without (see Holler et al., 2017, for a corpus study revealing that questions with gestures tend to receive quicker responses). One possible explanation of why the present study found this effect to be greater for difficult-to-name shapes is that verbal descriptions of easy-to-name shapes tended to be quicker to disambiguate (e.g., “number 6” or “letter K”, compared to “curvy line and two lines below”), rendering gestures of these shapes quickly redundant for comprehension.

5.3.3 Eye-tracking and gestures

The present study highlights some of the difficulties involved in research into multi-modal comprehension. One specific issue is that the effects under investigation here are fleeting: Speech with gesturing means ‘the same thing’ as speech without, rather than leading to different lasting interpretations. This is also true of comparable studies in speech disfluency, but causes less of a problem as there is only one input stream (speech). In the present study, gesture is on-going throughout the window in which these fleeting effects may be measured, it is difficult to distinguish effects on comprehension from those on visual attention (see also Huettig et al., 2011).

This is made even more difficult due by the fact that in language production, gestures tend to be tightly temporally coupled with speech (see Bergmann et al., 2011; McNeill, 1992, 2005). The stimuli used in the present study were constructed such that many gestures began over 1000 ms before referent-onset in speech, which has been suggested as a rough maximum for which a gesture tends to precede the onset of its lexical affiliate (see Butterworth & Beattie, 1978; De Ruiter, n.d.; Morrel-Samuels & Krauss, 1992). By contrast, it may be less important exactly when a disfluency occurs: For example, studies have found effects of disfluency on judgements of deception to be similar for disfluency occurring in utterance-initial

and utterance-medial positions (see Loy et al., 2017). Construction of speech-gesture stimuli therefore is likely to face a significant trade-off between realistic credibility (in terms of synchrony) and control (over the relative timings) required to disentangle the effects on comprehension from each modality.

5.4 Does gesturing guide listeners' predictions of upcoming message content?

Research suggests that when a gesture conveys something more than what is conveyed in speech, listeners' expectations are influenced by the meaning conveyed in gesture. This is evidenced by studies like Holle and Gunter (2007), in which disambiguating information presented in gesture was found to influence ERP responses at a target word in a subsequent clause. Chapters 4 and 5 have investigated whether listeners' predictions are influenced by presence (rather than content) of gesturing as a signal that the speaker is having difficulty planning speech. This idea is predicated on research which suggests that speakers produce more (and longer) iconic gestures when producing verbal descriptions is more conceptually demanding (i.e., Experiment 3.1, Hostetter et al. 2007b; Morsella and Krauss 2004). It also draws parallels with research showing that listeners' anticipations of referents are sensitive to signals in the speech channel such as disfluencies (Arnold et al., 2007; Barr, 2001).

Experiment 4.1 suggests that listeners may rely on the presence of iconic gesturing to inform their explicit predictions about what a fragment of speech and gesture was describing. However, we found little evidence to suggest that gesture informs listeners' anticipations of upcoming message content alongside the unfolding perceptual input. Although we have provided possible explanations for this

1605.4 Does gesturing guide listeners' predictions of upcoming message content?

discrepancy (the lack of time pressure in Experiment ??; participants' task-strategy in Experiment 5.1), there are two further explanations that warrant discussion.

The first point concerns the shapes and gestures used in the studies presented here. To investigate listeners' perception of the presence and duration of gesturing as signals about a speaker's message required us to ensure that the content of gesturing did not refer to one shape more than the other (else we would be unable to discern between whether participants were responding to gesturing as a signal or as a mode of communication similar to spoken language). To do this, we carefully constructed shapes and gestures such that the gestures were temporarily ambiguous between shapes—e.g., with the two shapes sharing a section, the initial fragment of gesture would represent both shapes equally. However, this means that we were presenting participants with gestural information which is likely to increase fixations to both shapes, much like the phonological onset of 'camel' increases fixations on images of both a camel and a candle (Allopenna et al., 1998). This also meant that while the shape and trajectory of gestures could be well controlled, this was at the expense of how naturalistic the gestures were—the careful finger tracing of shapes used in these comprehension experiments (4.1 and 5.1) are at odds with the imprecise gestures elicited in the production study (Experiment 3.1). One avenue for future research would be to explore how listeners' predictions of upcoming content are influenced by complex, squiggly iconic gestures which do not resemble either shape. Stimuli construction could be more flexible, but it is feasible that these gestures could be perceived as intentional signals from the speaker that they are attempting to visually represent a difficult-to-name shape, or as unintentional signals of the efforts of planning speech.

A second explanation of the discrepancy between experiments 4.1 (truncated) and 5.1 (time-pressured) is that for listeners to draw on gestures to inform fleeting predictions such as that of an upcoming referent may be needlessly demanding in situations where predictions are soon made irrelevant by the presence of

disambiguating information. In other words, the information (in both speech and gesture) prior to disambiguation is ultimately redundant in listeners' interpretation of the message. This does not refute the value of pre-disambiguation to participants in Experiment 5.1 in terms of the benefits for response times. However, it may indicate that the uptake (or use) of non-verbal information may be optional and situation specific.

The experiments presented in Part I of this thesis have focussed on how gesturing may be perceived as a signal about the literal meaning of a speaker's message. Non-verbal behaviours can also signal information about a speaker's intentions and goals, however. In Part II, we turn to how speakers' non-verbal behaviours may influence *pragmatic* comprehension—i.e., a listener's interpretation of what a speaker intends—rather than of the literal meaning of the words and gestures. Drawing on a parallel literature in speech disfluency, we investigate how manner of non-verbal delivery can ultimately have lasting repercussions for a listeners' final interpretation of an utterance.

1625.4 Does gesturing guide listeners' predictions of upcoming message content?

Part II

Markers of deception

Chapter 6

Gesturing informs pragmatic judgements: Interpreting non-verbal cues to deception in real time¹

In Part I we investigated the extent to which a speaker's gestures function as signals of difficulty in the planning and production of speech. In Chapter 3, we established that the production of iconic—or representational—gesturing relative to speech varies according to the conceptual demands required to describe an object. The contra-position of this finding was subsequently investigated in Chapter 4, in which we asked whether listeners interpret the occurrence of gesturing as an indication of upcoming semantic content. Results from Experiment 4.1 suggested that listeners associate the presence of iconic gesturing (in which the gestural content remains ambiguous) with less easily named shapes. This finding patterns

¹A revised version of this chapter is currently in press. Model results are detailed in Appendix A

with previous research suggesting that listeners are sensitive to other non-linguistic cues (specifically speech disfluency) in anticipating upcoming referents (see e.g., Arnold et al., 2007).

The experimental paradigm used in Chapter 4 presented participants with truncated audiovisual utterances and tasked them with guessing the object about to be described. This was subsequently developed in Chapter 5 to include the full spoken (and gestured) descriptions of shapes to examine whether gesture-based predictions of upcoming content are formed alongside the moment-to-moment processing of speech/gesture streams. Results were inconclusive as to whether an association between iconic gesturing and less nameable objects is borne out in listeners' on-line expectations of upcoming semantic content. We suggested a number of possible explanations for our findings, including the strategy employed by participants; the broader difficulties in disentangling the effects of gesturing on comprehension from those on visual attention; and the fact that listeners' on-the-fly predictions have comparatively minor consequences in terms of message comprehension due to subsequent disambiguating information (in contrast to the truncated paradigm). What remains an open question is whether and how different non-linguistic cues can have an impact on listeners' global interpretations of the speaker's message, beyond their effect on any short-lived predictions about upcoming semantic content.

Part II addresses this question by studying the influence of non-linguistic behaviours on listeners' pragmatic comprehension. Specifically, in the present chapter, we focus on if and when the presence of different types of non-verbal behaviour modulates listeners' judgements of message truth (i.e., whether or not a speaker is being deceptive). Subsequently, Chapter 7 will examine the differences between potential signals of deception in different modalities, and Chapter 8 uses manner of spoken delivery to investigate how listeners' perceptions of speech disfluency as markers of deception are modulated by the availability of competing explanations

of its cause. First, we provide an introduction to the field of deception research, before discussing in more depth the distinction between literal (or semantic) and pragmatic comprehension.

6.1 Deception and delivery

Although people lie frequently in everyday discourse (see DePaulo & Kashy, 1998; DePaulo, Kashy, Kirkendol, Wyer, & Epstein, 1996), research suggests that we are notably deficient in detecting deceit. A 2006 meta-analysis of over 24,000 participants across 206 studies revealed a 54% accuracy rate in distinguishing truths from lies—only just above chance. All utterances, both lies and truths, are accompanied by various behavioural cues which may influence how the utterance is interpreted. In deception, some of these behaviours may be the result of the cognitive processes underlying the act of constructing and maintaining a falsehood, as unintended displays (or *leakage*) of emotional/affective and cognitive states. Possible causes of behavioural variation between lies and truths are varied: One explanation (Ekman, 1992) is that the emotions which are experienced when telling a lie (such as fear, shame or excitement) result in cues to deception that are indicative of these emotional states (such as gaze aversion or a closed posture, DePaulo et al., 2003). Alternatively, cues may reflect the cognitive demands of formulating a lie (see Vrij, Kneller, & Mann, 2000), resulting in, for instance, increases in speech disfluencies (Zuckerman, DePaulo, & Rosenthal, 1981). Lastly, liars may attempt to control their behaviour when lying (see Buller & Burgoon, 2006), suppressing commonly assumed behavioural correlates of deception, and leading to fewer representational gestures (D. Cohen, Beattie, & Shovelton, 2010) and greater overall rigidity (Vrij, 1995).

Although prior research has identified many potential cues to deceit, there is often

disagreement with regard to both the reliability as well as sometimes the direction of these associations (for a comprehensive meta-analysis, see DePaulo et al., 2003). Disagreement may in part reflect, for example, differences between the types of lie being studied (exaggerations, outright falsehoods, omissions etc., see DePaulo et al., 1996), as well as differences in participant populations and the variable cues those liars tend to display (Hart, Fillmore, & Griffith, 2009).

In detecting deception, there appears to be a disparity between those cues which are perceived to be indicative of deceit and those cues which actually are (see DePaulo, Rosenthal, Rosenkrantz, & Green, 1982; Zuckerman, Koestner, & Driver, 1981). It may be that listeners are relying on the wrong cues when forming judgements of deception, or simply that associations between behavioural cues and lying are weak (Hartwig & Bond, 2011). Regardless, listeners appear to form these associations independent of cue validity: For example, a meta-analysis of 33 studies, Zuckerman, DePaulo, and Rosenthal (1981) found a wide range of behaviours which people believe to be associated with deception, from response length and latency to eye-gaze, postural shifts and self-adaptive movements (with twice as many cues *believed* to indicate dishonesty as were *actually* associated with lying). Beliefs about deceptive behaviour have even, in some cases, been shown to be so pervasive as to influence peoples views of their own behaviour when lying: In a study in which participants were asked to produce two interviews (for one of which they lied), participants' subsequent judgements of their own behaviour showed that they believed themselves to have moved more when lying, despite a decrease in movements actually occurring (Vrij et al., 1996). Furthermore this did not change when participants were informed prior to the interviews that lying is usually associated with a decrease in movements.

Meta-analyses such as Zuckerman, DePaulo, and Rosenthal (1981) and Hartwig and Bond (2011) have shown that across studies there are certain cues which listeners reliably associate with deception. Along with a number of cues in the speech

stream in both content and delivery (such as filled pauses), several non-verbal behaviours have emerged as reliable correlates with perceived dishonesty—namely postural shifts, and increased arm, foot and leg movements.² Listeners' uptake of this information to form judgements of message truth is a form of pragmatic comprehension, in that it involves drawing on contextual information in order to gain an interpretation of a speaker's intentions or goals.

6.2 Literal vs. pragmatic comprehension

Theories of cooperative communication emphasise a distinction between the conventional (or literal) meaning of words uttered and the meaning of these words in the wider context (Grice, 1975). Understanding this latter form of meaning—termed *pragmatic*—involves the listener drawing on information from any number of contextual sources to infer what the speaker intends by their utterance. Pragmatically relevant information can be anything from knowledge of the speaker (their goals, background and identity), the prior discourse, or the surrounding environment (e.g., Hagoort et al., 2004; Van Berkum, van den Brink, Tesink, Kos, & Hagoort, 2008) to the speaker's intonation (Brennan & Williams, 1995) or co-speech movements (Kelly, 2001; Kelly et al., 1999).

Under a pragmatic view, pragmatic meaning is derived by taking the semantic content of an utterance and applying a potentially complex inferential process. This *standard pragmatic model* (see Grice, 1975; Searle, 1969) has resulted in an assumption that comprehension of pragmatic meaning emerges at a later stage of language processing, only after a literal, context-independent meaning has been computed. A growing body of research, however, suggests that this may not be the

²We note there is some discrepancy as to how these cues are defined, leaving potential overlaps between, for instance, any of 'arm movements', 'hand movements', 'adaptors' and 'fidgeting'.

case, with studies investigating the relative time course of literal and pragmatic processing indicating that pragmatic interpretations can—given enough contextual support—be derived as quickly as literal meaning (for an overview, see Gibbs Jr & Colston, 2012).

A number of studies have explored how information in the wider context informs understanding by studying the time-course of pragmatic comprehension directly. For instance, ERP research by Nieuwland and Van Berkum (2006) found that an N400 effect—typically associated with various forms of unexpected input (Ganis, Kutas, & Sereno, 1996; Kutas, Neville, & Holcomb, 1987), and localised to within 200–600 ms of stimulus presentation Kutas and Federmeier 2011—associated with verb-object animacy violations (e.g., “The girl comforted the clock”) disappeared when the sentence was embedded within an appropriate context (a cartoon of a girl speaking to a clock about his depression). Similarly, N400 effects have been shown to be influenced by information about a speaker’s identity inferred from their voice: Van Berkum et al. (2008) found that N400 activity following a given utterance was modulated by whether it matched or mismatched what was likely given the age and gender of the speaker’s voice (e.g., “If only I looked like Britney Spears” in a male voice). Such research shows that the processes involved in comprehension are able to make rapid use of information beyond the literal meaning of an utterance, suggesting that local semantics and global context are immediately integrated to influence interpretation of a speaker’s message.

Research into the time course of comprehension of gestures has been focused on the relative distribution of semantic content between modalities, and how this influences comprehension of the literal message. For instance, Kelly et al. (2004) measured participants’ ERP responses when faced with gestures which were either semantically congruent or incongruent with subsequent spoken adjectives. Relative to trials in which gesture and speech conveyed the same meaning, Kelly et al. found an N400 effect when gestures strongly mismatched the speech, reflecting

semantic integration of the two modalities at an early stage of comprehension. By contrast, the roles that non-verbal behaviours play in pragmatic comprehension have garnered little attention. Studies have shown that the accompaniment of speech by a relevant pointing gesture results in a greater likelihood of both adults and children to interpret the utterance as an indirect request (Kelly, 2001; Kelly et al., 1999), but have been limited to post-hoc measures of comprehension. For example, after viewing an exchange between two characters about sandwiches which ends with the utterance “Actually, I’m still pretty hungry” either with or without a pointing gesture to the other character’s sandwich, Kelly et al. (1999) questioned participants about how they thought characters in the videotape would react.

In the field of deception research, associations between speakers’ goals (e.g., to deceive) and their non-linguistic behaviours have been more widely studied. The majority of research into perception of deception has tended to rely on after-the-fact judgements, or by assessing listeners’ explicit beliefs about cue validity (see Vrij & Semin, 1996; Zuckerman, Koestner, & Driver, 1981), and it is only recently that research has begun to investigate the processes by which cues are incorporated into listeners’ judgements of deception. Work from Loy et al. (2017) has shown that the association between speech disfluency (in both utterance-initial and utterance-medial positions) and deception is evident during the early moments of comprehension. Framed as a lie-detection game, participants in Loy et al.’s (2017) study made implicit judgements of message truth for utterances describing the location of some treasure (“The treasure is behind the <referent>”). Eye and mouse-tracking measures indicated that the presence of disfluency in speech influenced listeners’ judgements early on, alongside the unfolding linguistic input.

6.2.1 Negation

Judging messages to be dishonest such as those made by participants in Loy et al.'s (2017) study requires negating the content of the speaker's message. Research has indicated that processing negations tends to be more cognitively demanding, indicated by longer reaction times and higher error rates when verifying negative sentences against pictures compared to verifying positive ones (e.g., Carpenter & Just, 1975). One explanation for this is that negations are harder to process because listeners must first invoke a representation of the positive argument before subsequently negating it (Carpenter & Just, 1975; Kaup, Ludtke, & Zwaan, 2007). The speed with which listeners can draw on contextual cues such as fluency of speech to inform judgements of message truth (as in Loy et al., 2017) appears at odds with this account.

An alternative dynamic pragmatic account suggests that the initial representation of the positive is not mandatory (see Tian & Breheny, 2016), for instance in situations where cues in the sentence project a negative *Question Under Discussion* (*QUD*). *QUD* can be thought of as the contextual relevance of an utterance (see Ginzburg, 2012; Roberts, 2012), or what question the utterance is addressing. Tian and Breheny (2016) found that negative sentences with structures which project a negative *QUD* (e.g, *who didn't iron their shirt?*) as opposed to a positive one (*did John iron his shirt?*) reversed patterns of response times, suggesting that the positive counterpart need not always be represented prior to its negation. It is possible that non-linguistic cues, like linguistic ones, may have a similar effect, triggering the *QUD* of *the treasure isn't where?* as opposed to *is the treasure behind X?*. In the context of deception detection studies such as Loy et al. (2017), this may explain the speed with which effects are seen to emerge.

6.3 Experiments 6.1 and 6.2

The present chapter extends Loy et al.'s (2017) study to the visual domain of non-verbal delivery: Two experiments investigate the influence of a speaker's non-verbal behaviours on judgements of deception, asking if and when associations between these cues and deception emerge.

6.3.1 Abstract

When determining the veracity of an utterance, we perceive certain non-linguistic behaviours to indicate that a speaker is being deceptive. Recent work has highlighted that listeners' associations between speech disfluency and dishonesty are detectable at the earliest stages of reference comprehension, suggesting that the manner of spoken delivery influences pragmatic judgements concurrently with the processing of lexical information. Here, we investigate the influence of a speaker's gestures on judgements of deception, and ask if and when associations between non-verbal cues and deception emerge. Participants saw and heard a video of a potentially dishonest speaker describe treasure hidden behind an object, while also viewing images of both the named object and a distractor object. Their task was to click on the object behind which they believed the treasure to actually be hidden. Eye and mouse movements were recorded. Experiment 6.1 investigated listeners' associations between visual cues and deception, using a variety of static and dynamic cues. Experiment 6.2 focused on *adaptor* gestures (touching behaviours and movements directed towards the self, objects or others). We show that a speaker's non-verbal behaviour can have a rapid and direct influence on listeners' pragmatic judgements, supporting the idea that communication is fundamentally multi-modal.

6.3.2 Introduction

In natural communication, speakers can convey information via multiple channels. Along with spoken delivery, a speaker's gestures, postures and facial expressions can all offer extra-linguistic information about the speaker or message. Listeners can be affected by such information in a number of ways. They may, for example, make inferences about the speaker's emotion (Busso et al., 2004; Gregersen, 2005). Alternatively, their interpretation of the message itself may change, for example if extra-linguistic information causes them to believe that the speaker may be being dishonest (Zuckerman, DePaulo, & Rosenthal, 1981). The present paper focuses on this latter circumstance. In particular, we investigate whether, and how, speakers' postures or adaptor gestures affect listeners' judgements of veracity.

This is especially relevant in light of recent work investigating the manner in which utterances are spoken. Work focusing on the auditory modality has established an association between spoken disfluency and deceit that emerges from the early stages of comprehension. Loy et al. (2017) used a visual world eye- and mouse-tracking paradigm in which participants were presented with images of two objects, and heard utterances describing the location of some treasure purportedly hidden behind one of the objects. These utterances were presented as having been elicited in a previous experiment, in which the speaker was said to have been lying some of the time. Crucially, Loy et al. (2017) manipulated the manner of spoken delivery, with half of the experimental items containing a speech disfluency. Participants were tasked with clicking on the object they *believed* to be concealing the treasure, choosing either the object named in the utterance (indicating a judgement of honesty), or a distractor (dishonesty). They were more likely to judge disfluent utterances as dishonest than fluent ones (as indicated by a greater probability of clicking on the distractor in a disfluent trial). Importantly, disfluency resulted in an early bias in both eye and mouse movements towards the not-referred-to

object. This suggests that speech disfluency is incorporated into listeners' ideas concerning deceptive speech, and has an immediate effect on their interpretation of an utterance.

Turning from the auditory to the visual modality, research suggests that many non-verbal aspects of delivery are associated by listeners with deception. In an analysis of 33 studies, Zuckerman, DePaulo, and Rosenthal (1981) found that nine out of the ten visual cues-to-deception that were investigated were believed to be indicative of deceit. In 13 studies reporting relationships between cues and subsequent deception judgements (rather than explicit beliefs about cues), three (smiling, gaze, and postural shifts) of the four available visual cues were associated with perceived dishonesty. However, links between non-verbal behaviour and perceived deception have been studied only in terms of after-the-fact judgements, or by assessing listeners' explicit beliefs about cue validity (see Vrij & Semin, 1996; Zuckerman, Koestner, & Driver, 1981). How and when these cues are incorporated into judgements of deception remains unclear.

Research in the field of gestures has explored the time course of comprehension of iconic gestures (movements which visually represent content). This work suggests that information presented in the visual modality is integrated into language comprehension along a similar time course as the processing of speech (see e.g., Kelly et al., 2004; Özyürek et al., 2007). For instance, iconic gestures which are incongruent with sentential context have been associated with electrophysiological responses which are similar in latency, amplitude, and topography to those elicited when the incongruity is presented in speech (Özyürek et al., 2007).

Studies exploring the influence of a speaker's body language on pragmatic comprehension (e.g., Kelly et al., 1999) have not investigated the time-course during which this influence emerges, and a speaker's non-verbal behaviours are substantially more varied than speech hesitations, serving both as potential markers

of meta-cognitive states and planning processes, and as an alternative modality through which the speaker conveys semantic information (see e.g., Ekman & Friesen, 1969; McNeill, 1992). Therefore any process linking a speaker's movements with deception must be subtle enough to discriminate types of non-verbal behaviours, or risk over-attribution by labelling irrelevant cues as signs of deceit. Furthermore, listeners associate static visual cues with deception (for instance, averted eye-gaze, see Zuckerman, Koestner, & Driver, 1981), suggesting that judgements of deception are not linked just to variations in body movement, but to an array of non-verbal cues. Here, we aim to shed light on this wider question of how visual information about a speaker is integrated into the pragmatic interpretation of language, by investigating if and when the time course of listeners' judgements of deception is influenced by a variety of non-verbal behaviours in a similar way to hesitations and other auditory aspects of the manner of speech.

The two experiments presented here adapt the 'treasure game' paradigm from Loy et al. (2017) to include a video of a potentially deceptive speaker describing the location (behind one of two objects) of some hidden treasure on the screen (see Fig 6.1). Crucially, we manipulate the presence or absence of potential visual cues to deception in the video. Listeners hear and watch the speaker, attempting to guess, and click on, the true location of the treasure, which allows us to infer whether they believe the speaker to be lying or telling the truth. If listeners associate a given visual cue with deception, then following these cues they should be more likely to click on the object which has not been mentioned. By measuring listeners' eye and mouse movements as the speaker's descriptions unfold, we can investigate their interpretations of what is being said over time.

In Experiment 6.1, we focus on how trunk movements (postural shifts) influence judgements of deception, with filler trials presenting two further types of non-verbal behaviour (adaptor gesturing, and different static postures). Our focus on trunk movements is based on previous research indicating that listeners perceive

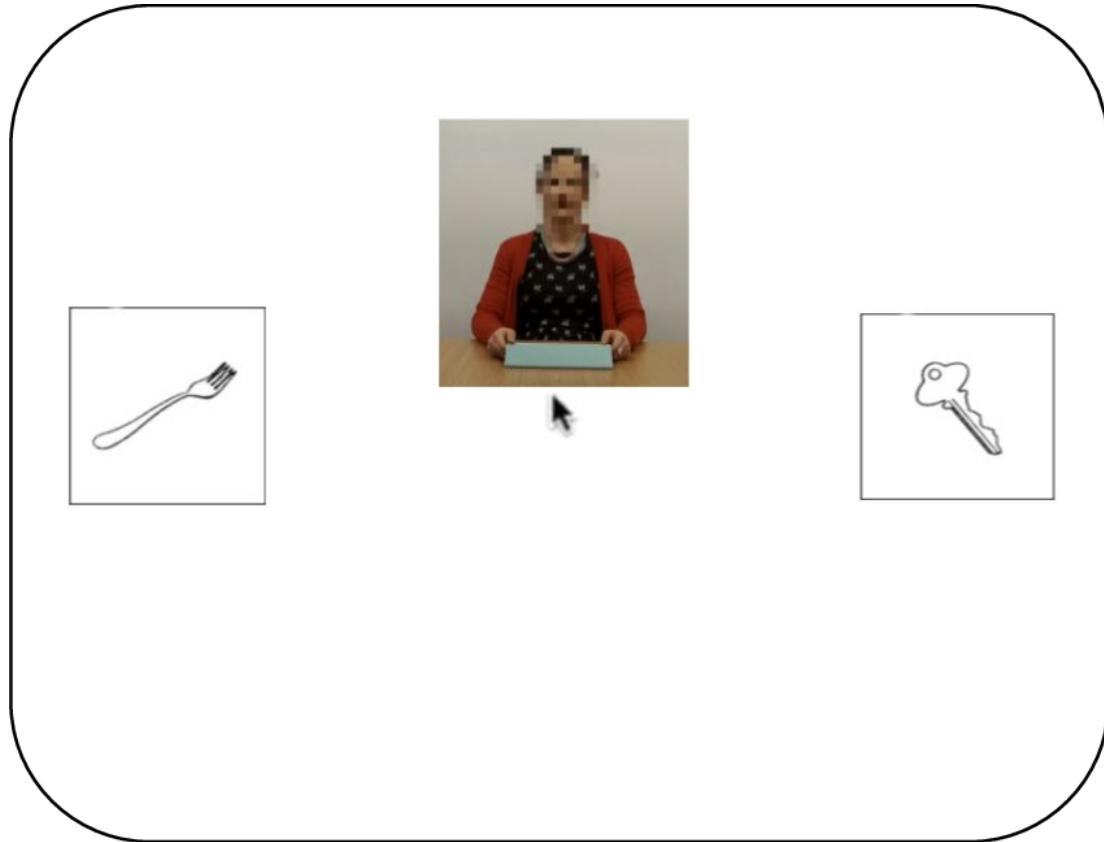


Figure 6.1: Layout of experimental display: Visual-world-paradigm with video stimulus.

these movements as cues to lying (Vrij & Semin, 1996; Zuckerman, DePaulo, & Rosenthal, 1981). Trunk movements are also a plausible utterance-initial gesture (see Cassell, Nakano, Bickmore, Sidner, & Rich, 2001), allowing us to ensure that gestures can be viewed in their entirety before visual targets are referred to. Based on a post-hoc analysis of filler trials which suggested that listeners' judgements were in fact most strongly influenced by the speaker's adaptor gestures, we designed Experiment 6.2 to replicate this latter effect.

6.3.3 Experiment 6.1

Experiment 6.1 makes use of eye- and mouse-tracking to investigate whether a speaker's non-verbal behaviours affect a listener's judgements of deception over time. The experiment was presented as a 'lie detection game'. Each trial included a video and audio recording of a potentially deceptive speaker describing the location of some hidden treasure. Throughout a trial, two images, depicting potential treasure locations, remained visible on the screen. Participants were tasked with using the mouse to click on the object they believed to be concealing the treasure. Critical trials presented videos of the speaker either producing a trunk movement immediately prior to utterance playback, or sitting motionless (no cue) for the equivalent amount of time. Filler trials presented videos of the speaker producing no cue, sitting in a different posture, or producing an adaptor gesture. Our aim was to investigate whether and when these non-verbal cues would be associated with falsehood.

Participants

Twenty-four self-reported native speakers of English were recruited from the University of Edinburgh community, and took part in the experiment in return for a payment of £4. Consent was obtained in accordance with the University of Edinburgh's Psychology Research Ethics Committee guidelines (ref number: 9-1718/1). Participants all had normal or corrected-to-normal vision, and were all right-handed mouse users.

Materials

Visual stimuli consisted of the same 120 line drawings from Snodgrass and Vanderwart (1980) which were used in Loy et al. (2017), sixty of which served as the object named as hiding the treasure (referents) and the other sixty as distractors. Referents were randomly paired with distractors and presented across sixty trials (20 critical trials and 40 fillers). Critical referents and distractors were matched for both ease of naming and familiarity. Each pair of referents was associated with an audio recording of fluent speech specifying the image as the object that the treasure was hidden behind (“The treasure is behind the <referent>”), taken from Loy et al. (2017).

To create the video recordings to use with the previously-recorded audio stimuli, we recorded a volunteer repeating the phrase “the treasure is behind the <object>” while either sitting motionless or performing a given gesture (trunk movement, adaptor gesture, different static posture). Videos showed the speaker in front of a plain white background, seated at a table with a tablet computer on it (where the referent, distractor, and treasure were purported to be displayed). The face shown in each video was pixelated, to allow different videos to be associated with different audio recordings without providing evidence that the visual and auditory channels had been recorded separately.

The video recordings were paired with the fluent audio recordings from Loy et al. (2017). In the 20 critical trials, the audio recordings were paired with 10 videos showing the speaker producing no cue, and five different videos of trunk movements (each used in two different critical trials). The critical trials were counterbalanced across two lists, such that audio recordings paired with a motionless speaker in one list were paired with trunk movements in the other. Forty filler trials were added to each list. In these trials, the 10 videos showing no cue from critical trials were each presented in two trials, and 20 videos of other gestures (10 showing adaptor

gesturing, and 10 showing the speaker motionless but in a different posture) were presented once each. For each participant, each of these 40 videos was randomly paired with a pair of images and an audio track, with no repetition of referents across items. Over the course of the experiment each participant saw 30 videos in which the speaker produced no gestural cue, and 30 videos of the speaker producing a potential cue to deception (10 in critical trials, showing trunk movements; 10 fillers with adaptor gestures; 10 fillers showing different postures).

We identified a time-point in each video recording at which, according to our judgement, it would be natural for audio to begin. For videos showing a trunk movement, this was the frame of the video at which the movement ended, meaning that there was no overlap between the gestural cue and the ensuing speech. The time to audio onset was matched in videos showing no cue, thus controlling for any sensitivity to the duration of video prior to speech. For videos showing an adaptor gesture the amount of overlap between the visual cue and speech varied according to the experimenters' judgements of what appeared natural; time to audio onset was matched in videos showing the speaker in different static postures.

Procedure

The experiment was presented using OpenSesame version 3.1 (Mathôt et al., 2012). Stimuli were displayed on a 21 in. CRT monitor with a resolution of 1024×768 , placed 850 mm from an Eyelink 1000 Tower-mounted eye-tracker which tracked eye movements at 500 Hz (right eye only). Audio was sampled at 44100 Hz and presented in stereo from speakers on either side of the monitor. Videos were presented at 25 fps, and mouse coordinates were sampled at every frame (every 40 ms). Eye movements, mouse coordinates and object clicked (referent or distractor) were recorded for each trial.

Fig 6.2 represents a sample trial from the experiment. Between trials, participants underwent a manual drift correction to ensure accurate recordings from the eye-tracker. After this, the central fixation dot turned red for 500 ms to signify progression to the trial. This was replaced by two images corresponding to the referent and distractor, each measuring 150 × 150 pixels, centered vertically and positioned such that the center of each object was 15% from either edge of the display. The positions (left vs. right) of referents and distractors were randomly chosen, with the constraint that for each participant, referents occurred equally often on each side, separately for critical and filler trials. 2000 ms after the onset of the image display, a video was added to the screen, and the mouse pointer was centred and made visible. The video, measuring 266 × 284 pixels, was displayed with the bottom edge at the vertical midpoint of the screen and centered horizontally. Playback of the audio recording began at the assigned frame of the video (see materials above). The trial ended once the participant clicked on either object, or timed-out 5000 ms after onset of the referent noun, at which point participants saw a message telling them to click on subsequent objects faster.

Participants were told that they were watching recordings taken from a previous experiment, in which one participant was tasked with describing the location of some hidden treasure with the aim of misleading another participant into choosing the wrong location. To emphasise this, the instructions included a photograph of two people purportedly participating in this previous experiment. Participants were told that the speakers in the previous experiment had lied approximately half of the time. Participants were instructed to click on the object behind which *they believed* the treasure to be hidden, with the overall aim of accumulating as much treasure as they could across the experiment. Participants received no feedback after their object clicks, except on bonus trials, which are described in the next section. They were told that the top scorers would be able to enter their names on a high-score table, which was shown at the beginning of the experiment.

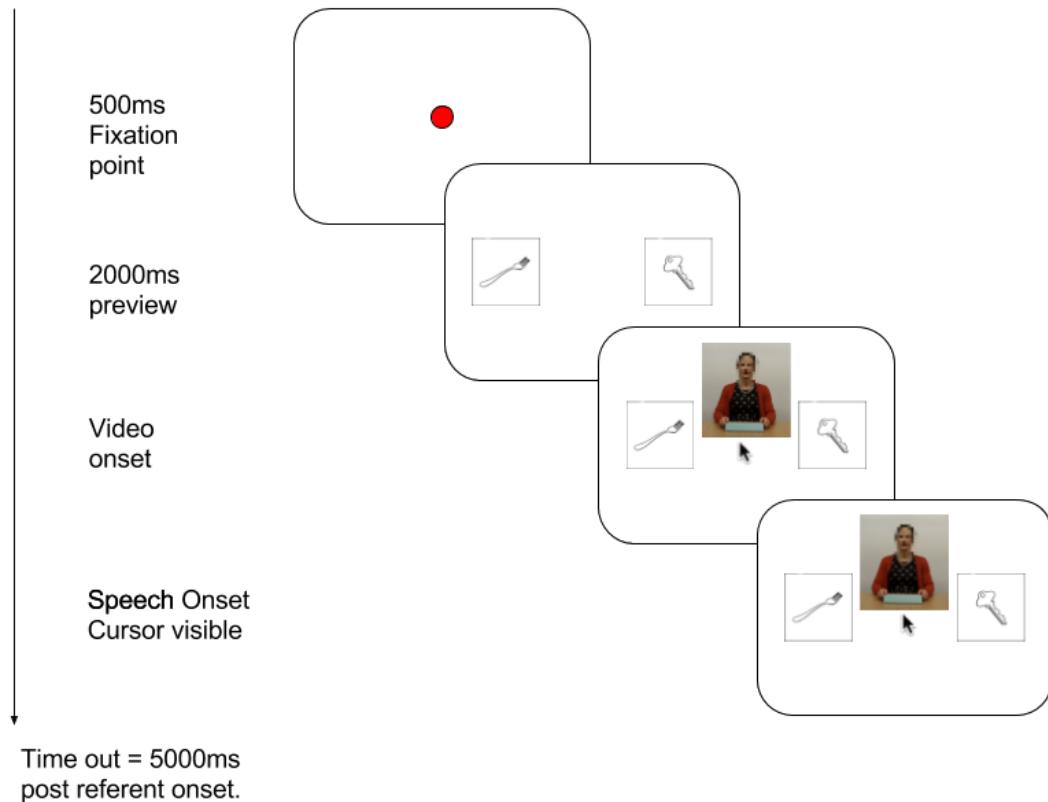


Figure 6.2: Procedure of a given trial, Experiments 6.1 and 6.2

The order of trials was randomly assigned on each run of the experiment. Participants completed five practice trials (one of which was presented as a bonus trial—see below) prior to the main experiment. Two of these presented a video showing no cue, two displayed a video of the speaker in different postures, and one displayed a video of the speaker making a trunk movement.

Bonus Trials

To maintain motivation throughout the study, participants were told that there were a number of ‘hidden bonus rounds’ which offered more treasure than regular rounds. 25% of filler trials (half including a gestural cue; half presenting a video showing no cue) were randomly designated as bonus rounds for each participant.

These trials were visually identical to regular trials. However, following the mouse click (regardless of the object chosen), a message was displayed informing participants that they had successfully located bonus treasure.

Post-test Questionnaire

Participants were asked to complete a short post-test questionnaire which asked whether they had noticed anything odd about the visual or audio stimuli. Any participant who indicated that they had noticed anything unusual was then questioned further, to decide whether they believed that the speech and gesture had been produced naturally and simultaneously. All participants were subsequently debriefed, during which they were told that the audio and video were created separately and stitched together, and asked again verbally if they had noticed anything unusual in that respect. Responses to the questionnaire and debrief were used to determine whether participants should be excluded from the analysis.

6.3.4 Results

Analysis

Analysis of critical trials was carried out in R version 3.5.1 (R Core Team, 2018), using the lme4 package version 1.1-17 (Bates et al., 2015). Data from four participants who indicated suspicion of the supposed origins of the audiovisual stimuli based on the post-test questionnaire and/or debrief were removed from all analyses, leaving data from twenty participants. Of the resultant 400 critical trials, one trial, in which the participant did not click on either the referent or distractor, was excluded from all analyses.

The object clicked (referent or distractor) was modelled using mixed effects logistic regression, with a fixed effect of non-verbal behaviour in the video (no cue vs. trunk movement, dummy coded with no cue as the reference), and with random intercepts and slopes for non-verbal behaviour both by-participant and by-item. Time taken to click an object (measured from referent onset) was log transformed and modelled using mixed effects linear regression with fixed effects of object clicked (referent vs. distractor, dummy coded with referent as the reference) and non-verbal behaviour (no cue vs. trunk movement, dummy coded with no cue as the reference). Random intercepts and slopes for non-verbal behaviour were included both by-participant and by-item, as well as random slopes by-participant for object clicked.

Eye fixation data was averaged into 20 ms bins (of 10 samples) prior to analysis. For each bin, we calculated the proportions of time spent fixating each of the referent and the distractor. The position of the mouse was sampled every 40 ms. Using the X coordinates only, we calculated the number of screen pixels moved and the direction of movement (towards either referent or distractor). We then calculated the cumulative distance travelled towards each object over time as a proportion of the cumulative distance travelled in both directions from referent onset up until that time bin. Movements beyond the outer edge of either object were considered to be ‘overshooting’ and were not included in calculations (0.8% of samples).

The proportions of fixations and mouse movements to either object were empirical logit transformed (Barr, 2008), yielding measures for which a value of zero indicates no bias towards either object, and positive and negative values indicate a bias towards the referent and distractor respectively.

As in previous studies using the treasure game paradigm (King, Loy, & Corley, 2018; Loy et al., 2017), eye- and mouse-tracking analyses were conducted on a

time-window beginning at referent onset and extending for 800 ms, just beyond the duration of the longest critical referent name (776 ms). Transformed fixation proportions for eye and mouse movements were modelled over this time window using linear mixed effects models, with fixed effects of time from referent onset (seconds), non-verbal behaviour (dummy coded with ‘no cue’ as reference level), and their interaction. Random intercepts and slopes for time and non-verbal behaviour were included both by-item and by-participant. Following Baayen (2008), we considered effects in these models to be significant where $|t| > 2$.

Object clicks

Participants clicked on the referent (named object) in 56% of critical trials and the distractor in 44%. Table 6.1 shows the numbers of clicks across all participants to either object split by whether the video showed no cue or a trunk movement. Participants were more likely to click on the referent than the distractor following a video showing no cue. There was a marginal reduction of this bias following videos of the speaker producing a trunk movement ($\beta = -0.56$, SE = 0.32, $p = .08$). There was no effect of non-verbal behaviour on the times taken by participants to click on an object.

Table 6.1: Objects clicked in critical trials in Experiment 6.1: Clicks recorded on each object (referent or distractor) split by condition (no cue vs. trunk movement).

	No Cue	Trunk Movement
Clicks to Referent	125 (62.5%)	99 (49.7%)
Clicks to Distractor	75 (37.5%)	100 (50.3%)

Eye movements

Fig 6.3 shows the time course of fixations to referents, distractors and videos in critical trials for the 2000 ms from referent onset, split by whether the video showed no cue or a trunk movement. Analysis conducted over the 800 ms period from referent onset showed that, following videos showing no cue to deception, participants became increasingly likely to fixate the referent over the distractor as this window progressed (as indicated by a main effect of time $\beta = 2.23$, SE = 0.67, $t = 3.34$). Importantly, there was no interaction of time with non-verbal behaviour, indicating that the presence of trunk movements did not have an early influence on participants' increasing fixation bias toward the referent.

Mouse movements

Fig 6.4 shows the time course of the proportions of cumulative distance the mouse moved towards the referent and distractor in critical trials for the 2000 ms period from referent onset, split by whether the video showed either no cue or a trunk movement. Analysis of the 800 ms following referent onset showed that participants' mouse movements patterned with their eye movements: Over the course of the window participants were increasingly likely to have moved more towards the referent than the distractor following videos of no cue (main effect of time: $\beta = 0.30$, SE = 0.12, $t = 2.41$). Videos showing the speaker producing a trunk movement did not influence this increasing referent bias.

6.3.5 Additional Analyses of Filler trials

In the post-test verbal questioning, 8 participants (40%) specifically mentioned responding to the speaker's hand-movements in their judgements of whether or

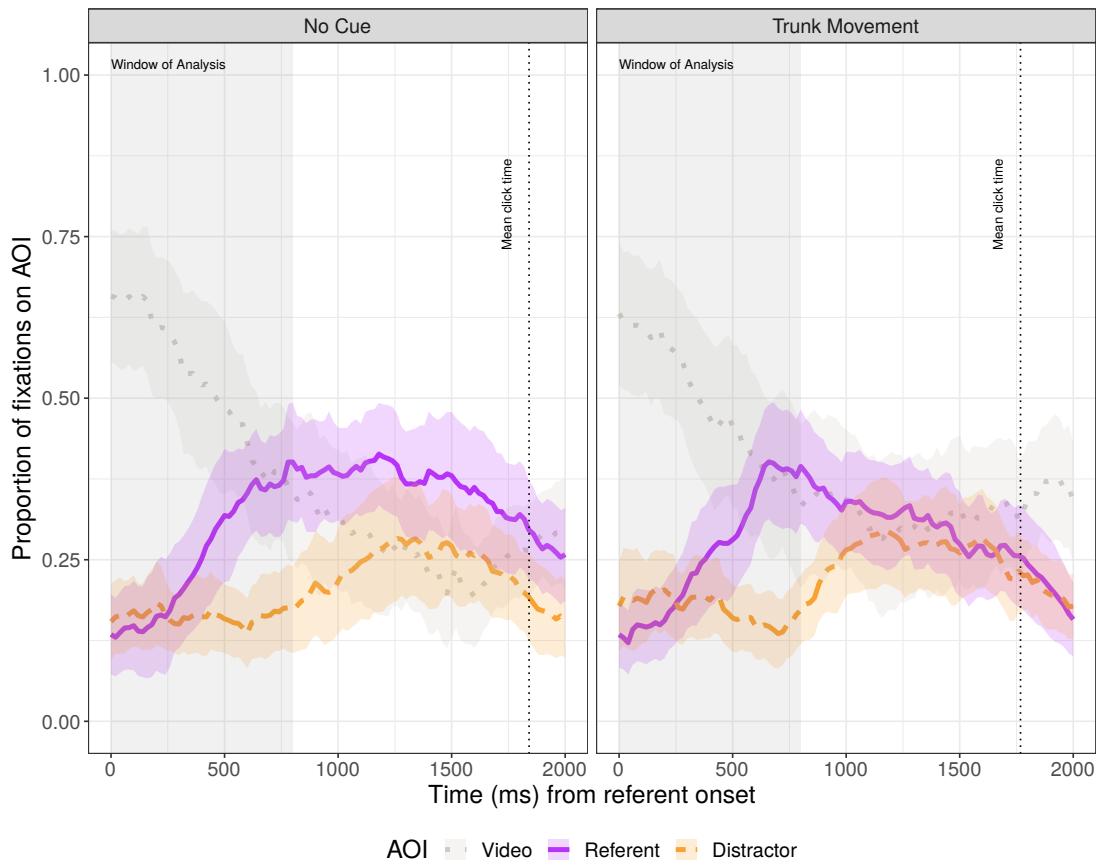


Figure 6.3: Eye-tracking results for critical trials in Experiment 6.1: Proportion of fixations to each object (referent or distractor) and the video, from 0 to 2000 ms post-referent onset, calculated out of the total sum of fixations for each 20 ms time bin. Shaded areas represent 95% confidence intervals derived via bootstrapping subject data ($R=1000$). Dotted lines indicate mean click time by condition.

not the speaker was deceptive. We therefore conducted post-hoc analyses on filler trials to investigate whether the types of non-verbal behaviours presented in these trials (different postures and adaptor gesturing) were influencing participants' judgements of deception. Analysis of filler trials was conducted on 797 trials (3 trials were excluded from analysis due to no mouse click on either object), with non-verbal behaviour comprising three levels: No cue, different posture and adaptor gesture (dummy coded in all analysis, again with 'no cue' as the reference). The time-window of analysis for eye and mouse movements was extended to

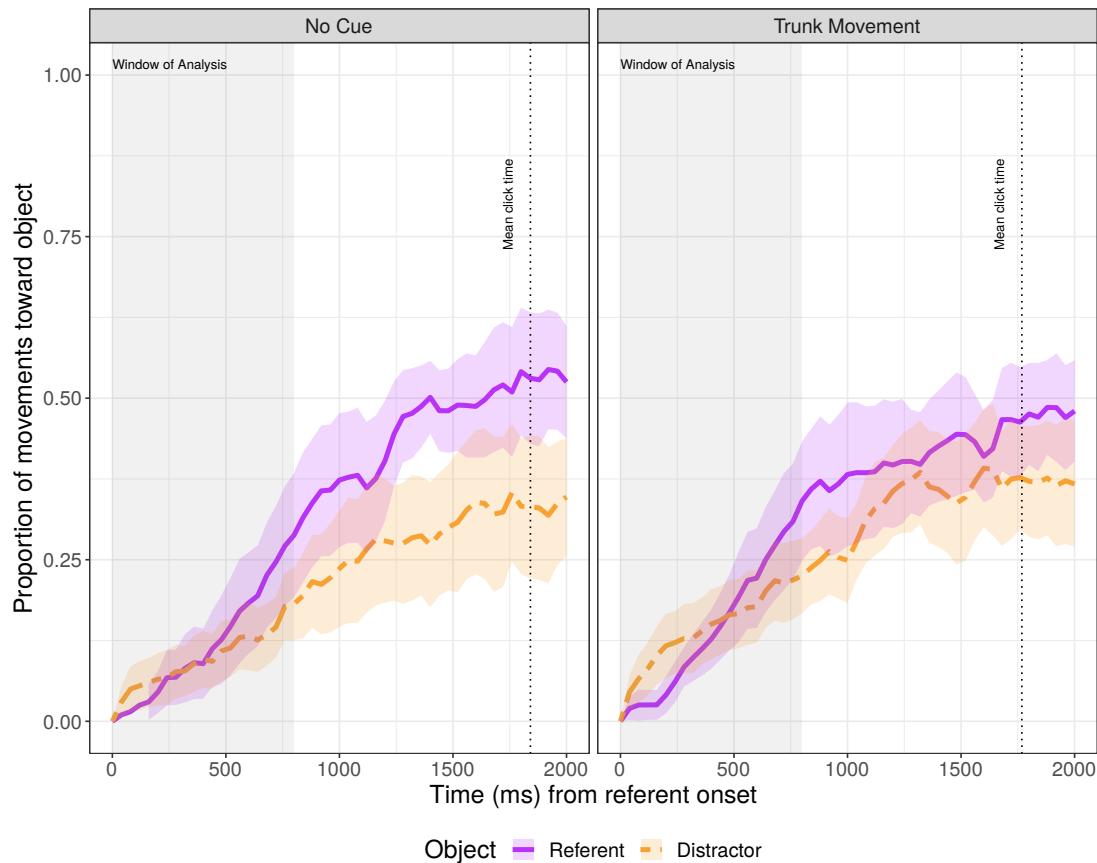


Figure 6.4: Mouse-tracking results for critical trials in Experiment 6.1: Proportion of cumulative distance travelled toward each object from 0 to 2000 ms post-referent onset. Proportions were calculated from the total cumulative distance participants moved the mouse until that time bin. Shaded areas represent 95% confidence intervals derived via bootstrapping subject data ($R=1000$). Dotted lines indicate mean click time by condition.

1100 ms to include the duration of the longest referent in filler trials (1062 ms). The models were specified in the same way as those for the critical trials, with the exception that there were no by-item random effects of non-verbal behaviour, because the videos in filler trials were not counterbalanced across items.

Object clicks

Table 6.2 shows the numbers of clicks in filler trials across all participants to either object, split by the type of non-verbal behaviour presented in the video. For trials in which the video showed a speaker producing no cue, participants tended to click on the referent rather than the distractor ($\beta = 0.63$, SE = 0.16, $p < .001$). For trials in which the videos showed the speaker either in a different posture or producing an adaptor gesture, this bias to click on the referent was reduced ($\beta = -0.73$, SE = 0.31, $p = .02$) and $\beta = -1.03$, SE = 0.33, $p = .002$ respectively), suggesting that presence of these types of non-verbal cues influenced participants' final judgements of whether the speaker was truthful or dishonest.

Table 6.2: Objects clicked in filler trials in Experiment 6.1: Clicks recorded on each object (referent or distractor) split by each type of non-verbal behaviour presented in the video.

	No-Cue	Different Posture	Adaptor Gesture
Clicks to Referent	256 (64.5%)	96 (48.0%)	83 (41.5%)
Clicks to Distractor	141 (35.5%)	104 (52.0%)	117 (58.5%)

Eye movements

Fig 6.5 shows the time course of proportions of fixations to referent, distractor and video split by the type of non-verbal behaviour shown in the filler trials. Analysis conducted on the 1100 ms following referent onset revealed that, as in critical trials, participants tended to fixate the referent over the distractor more as time increased ($\beta = 1.05$, SE = 0.34, $t = 3.13$). However, in contrast to the critical

trials, this bias towards the referent over time was attenuated following videos showing the speaker in a different posture, and those in which the speaker was shown to produce an adaptor gesture ($\beta = -0.58$, $SE = 0.13$, $t = -4.42$ and $\beta = -0.96$, $SE = 0.13$, $t = -7.31$ respectively).

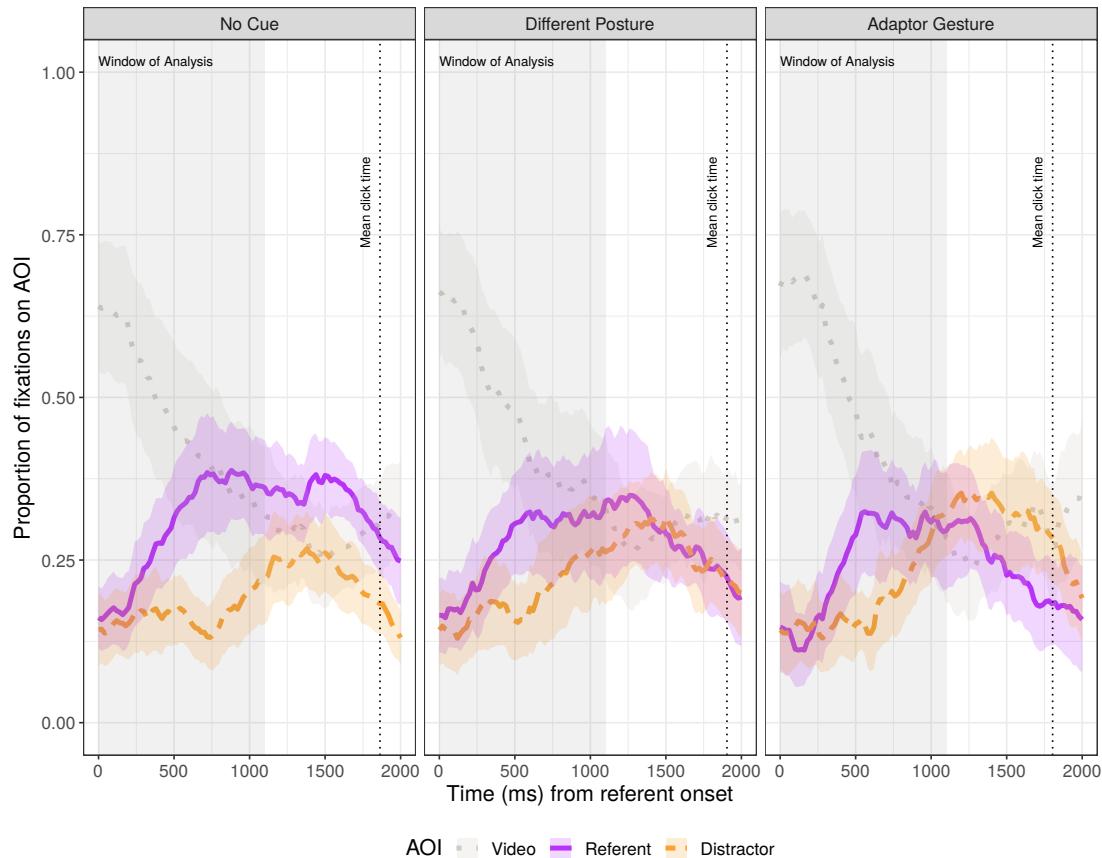


Figure 6.5: Eye-tracking results for filler trials in Experiment 6.1: Proportion of fixations to each object (referent or distractor) and the video, from 0 to 2000 ms post-referent onset, calculated out of the total sum of fixations for each 20 ms time bin. Shaded areas represent 95% confidence intervals derived via bootstrapping subject data ($R=1000$). Dotted lines indicate mean click time by condition.

Mouse movements

Fig 6.6 shows participants' mouse movements towards the referent and distractor split by the type of non-verbal cue shown in the filler trials. Analysis conducted on the 1100 ms following referent onset showed that mouse movements patterned with eye movements. The increasing bias over time to move the mouse towards the referent rather than the distractor following videos in which the speaker produced no cue (main effect of time: $\beta = 0.40$, SE = 0.08, $t = 4.71$) was reduced following videos showing the speaker in a different posture ($\beta = -0.22$, SE = 0.05, $t = -4.41$) or producing an adaptor gesture ($\beta = -0.35$, SE = 0.05, $t = -7.19$).

6.3.6 Discussion

Experiment 6.1 investigated how the pragmatic inferences listeners make about a speaker's honesty are influenced by the presence of non-verbal cues to deception, in the form of trunk movements. We presented videos of a potentially deceptive speaker making a statement about the location of some treasure. We measured the eye and mouse movements made by participants who were tasked with clicking on one of two possible treasure locations; one which was mentioned, and one which was not. Participants were thus making implicit decisions about the honesty of each utterance. As in previous studies using versions of this paradigm (King et al., 2018; Loy et al., 2017), participants showed a tendency to interpret an utterance as truthful (as indicated by more clicks to the named object) when there was no obvious cue to deception (i.e., speaking fluently, or sitting motionless). In the videos presented alongside utterances, the presence of a trunk movement prior to speech onset had only a marginal influence on participants' judgements of deception, as evidenced by the objects selected, in contrast to the existing literature (e.g., Vrij & Semin, 1996; Zuckerman, DePaulo, & Rosenthal, 1981).

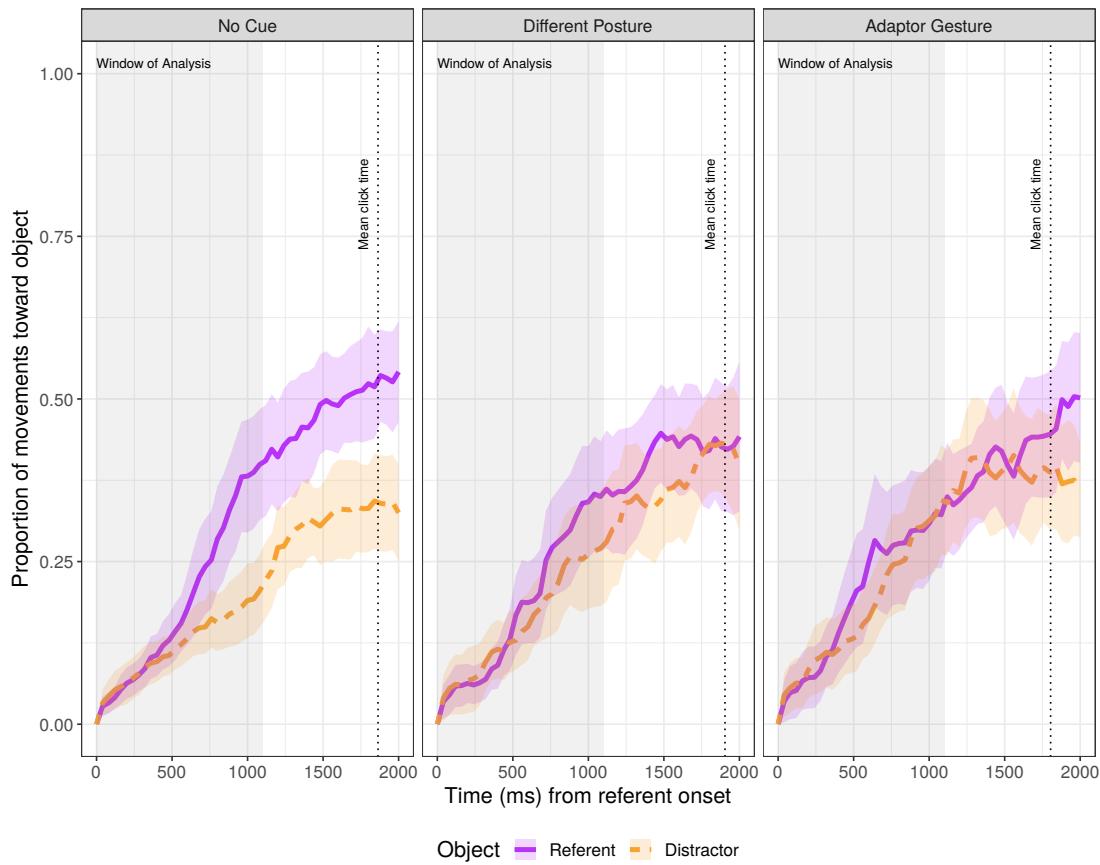


Figure 6.6: Mouse-tracking results for filler trials in Experiment 6.1: Proportion of cumulative distance travelled toward each object from 0 to 2000 ms post-referent onset. Proportions were calculated from the total cumulative distance participants moved the mouse until that time bin. Shaded areas represent 95% confidence intervals derived via bootstrapping subject data ($R=1000$). Dotted lines indicate mean click time by condition.

Similarly, participants' eye and mouse movements during the 800 ms following referent-onset were not affected by whether the video showed the speaker producing a trunk movement or no cue.

Additional analyses of filler trials suggested that participants may however have been influenced by the other types of non-verbal behaviour presented in the experiment: Videos showing the speaker producing either an adaptor gesture or

sitting in a different posture were associated with more judgements of deception than videos showing the speaker producing no cue. Furthermore, the influence of these non-verbal cues was evident soon after the onset of the referent noun, as indicated by a weakening of the biases to fixate, and move the mouse pointer towards, the referent over the distractor. However, the filler trials differed from experimental trials in three important ways. First, referents were not counterbalanced; any findings may have partially or wholly reflected differences between the plausibilities of particular objects as treasure locations. Second, the analysis window for fillers was 1100 ms, rather than the 800 ms used in previous studies (and for the critical items in the current experiment; King et al., 2018; Loy et al., 2017). This allows for a later influence of gesture than of disfluency, rendering direct comparison between modalities difficult. Third, since the trials under consideration here were filler trials, 25% of the items analysed were identified immediately after the mouse click as bonus trials; this may have made particular gestures more salient over the course of the experiment.

From a practical viewpoint, participants' eye and mouse movements in Experiment 6.1 support the compatibility of the visual world paradigm with a range of video stimuli: Viewing videos in which movements co-occurred with speech (e.g., adaptor gestures) did not prevent the emergence of a fixation bias. Fig 6.5 suggests that this bias may begin to emerge alongside the unfolding of the speech stream. Moreover, small movements such as finger tapping appeared to be salient enough to influence participants' final judgements of deception. However, the non-verbal behaviours that appeared to have the greatest influence on participants' judgements were never the intended focus of Experiment 6.1, and these trials differed from critical trials in a number of respects.

In addition to highlighting the salience of hand movements in making deception judgements, responses to the post-test questioning revealed that 4 participants (20%) claimed to rely on 'how relaxed the speaker looked' in making their

judgements, with two of these specifically mentioning that the videos in which the speaker produced no cue presented her in an unrelaxed posture. It is possible that the association between non-verbal behaviour and deception is driven by perceived anxiety. In this case, our findings are largely in keeping with the literature, in that adaptor gestures, but not shifts of posture, have been suggested to be associated with nervousness (Gregersen, 2005). With this in mind, and given that the effects of adaptor fillers in Experiment 6.1 were larger than those of posture changes, we designed Experiment 6.2 as a more controlled investigation of the association between adaptor gesturing and perceived dishonesty. New video stimuli were created to ensure that recordings showed the speaker either producing a typically nervous adaptor gesture, or sitting motionless and in a relaxed posture. There were no filler trials.

6.3.7 Experiment 6.2

Using the same paradigm as Experiment 6.1, participants in Experiment 6.2 heard utterances accompanied by a video of a speaker either producing an adaptor gesture or sitting motionless, and were tasked with making an implicit judgement on whether the speaker was lying or telling the truth.

The videos used in Experiment 6.2 showed adaptor gestures which have previously been suggested to be associated with anxiety (see Gregersen, 2005), and were pre-tested for perceived nervousness in the speaker. This ensured both that videos of gestural cues showed behaviours typically associated with nervousness, and that videos with no cue presented the speaker in a relaxed posture. As a manipulation check, after the treasure-game task, participants were asked to rate how nervous the speaker looked in each video (without audio).

Participants

Twenty-three self-reported native English speaking participants who were right-handed mouse users took part in exchange for £3 compensation. Consent was obtained in accordance with the University of Edinburgh's Psychology Research Ethics Committee guidelines (ref number: 9-1718/1).

Materials

Across 20 trials, 40 images were used (20 referents; 20 distractors). These were the same images as those used in critical trials in Experiment 6.1. As in Experiment 6.1, these images were displayed in referent-distractor pairs, with each pair shown alongside a recorded utterance naming the referent as the location of the treasure. The pairing of referents and distractors on each trial was randomised.

As in Experiment 6.1, each recorded utterance and pair of images was presented alongside a video clip of a person purported to be the speaker of the utterance. Twenty-eight new video clips were recorded (18 different adaptor gestures; 10 no-cue). Care was taken to ensure that the videos including no cue showed the speaker in a relaxed posture. Adaptor gestures were based on descriptions of anxious non-verbal behaviour from Gregersen (2005). All 28 videos were pre-tested for perceived nervousness of the speaker. Ten native English speakers, who did not take part in either of Experiments 6.1 or 6.2, were told that they were going to watch videos (without audio) of someone being questioned in a stressful situation. They were asked to rate how nervous the speaker looked in each video (1: very relaxed, 7: very nervous). The 10 videos showing adaptor gestures with the highest ratings for nervousness (Mean = 4.1, SD = 1.5) were included in the experiment, along with the 10 videos showing no cue (Mean = 1.9, SD = 1.1). Figure 6.7

shows stills from each of the critical videos showing an adaptor gesture and an example of a video showing no gesture.

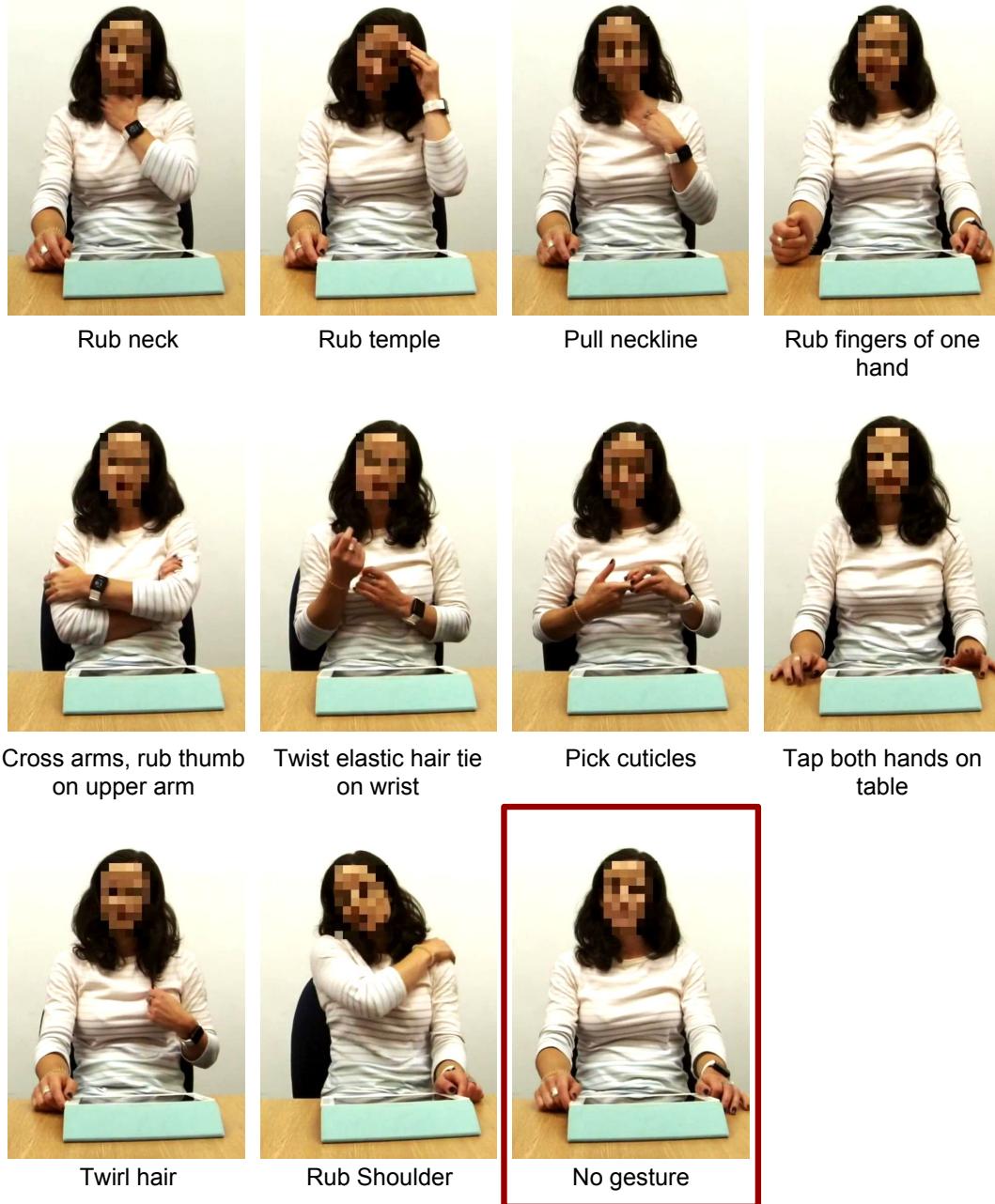


Figure 6.7: Adaptor gestures shown in videos used in critical trials in Experiment 6.2

The 20 referents were counterbalanced across two lists such that each referent that occurred with a video showing adaptor gesturing in the first list occurred with a

video showing no cue in the second. The pairings of referents with specific videos within each condition was randomised for each run of the experiment.

Procedure

The experimental procedure matched that of Experiment 6.1 in all aspects with the exception of the following changes. First, the size of the video stimuli changed slightly to 236×336 pixels, due to videos being recorded in a different room and cropped accordingly to include only the plain background and the speaker. Second, the duration of video presented prior to audio playback was fixed at 1400 ms (after the initiation of gestural cues in all videos) in order to control for participants interpreting the duration from video to speech onset as speech initiation time and in turn associating this with deceit. This was possible as we did not constrain non-verbal cues to be fully presented prior to speech (as we did for trunk movements in Experiment 6.1). Third, because there were no fillers, we did not include any ‘bonus’ trials, so participants did not receive any feedback during the experiment.

After the main task, participants were asked to watch all 20 videos again, without audio, and asked to rate how nervous they thought the speaker looked (using the 1–7 scale described above). Participants then completed the same post-test questionnaire as in Experiment 6.1, with data being excluded from analysis on the same basis.

6.3.8 Results

Data from three participants was excluded because they believed that the audiovisual stimuli were scripted, based on the post-test questionnaire and

questioning during debrief. Analysis was conducted on data from the remaining 20 participants.

Analysis

We followed the same analysis strategy as that used for the critical trials in Experiment 6.1. Of the 400 trials, those which did not result in a click to either object (3) were excluded from analyses. Eye- and mouse-tracking analyses were conducted on the 800 ms window following referent noun onset, just beyond the duration of the longest critical referent name (776 ms).

Participants' post-test ratings (1–7) of how nervous the speaker appeared in each video were analysed using mixed effects linear regression with fixed effects of non-verbal behaviour (no cue vs. adaptor gesturing), by-video and by-participant random intercepts and a by-participant random effect of non-verbal behaviour. Results confirmed that videos of gesturing were perceived as more nervous than videos showing no cue ($\beta = 3.20$, SE = 0.32, $t = 10.08$).

Object clicks

Across the experiment, participants clicked on the referent in 53% of trials and the distractor in the remaining 47%. Table 6.3 shows the numbers of clicks to either object for each type of non-verbal behaviour (no cue vs. adaptor gesturing). As in Experiment 6.1, participants who viewed videos of a motionless speaker were more likely to click on the referent than the distractor ($\beta = 1.53$, SE = 0.23, $p < .001$). The non-verbal behaviour shown in the video was found to influence participants' judgements of deception: Relative to videos showing no cue to deception, those showing adaptor gesturing in the video resulted in fewer clicks on the referent

$(\beta = -2.78, \text{SE} = 0.38, p < .001)$. The time participants took to click on an object was not influenced by the behaviour shown in the video.

Table 6.3: Objects clicked in critical trials in Experiment 6.2: Clicks recorded on each object (referent or distractor) split by condition (no cue vs. adaptor gesture).

	No Cue	Adaptor gesture
Clicks to Referent	161 (80.9%)	48 (24.2%)
Clicks to Distractor	38 (19.1%)	150 (75.8%)

Eye movements

Fig 6.8 shows the time course of fixations to referents, distractors and videos in critical trials for the 2000 ms from referent onset, split by presence of adaptor gesturing. Analyses conducted over a window extending 800 ms from the referent onset reveal a main effect of time ($\beta = 2.99, \text{SE} = 0.67, t = 4.48$), indicating that, as for Experiment 6.1, participants' fixations tend to favour the referent rather than the distractor over time. However, a significant interaction between time and non-verbal behaviour ($\beta = -2.94, \text{SE} = 0.21, t = -14.27$) indicates that the increasing referent-bias was attenuated in trials showing the speaker producing an adaptor gesture.

Mouse movements

Fig 6.9 shows the distance the mouse moved towards the referent and distractor over time, for 2000 ms from referent onset, split by condition. Mouse movements over the course of the 800 ms window from referent onset again patterned with the eye-tracking data: As time from referent onset increased, participants showed an increasing likelihood of having moved the cursor more toward the referent than

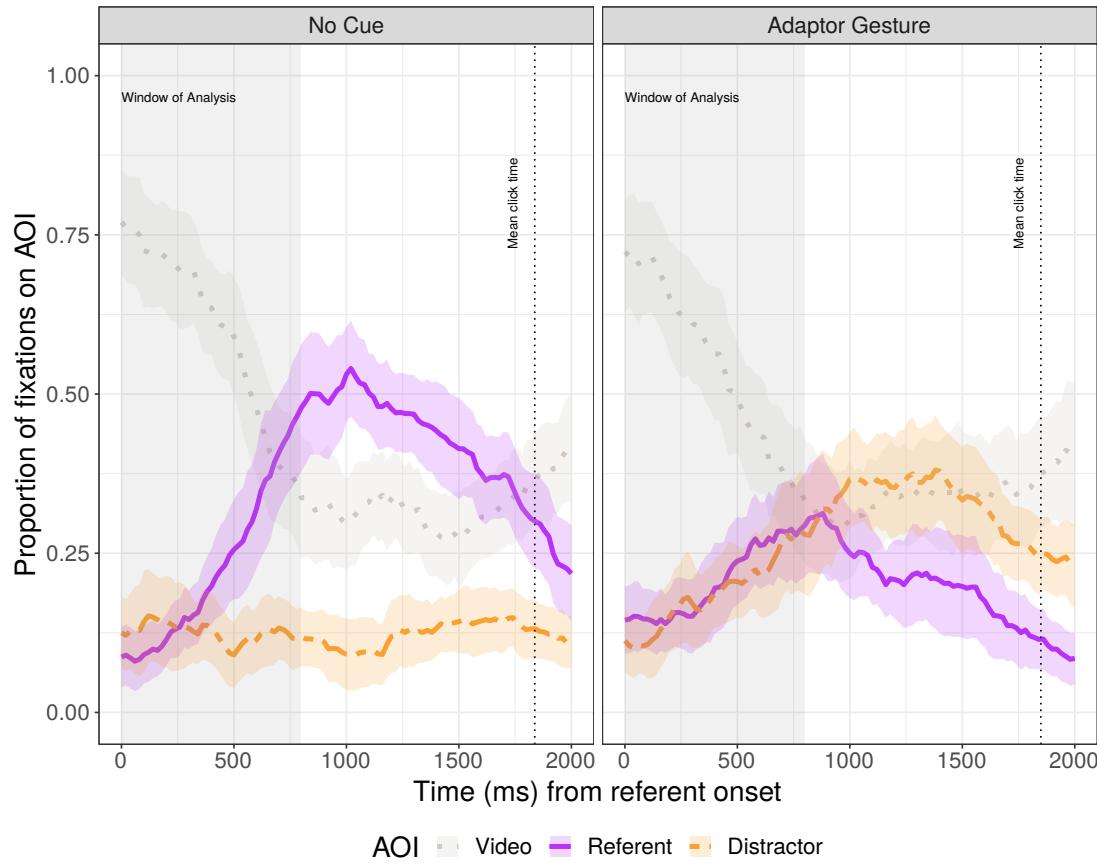


Figure 6.8: Eye-tracking results for Experiment 6.2: Proportion of fixations to each object (referent or distractor) and the video, from 0 to 2000 ms post-referent onset, calculated out of the total sum of fixations for each 20 ms time bin. Shaded areas represent 95% confidence intervals derived via bootstrapping subject data ($R=1000$). Dotted lines indicate mean click time by condition.

the distractor following videos showing the speaker producing no cue ($\beta = 0.61$, $SE = 0.10$, $t = 5.88$), but this was reduced following videos showing an adaptor gesture ($\beta = -0.78$, $SE = 0.08$, $t = -9.27$).

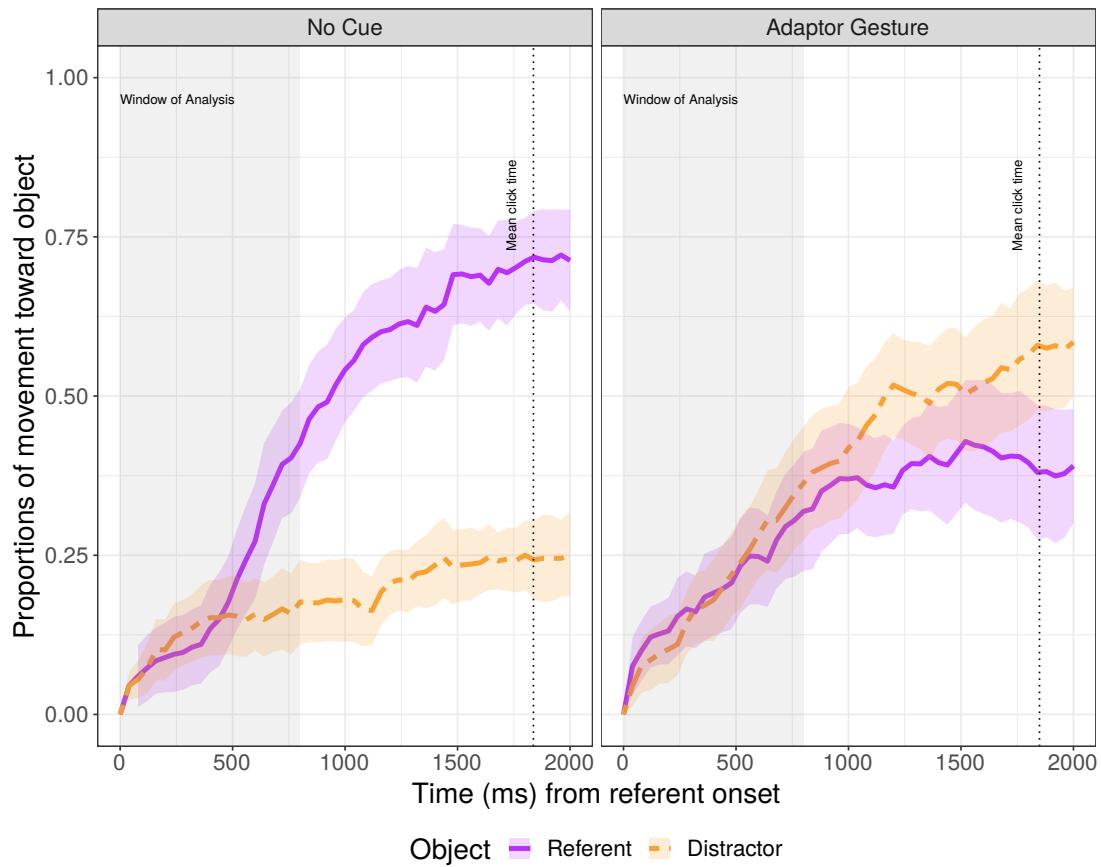


Figure 6.9: Mouse-tracking results for Experiment 6.2: Proportion of cumulative distance travelled toward each object from 0 to 2000 ms post-referent onset. Proportions were calculated from the total cumulative distance participants moved the mouse until that time bin. Shaded areas represent 95% confidence intervals derived via bootstrapping subject data ($R=1000$). Dotted lines indicate mean click time by condition.

6.3.9 General discussion

In Experiments 6.1 and 6.2, we investigated the influence of a speaker's non-verbal behaviour on judgements of deception, focusing respectively on trunk movements and on adaptor gestures. Recorded speakers referred to one of two objects as the location of some treasure. We manipulated the visual presentation of non-verbal cues while measuring listeners' eye and mouse movements towards images of either

the referent named by the speaker, or a distractor object. This allowed us to explore whether, and when, listeners began to associate non-verbal cues with deception.

Contrary to the effects of spoken hesitations (Loy et al., 2017), trunk movements in Experiment 6.1 did not influence the patterns of eye and mouse movements in the 800 ms following referent onset. Instead there was an overall early tendency to fixate on and move the mouse towards the referent over the distractor. The eventual object selected by listeners was only marginally affected by whether or not the video showed the speaker producing a trunk movement. The contrast of these findings with previous research (e.g., Vrij & Semin, 1996) may reflect differences between beliefs about cues to deception (as indicated in questionnaires) and those cues which listeners associated with deception when presented with them. Alternatively, the inclusion of additional non-verbal behaviours in filler trials may have weakened the association between trunk movements and deception which has been found in previous research (Vrij & Semin, 1996; Zuckerman, DePaulo, & Rosenthal, 1981). This is partly supported by studies which found a facilitative effect of illustrative gesturing on listeners' comprehension to be weakened for speakers who produce a lot of other, non-communicative movements (Holle & Gunter, 2007). Finally, evidence points to the importance of temporal synchrony in the integration of illustrative gesturing with speech (see Habets et al., 2011). In Experiment 6.1, trunk movements were presented before the onset of speech; this may have weakened any potential association between cue and interpretation.

Importantly, however, additional analyses of Experiment 6.1 suggested that other types of non-verbal behaviour used in filler trials (different static postures and adaptor gestures) were associated with judgements of dishonesty. Speakers were less likely to click on a referent than a distractor following either of these cues; and eye-tracking and mouse movement records suggest that this bias emerged early—alongside the unfolding of the referent noun in speech. Experiment 6.2 was

conducted to confirm the influence of adaptor gestures on judgements of deception in a study designed specifically to this end. Videos in Experiment 6.2 showed the speaker either producing a typically nervous adaptor gesture or sitting motionless. Results indicate a reliable association between adaptor gesturing and perceived dishonesty, as evidenced by the object selected. Furthermore, the influence of adaptor gesturing on listeners' judgements of deception was present early on: Adaptor gestures reduced the emerging bias to move the mouse towards, and fixate, the referent during the 800 ms post referent-noun onset.

The studies presented here provide a visual-modality parallel with the findings from Loy et al. (2017) which showed that fluency of speech influences judgements of whether a speaker is lying. In keeping with Loy et al. (2017), our results suggest that listeners may have an implicit bias to judge a speaker as honest in the absence of any obvious potential cue to deception—a trend which is present in other studies in deception detection (see Barres & Johnson-Laird, 2003; DePaulo, 1985). In both experiments, utterances presented with the speaker in a neutral posture and not gesturing biased listeners towards believing the speaker to be truthful, as shown by an increased tendency to fixate on, move the mouse towards, and eventually click on the object which was named by the speaker.

Similarly to the effect of manner of spoken delivery on these judgements found in Loy et al. (2017), we found that manner of non-verbal delivery influenced judgements of deception, in particular when the speaker was seen to produce typically anxious adaptor gestures alongside speech. Importantly, this was detectable in the initial stages of linguistic processing, with effects found in Experiment 6.2 during the same time window as that in which Loy et al. (2017) found effects of speech disfluency. However, visual inspection of Fig 6.8 suggests that when presented with a video showing a speaker producing an adaptor gesture, although the referent bias is reduced at an early stage, the tendency to fixate the distractor over the referent—signifying perceived dishonesty—emerges

approximately 1000 ms after the referent began. This contrasts with Loy et al. (2017), in which the visual inspection of the time-course of fixations (see Fig 2 p.1443, and Fig 4 p.1447, Loy et al., 2017) suggests that the bias emerged approximately 600 ms post referent onset following both utterance initial and utterance medial disfluencies. Whether this discrepancy suggests a difference in how audio and visual cues influence pragmatic judgements of deception, or whether this is simply a result of the presence of a video in the display delaying the emergence of a fixation bias, we cannot say. To better understand how information in different modalities affect comprehension, further research would require investigating the effect of spoken delivery when the visual channel is also available: For example, studying the time course of deception judgements when faced with one or both of a disfluency and an adaptor gesture.

Our findings are largely consistent with previous research on beliefs about, and judgements concerning, non-verbal cues to deception, suggesting that listeners perceive a range of non-verbal behaviours, both dynamic and static, as indicative of deceit (e.g., Akehurst, Kohnken, Vrij, & Bull, 1996; Vrij et al., 2000). The lack of a reliable association between trunk movements and judgements of deception shows that care should be taken when generalising from peoples' beliefs about cues to deception (for instance, Vrij & Semin, 1996; Zuckerman, DePaulo, & Rosenthal, 1981) to 'live' situations in which they are faced with a variety of possible cues. Additionally, the studies presented here indicate that the link between non-verbal behaviour and deception may be driven partly by those behaviours which the listener perceives as signalling anxiety in the speaker, although further research is needed to confirm whether this is the case.

It is worth noting that the studies presented here show that it is possible to extend the Visual World paradigm to include visual information about the speaker, and not just the extensional world. By including a video recording of a speaker alongside recorded speech, it is possible to measure the influence of non-verbal

behaviour on listeners' online processing of the unfolding message, even when listeners eventually fixate other images in the display. This is perhaps because listeners are able to extract information about gestures through peripheral vision (see e.g., Gullberg & Holmqvist, 2006). Overall, the studies here show that in utterance processing, the visual channel can have a rapid and direct effect on a listener's pragmatic judgements, supporting the idea that communication is fundamentally multi-modal: Speech and non-verbal behaviour interactively codetermine meaning.

6.4 Chapter discussion

This chapter set out to examine how the non-verbal behaviours presented alongside speech influence listeners' perceptions of deception. Extending a paradigm used previously to investigate the association between speech disfluency and dishonesty (Loy et al., 2017), we presented participants with a task in which they saw and heard a potentially dishonest speaker and, in a given trial, made an implicit judgement about the veracity of the utterance presented to them (by choosing between two objects). Similarly to effects of manner of spoken delivery, our results revealed that non-verbal cues displayed by the speaker influenced listeners' final judgements of deception. Utterances presented with the speaker sitting motionless biased listeners towards believing the speaker to be honest, while adaptor gesturing specifically resulted in a bias towards interpreting the speaker as dishonest.

The findings from this chapter inform this thesis in two respects. Firstly, they support an account that listeners can, in certain contexts, interpret the occurrence of gesturing as collateral signals about the speaker's cognitive and/or emotional states. In Chapter 4, we saw that listeners reliably drew on the occurrence of iconic gesturing to inform their explicit predictions about upcoming referents,

with the direction of this association suggesting that the act of producing iconic gesturing may be interpreted as a signal of speech production difficulty. Here, we saw that a comparable association holds between adaptor gesturing and lying when forming pragmatic judgements of deception. Additionally, the results from the present chapter provide evidence that this association is borne out alongside the lexical processing of speech. Aligning with research in manner of spoken delivery, these findings highlight that the manner of an utterance's non-verbal delivery can influence listeners' interpretations during early moments of comprehension.

Two interesting questions arose from the experiments presented here. The first is concerned with the time taken for biases between adaptor gesturing and deception to fully emerge—i.e., the point at which a preference is shown for the object which implicitly indicates judgements of dishonesty, rather than simply an attenuation of the bias to the named object—which appears (on visual inspection) to be greater than previous research suggests it is for comparable disfluency-dishonesty biases. Secondly, postural shifts, which have emerged from meta-analytical studies as cues which listeners reliably associate with deception (see Hartwig & Bond, 2011; Zuckerman, DePaulo, & Rosenthal, 1981) had only a marginal effect on listeners' final judgements about whether or not the speaker was deceptive, and there was no evidence of any association between this cue and deception in the initial stages of comprehension. It is possible that the variety of cues presented in Experiment 6.1 weakened biases between specific cues and deception, with previous research reporting these strongly held associations perhaps because of the tendency to investigate listeners' explicit beliefs about cue validity rather than those cues which they rely on in making judgements (see Zuckerman, DePaulo, & Rosenthal, 1981). In the next chapter, we address these questions, by examining whether the effects of disfluency and adaptor gesturing on listeners' inferences about deception remain relevant when presented in contexts where speakers present different cues in different modalities. This approach also offers a potential opportunity to examine

differences in the relative time courses of the biases between deception and both gesture and disfluency when both are presented within a multi-modal context.

Chapter 7

Competing cues: Gestural vs disfluent signals to deception

In exploring how meaning is interpreted in the non-linguistic behaviours produced alongside speech, this thesis has thus far investigated two avenues. Firstly, in Part I, we saw evidence suggesting that listeners perceive the presence and duration of iconic gesturing to signal that the speaker is experiencing difficulty in producing speech. We also saw some of the difficulties encountered in using the visual world paradigm to study listeners' real-time predictions alongside the unfolding of gesture and speech.

We subsequently turned from the influence of non-linguistic behaviours on listeners' transient predictions to a way in which they can have a lasting impact on the global interpretation of a message. In Chapter 6, we established that certain non-verbal behaviours can have a reliable influence on listeners' pragmatic comprehension via their interpretation as cues to deception. In doing so, we saw the influence of non-linguistic information on listeners' comprehension emerging rapidly, alongside the unfolding of the linguistic input. Here, we extend this to a context in which a

speaker produces potential cues to deception in both modalities. By manipulating the presentation of two different types of non-linguistic behaviours which have previously been associated with perceived deception—filled pauses in speech (see Loy et al., 2017) and adaptor gestures (Experiment 6.2)—we investigate the relative salience and time course of different non-linguistic cues on listeners' judgements of deception.

Experiments 6.1 and 6.2 presented participants with recordings of speakers referring to one of two objects (displayed on screen) as the location of some hidden treasure. Participants were tasked with clicking with the mouse on the object which they *believed* the treasure to be behind. Also in the visual display was a video purporting to show the speaker producing the utterances heard by participants. Building on a study from Loy et al. (2017) which manipulated the fluency of the utterances presented to listeners, we manipulated the visual presentation of non-verbal cues (trunk movements, adaptor gesturing, and different static postures) while measuring listeners' eye and mouse movements towards images of either the referent named by the speaker, or a distractor object. Similar to the effects of manner of spoken delivery on judgements of deception, non-verbal cues (specifically adaptor gesturing) influenced participants' interpretations of whether or not the speaker was lying. This was borne out in listeners' final judgements of message truth, as well as in their eye- and mouse-movements along a similar time course to the influence of manner of spoken delivery seen in previous research (Loy et al., 2017).

In most human communication, however, listeners are presented with a speaker's behavioural cues in both streams of perceptual input: audio and video. To date, research has focused on one modality at a time (Chapter 6; Loy et al., 2017). Similarly, looking back to Part I of this thesis, Arnold et al. (2007) and Chapters 4 and 5 presented studies of listeners' interpretation of non-linguistic markers of speech planning difficulty in the spoken and visual modalities respectively. By

controlling other aspects of non-linguistic delivery, these studies allow us to isolate listeners' associations between specific behaviours and, for instance, deception.

Research suggests that there are many non-linguistic behaviours, both visible and audible, that may be linked with deception (see DePaulo et al., 1982; Zuckerman, DePaulo, & Rosenthal, 1981). For a listener, both channels offer non-linguistic signals about speakers and the messages they convey. Comparisons between judgements of deception from audio, video and audiovisual stimuli indicate that people tend to make more accurate judgements when presented with the audio channel (see Bond & DePaulo, 2006), but this may be due to a number of reasons: For example, speakers may simply produce fewer non-verbal cues, or those non-verbal behaviours speakers do produce may be less reliable indicators of deception. Little is known about relative weights listeners assign to cues in different modalities in judging the veracity of a statement.

By presenting participants with utterances in which we orthogonally manipulated the presence or absence of audio and visual cues, Experiment 7.1 investigates the relative salience of manner of spoken delivery and manner of non-verbal delivery on judgements of deception. Specifically, we explore how perception of deception is influenced by different combinations of speech (dis)fluency and adaptor gesturing. We also explore the possibility that these non-linguistic cues are additive in their effect on listeners' judgements of deception: Does disfluency in speech *as well as* adaptor gesturing increase the likelihood of an utterance being interpreted as dishonest? Alternatively, do listeners simply judge any deviation from stereotypically normal behaviour (fluent, no gesturing) as indicative of deception, regardless of whether there are single or multiple cues? Finally, Experiment 7.1 is also concerned with the relative time courses of how information about manner of spoken delivery and manner of non-verbal delivery influence judgements of deception.

7.1 Perception of deception in different modalities

Listeners are affected by non-linguistic information in various ways. The manner in which an utterance is delivered can aid comprehension of the literal message, for instance in evaluating syntactic ambiguity (speech disfluency prior to an ambiguous noun phrase biases listeners to assume the noun phrase is the subject of a new clause, see Bailey & Ferreira, 2003) or predicting semantic content (speech disfluency biases listeners to anticipate less familiar words, see Arnold et al., 2007; Barr & Seyfeddinipur, 2010; Corley et al., 2007). They can also have a lasting effect on the global interpretation of a message, for instance in causing a listener to believe that the speaker is telling a lie (Zuckerman, DePaulo, & Rosenthal, 1981).

While there may not be a clearly defined set of behavioural correlates with *actual* lying, a wide range of non-linguistic behaviours have been found to be reliably perceived as signalling deception (for discussions of actual and perceived cues to deception, see DePaulo et al., 1982; Hartwig & Bond, 2011). Aspects of the manner of spoken delivery (such as speech rate, speech errors, and vocal pitch) and many non-verbal behaviours (foot and leg movements, self-adaptive gestures, gaze direction and postural shifts) have been found to reliably *be believed to* indicate lying Zuckerman, DePaulo, and Rosenthal (1981). As well as showing listeners' beliefs about cues to deception, research has shown how certain non-linguistic behaviours—including increases of speech disfluency and arm movements—are reliably used in judgements of whether or not a speaker is lying (Hartwig & Bond, 2011, , for a recent meta-analysis of 128 studies).

Recent research (Chapter 6, Loy et al., 2017) has shown that listeners' associations between different aspects of delivery and perceived deception can have a rapid

effect on comprehension. However, in both Loy et al. (2017) and Chapter 6, possible cues to deception were presented in one modality. Loy et al. (2017) manipulated the presence of speech fluency (both utterance-initial and utterance-medial, in Experiments 1 and 2 respectively), with disfluent utterances created by splicing a disfluency into the initial utterance fragment used in the fluent utterances. Analogously, Experiments 6.1 and 6.2 manipulated whether the video accompanying a fluent utterance showed the (purported) speaker producing a gesture or sitting motionless.

Comparisons between deception judgements based on audio, video, or audiovisual stimuli, have indicated that the audio channel may be more salient to listeners forming these judgements. For example, relative to stimuli presented in both audio and video, listeners are less accurate in detecting deceit when responding to video-only stimuli, but not when responding to audio-only stimuli (see Bond & DePaulo, 2006). Often, however, studies of perceived deception have tended not to directly manipulate the presence or absence of specific cues (as is done in Loy et al., 2017), instead comparing judgements of utterances from speakers naturally varying in their non-linguistic behaviours. This makes it difficult to discern between whether between-modality differences in the accuracy of deception judgements are due to the relative frequency, reliability or salience of cues: For example, lower rates of accuracy in detecting lies using video-only stimuli may be a result of the stimuli containing fewer, or less reliable cues, than the audio channel.

The present experiment tests listeners' sensitivity to different types of non-linguistic cues to deception in a multi-modal setting. Combining aspects of Loy et al. (2017) and Experiment 6.2, we study listeners' judgements of deception when faced with different combinations of non-linguistic cues across modalities. Following a two (Fluent vs. Disfluent) \times two (No Gesture vs. Adaptor Gesture) design, we present participants with multi-modal statements about the location of some hidden

treasure, and—as before—task them with clicking on the object they believe the treasure to be behind.

Experiment 7.1 asks two questions. Firstly, how do listeners' associations between visual and spoken cues and deception hold in situations in which speakers produce multiple cues in different modalities? By directly manipulating the presence and absence of filled pauses and adaptor gestures, we investigate whether listeners associate these cues with deception to different extents.

Secondly, the present experiment studies the time course over which non-linguistic cues in different channels influence judgements of deception, to test whether cues vary in how quickly they are linked with deception. Loy et al. (2017) and Experiment 6.2 both found early effects (of disfluency and adaptor gesture respectively) on participants' eye and mouse movements—within 800 ms of playback of the referent-noun. However, visual inspection of the later time course suggests that the point at which fixations to the distractor overtook those to the referent emerged approximately 1000 ms post referent-noun onset in Experiment 6.2—400 ms later than in Loy et al. (2017) (see Fig 2 p.1443 and Fig p.1447 in Loy et al. 2017 and Figure 6.8 in Chapter 6 of the present thesis). This difference could be due to the video component in Experiment 6.2 detracting from participants fixating on other objects in the display. Alternatively, it could reflect differences in how these cues are associated with lying. By presenting participants with both of these cues in the same multi-modal setting, Experiment 7.1 investigates the relative time courses during which filled pauses and adaptor gesturing influence judgements of deception.

7.2 Experiment 7.1

Experiment 7.1 investigates the influence of manner of spoken delivery and manner of non-verbal delivery on listeners' judgements of deception. As in Loy et al. (2017) and Experiments 6.1 and 6.1, Experiment 7.1 presents participants with recordings of a speaker referring to one of two objects (displayed on screen) as the location of some hidden treasure. Participants are tasked with using the mouse to click on the object which they *believe* the treasure to be behind. Also presented in the visual display is a video purporting to show the speaker producing the utterances heard in the recordings. The utterances participants hear are either fluent or disfluent ("The treasure is behind [the]/[thee, uh] <referent>"). The videos present a subject (purported to be speaker of the utterances) either sitting motionless or producing an adaptor gesture (fidgeting, tapping, adjusting hair or clothing). Investigating the objects participants click on in the experiment allows us to test the relative impact of different combinations of spoken and gestural cues on listeners' final judgements of whether an utterance is a truth or a lie. Participants' eye movements and mouse coordinates were recorded throughout the experiment to investigate the time course over which these cues influence perception of deception.

7.2.1 Participants

Twenty-seven self-reported native speakers of English were recruited from the University of Edinburgh community, and took part in the experiment in return for a payment of £4. Participants all had normal or corrected-to-normal vision, and were all right-handed mouse users. Consent was obtained in accordance with the University of Edinburgh's Psychology Research Ethics Committee guidelines (ref number: 205-1718/1) The experiment was pre-registered at <https://osf.io/4x9ab/>

7.2.2 Materials

A set of 120 line drawings from Snodgrass and Vanderwart (1980) made up the images used in the experiment. This was the same set as used in previous studies using the ‘treasure-game’ paradigm (King et al. 2018; Loy et al. 2017, Experiments 6.1 and 6.2 in this thesis). Sixty of these were used as referents—objects named in trials as hiding the treasure—and the remaining sixty were used as distractors. Across 60 trials (20 critical trials and 40 filler trials), referents were randomly paired with distractors. Referents and distractors in critical trials were matched for ease of naming and familiarity (see Loy et al., 2017, for details), minimizing participants’ biases from utterance (dis)fluency to relative difficulty with which an object can be described (see Arnold et al., 2007).

Each critical referent was associated with recordings of a female speaker naming that image as the object which the treasure was hidden behind. These were the same recordings used in Loy et al. (2017), Experiment 2, and were of either fluent or disfluent utterances. A disfluent segment comprising a prolonged article and a filled pause was spliced into each fluent utterances (“The treasure is behind the <referent>”) to form the disfluent ones (“The treasure is behind thee - uh - <referent>”). An additional 40 utterances were used in filler trials. Half of these included either a form of disfluency or discourse manipulation (see Table 7.1), in order to present participants with a set of utterances which could be believed to be unscripted, and to distract them from the manipulation of the filled pause in critical trials.

Sixty different video recordings were used across the 60 trials. Videos showed a female volunteer sitting at a table on which there was tablet computer angled towards her. In all videos, the volunteer’s face was pixelated to ensure that, when presented alongside recordings of spoken utterances, it appeared believable that the video showed a speaker producing these utterances. The twenty videos used in

Table 7.1: Disfluencies and discourse manipulations in filler items in Experiment 7.1, reproduced with permission from Loy et al. (2017).

Filler type	Manipulation	No. of Utterances	Example
Fluent	None	20	The treasure is behind the <referent>.
Disfluent	Prolongation	3	The treasure is behind <i>thee...</i> <referent>.
	Repetition	4	The treasure is behind <i>the-the</i> <referent>.
	Filled pause (utterance-initial)	3	<i>Umm..</i> The treasure is behind the <referent>.
Other	Discourse marker	5	<i>Okay,</i> the treasure is behind the <referent>.
	Modal	3	The treasure <i>could be</i> behind the <referent>.
	Combination	2	<i>Right,</i> the treasure <i>might be</i> behind the <referent>.

critical trials were the same as those used in Experiment 6.2, showing ten different adaptor gestures (self- or object-adaptive movements such as fidgeting or tapping) and ten videos of the volunteer sitting motionless. An additional 40 videos were used in filler trials. Matching the variation in manner of spoken delivery in filler utterances, half of the filler videos showed the volunteer producing a different adaptor gesture (see Table 7.2).

The 20 critical referents were counterbalanced across four lists, each containing 10 videos of adaptor gestures (5 with fluent utterances; 5 disfluent) and 10 videos showing no gesture (5 fluent; 5 disfluent), such that referents occurring in the first

Table 7.2: Non-verbal manipulations in filler videos in Experiment 7.1

No. of Videos	Manipulation
20	None [Sitting motionless]
2	Finger-tapping on table
2	Hands on table, rubbing fingers of one hand
2	One hand rubs opposite shoulder
2	One hand twirls hair
2	One hand puts hair behind ear
1	Finger-tapping on tablet computer
1	Hands on table, rubbing fingers of both hands
1	Hands in front of torso, rubbing hands together
1	Hands in front of torso, one hand pulls on elastic band on other wrist
1	Hands in front of torso, rubbing hands together
1	Arms crossed, one hand rubs other arm
1	One hand rubs back of neck
1	One hand strokes chin
1	One hand adjusts neckline of top
1	One hand adjusts shoulder of top
1	One hand adjusts sleeve of top

list in one condition occurred in the other conditions in the second, third, and fourth lists. Each of these lists included 40 filler trials. Twenty of the filler trials presented participants with a fluent utterance naming a referent as hiding the treasure, and the remaining 20 presented an utterance with either a disfluency or a discourse manipulation (Table 7.1). In each of these sets of 20, 10 were presented with a video showing an adaptor gesture, and 10 were presented with a video showing no gesture (Table 7.2). Therefore, the fluency and gesture manipulations in the filler stimuli followed the same distribution as that of the critical set. The

specific pairings of videos to referents in filler trials were randomly assigned on each run of the experiment.

Bonus rounds

The aim of the bonus trials was to maintain motivation throughout the study. Twenty percent of filler trials were designated as ‘hidden bonus rounds’, offering more treasure than regular rounds. These trials were identical to regular trials, apart from a message informing participants that they had successfully located a bonus treasure chest. This message was presented immediately following their mouse click, regardless of whether they chose the referent or the distractor. On each run of the experiment, filler trials were randomly assigned as bonus rounds, with the constraint that an equal number of these trials presented all combinations of gesture vs. no gesture videos and fluent vs. disfluency/discourse manipulations utterances in the filler trials.

Procedure

The experiment was presented using OpenSesame version 3.2 (Mathôt et al., 2012) and stimuli were displayed on a 21 in. CRT monitor with a resolution of 1024×768 . The monitor was placed 850 mm from an Eyelink 1000 Tower-mounted eye-tracker which tracked eye movements at 500 Hz (right eye only). Audio was presented in stereo from speakers on either side of the monitor and sampled at 44100 Hz. Videos were presented at 25 frames per second, with mouse coordinates sampled at every frame (every 40 ms). Eye movements, mouse coordinates and object clicked (referent or distractor) were recorded for each trial.

Fig 7.1 represents a sample trial from the experiment. Participants underwent a manual drift correction between each trial to ensure accurate recording of

eye movements. The central fixation dot then turned red for 500 ms signifying the beginning of the trial. Two images of the referent and the distractor, each measuring 150×150 pixels, were presented centered vertically in the display with the center of each image positioned 15% of the screen-width in from the outside edge. The relative positions (left or right) of referents and distractors were randomly assigned for each trial with the constraint that referents occurred equally often on either side. After 2000 ms the video (236×336 pixels) was added to the display, centered horizontally with the bottom edge at the vertical midpoint. After 1400 ms from video onset, playback of the audio recording began and the mouse cursor was centered and made visible. The trial ended upon mouse click on either image or 5000 ms after the onset of the referent noun in the utterance (in the case of a time out, a message was displayed telling participants to click on subsequent objects faster).

With the exception of the bonus trials, which only occurred on filler trials, participants received no feedback after their mouse clicks on objects.

Participants were told that the audiovisual recordings were taken from a previous experiment involving two players, and in which one player was tasked with describing the location of hidden treasure, and attempted to mislead the other player into choosing the wrong location. To support this cover story, the instructions presented a photograph of two people purportedly taking part in this two player game (seated at a table facing each other with tablet computers in front of them).

Participants were told that the players in this previous experiment lied approximately half of the time. We tasked participants with clicking on the object behind which *they believed* the treasure to be hidden in each trial, thereby collecting the located treasure themselves. A high score table was shown at the beginning of the experiment, purporting to show players who had successfully located the most

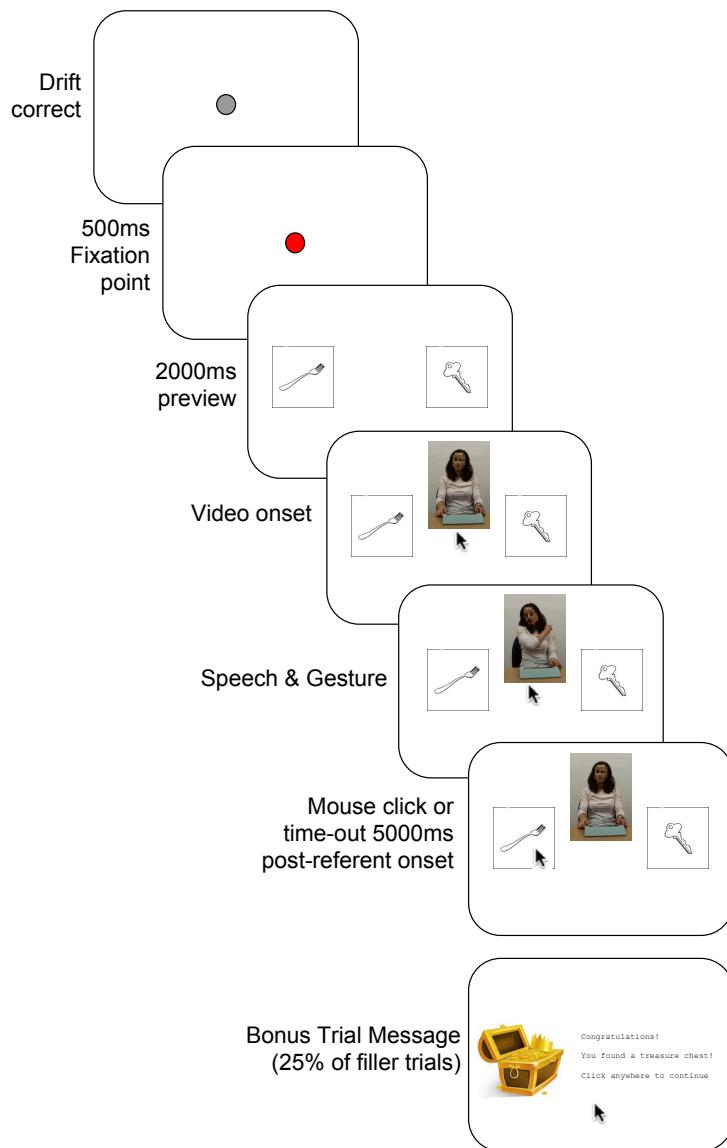


Figure 7.1: Procedure of a given trial in Experiment 7.1

treasure, and participants were informed that they would be able to enter their names on this table should they score highly enough.

The order of trials was randomly assigned on each run of the experiment. After the instructions, participants completed five practice trials. One of these included a disfluent utterance, one included a video of an adaptor gesture, and the remaining

three presented fluent utterances and no gesture. One of these three was presented as a bonus trial.

All participants received the same score putting them at the top of the scoreboard. After the treasure-hunt game (and after submitting their name on to the scoreboard), participants completed a short questionnaire which asked them if they noticed anything odd about the stimuli. Participants were then verbally questioned to establish whether it had occurred to them that audio and video had been scripted or artificially constructed. After being debriefed (and informed that audio and video stimuli were created separately and stitched together), participants were asked again (verbally) if they had noticed anything to that effect.

7.2.3 Results

Data from three participants were removed because they indicated in the questionnaire or during verbal questioning that they had held suspicions about the audiovisual stimuli during the eye-tracking task. This left data from 24 participants. Of the resulting 480 trials, 4 (0.8%) ended with no mouse click on either object and were removed from the analysis.

Analysis

Analysis was run in R version 3.5.2 (R Core Team, 2018), using the lme4 package version 1.1-17 (Bates et al., 2015).

Mouse clicks to the distractor were modelled using mixed effects logistic regression, with fixed effects of fluency (fluent vs. disfluent, deviation coded), gesture (no gesture vs. adaptor gesture, deviation coded) and their interaction. Random intercepts and slopes for fluency and gesture were included by-participant, along

with by-referent random intercepts. Pairwise comparisons of estimated marginal means with multivariate *t*-distribution adjustment (see Genz & Bretz, 1999) were conducted to investigate differences between combinations of (dis)fluency and gesturing on participants' judgements of deception.

The time taken for participants to click on an object (measured from referent noun onset) were log transformed and modelled using mixed effects linear regression with fixed effects of object clicked (referent vs. distractor, deviation coded), fluency (fluent vs. disfluent, deviation coded), gesture (no gesture vs. adaptor gesture, deviation coded) and all interactions. Random intercepts were included both by-participant and by-referent, along with by-participant random slopes of fluency, gesture, and object clicked.

Eye fixation data was averaged into bins of 20 ms (10 samples) prior to analysis. The proportion of time in each bin spent fixating either object (referent or distractor) was calculated. Mouse coordinates were sampled every 40 ms. Using only *X* coordinates, we calculated the number of screen pixels moved in each 40 ms sample and the direction of movement (towards either referent or distractor). From these, we calculated the cumulative distance travelled towards each object from referent noun onset up until that time as a proportion of the cumulative distance travelled in either direction. Any movements beyond the outside edge of either referent or distractor were considered to be 'overshooting' and were not included in calculations (2.1% of samples). The proportions of fixations and mouse movements to either object were empirical logit transformed (Barr, 2008), yielding measures for which positive and negative values indicate a bias towards the referent and distractor respectively, and a value of zero indicates no bias towards either object.

Eye and mouse movement data was analysed using two methods. Firstly, in the time window beginning at referent-noun onset and extending for 800 ms,

we conducted a growth curve analysis (Mirman et al., 2008) on the empirical logit transformed bias to the referent over the distractor. Models for both eye- and mouse-tracking data included fixed effects of fluency (fluent vs. disfluent, deviation coded), gesture (no gesture vs. adaptor gesture), time (Z-scored) and all interactions. No higher order polynomials of time were included in either eye- or mouse-tracking models, due to their inclusion failing to improve model fit (dependent upon both likelihood ratio test and a Bayesian information criterion decrease of ≥ 10 , following Raftery 1995). By-referent random intercepts were included, along with by-participant random intercepts and random slopes of fluency, gesture, and time. Following Baayen (2008), we considered effects in all models to be significant where $|t| > 2$.

The second analysis was conducted on the empirical logit transformed bias to the referent over the distractor on a point-by-point basis, with linear mixed effects models constructed for individual time bins. For the 2000 ms following referent noun onset, a model was run on each bin of 60 ms and 120 ms for eye and mouse movements respectively, with fixed effects of fluency, gesture and their interaction, and random intercepts by-participant and by-referent. Random slopes were not included due to problems with model convergence on several time bins. We note that this analysis involves an increase in family-wise error rate. Previous studies using this methodology have accounted for this by defining a minimum number of adjacent significant bins to indicate effects (see Borovsky, Elman, & Fernald, 2012), or—as we do here—combining it with modelling a longer time window to establish the presence of an effect (Ito, Pickering, & Corley, 2018), with bin-wise analyses indicating the point of divergence.

Object clicks and response times

Over the course of the experiment, participants clicked on the referent in 45% of critical trials, and on the distractor in 55%. Table 7.3 shows the numbers of mouse clicks on either object in critical trials, split by the presence and absence of speech disfluency and adaptor gesturing. A main effect of fluency ($\beta = 1.39$, SE = 0.43, $p = .001$) replicated the findings in Loy et al. (2017), and a main effect of gesture ($\beta = 1.92$, SE = 0.42, $p < .001$) replicated the results from Experiment 6.2. Both non-linguistic cues resulted in increased likelihood of participants judging the speaker to be dishonest (as indicated by increased clicks to the distractor). A significant interaction between fluency and gesture ($\beta = -1.27$, SE = 0.49, $p = .009$) suggests that the addition of multiple non-linguistic cues resulted in a weaker effect on listeners' judgements of deception than a single cue does relative to no cue. This is not surprising as any other result would entail trials with multiple cues approaching a ceiling effect. Model results for mouse clicks are shown in Table 7.4.

Table 7.3: Objects clicked and time taken to click in critical trials in Experiment 7.1: Clicks recorded on each object (referent or distractor) by presence of disfluency and adaptor gesture

	Clicks to Referent	Clicks to Distractor	Time (ms) from referent-noun onset
Fluent-No Gesture	95 (79%)	25 (21%)	1925
Fluent-Gesture	38 (32%)	81 (68%)	2106
Disfluent-No Gesture	51 (43%)	67 (57%)	2058
Disfluent-Gesture	31 (26%)	88 (74%)	1955

Pairwise comparison of estimated marginal means (see Table 7.5) revealed significant differences between the fluent-no gesture condition and all other conditions, and between the disfluent-no gesture condition and disfluent-gesture condition. These results suggest that while the addition of an adaptor gesture cue to a disfluent utterance increases the likelihood of it perceived as dishonest, the

Table 7.4: Model results for clicks to distractor over referent in Experiment 7.1

	β	SE	p
(Intercept)	0.323	(0.180)	.07
Disfluency	1.386	(0.432)	.001
Gesture	1.920	(0.424)	<.001
Disfluency × Gesture	-1.267	(0.486)	.009
Var(1—Participant)	0.345		
Var(Disfluency—Participant)	2.694		
Var(Gesture—Participant)	2.503		
Var(1—Referent)	0.038		
Total	476		
Participant	24		
Referent	20		

opposite (the addition of a disfluent cue to an utterance with adaptor gesturing) is not the case.

Table 7.5: Pairwise comparisons of estimated marginal means of object-clicks in Experiment 7.1

	Mean difference (Log odds)	SE	p [†]
Fluent-No Gesture : Disfluent-No Gesture	-2.02	0.49	<.001
Fluent-No Gesture : Fluent-Gesture	-2.55	0.49	<.001
Fluent-No Gesture : Disfluent-Gesture	-3.31	0.72	<.001
Disfluent-No Gesture : Fluent-Gesture	-0.53	0.47	.64
Disfluent-No Gesture : Disfluent-Gesture	-1.29	0.49	.040
Fluent-Gesture : Disfluent-Gesture	-0.75	0.50	.41

[†] Adjustment method: multivariate t

Analysis of the times taken to click the mouse revealed significant interactions between object clicked and both forms of non-linguistic cue (disfluency and adaptor gesture). Participants were quicker to click the mouse when they were clicking the distractor and a) the utterance contained a disfluency ($\beta = -0.14$, SE = 0.06, $t = -2.12$) or b) the video showed an adaptor gesture ($\beta = -0.19$, SE = 0.06, $t = -2.91$). Model results for times taken to click the mouse are shown in Table 7.6.

Table 7.6: Model results for times taken to click the mouse in Experiment 7.1

	β	SE	t
(Intercept)	7.559	(0.041)	186.01
Disfluency	-0.019	(0.037)	-0.51
Gesture	0.011	(0.033)	0.32
Clicked: Distractor	0.000	(0.036)	0.00
Disfluency × Gesture	-0.049	(0.061)	-0.81
Disfluency × Clicked: Distractor	-0.136	(0.064)	-2.12
Gesture × Clicked: Distractor	-0.187	(0.064)	-2.91
Disfluency × Gesture × Clicked: Distractor	0.077	(0.124)	0.62
Var(1—Participant)	0.034		
Var(Disfluency—Participant)	0.012		
Var(Gesture—Participant)	0.005		
Var(Clicked: Distractor—Participant)	0.007		
Var(1—Referent)	0.000		
Total	476		
Participant	24		
Referent	20		

Eye movements

Figure 7.2 shows the time course of fixations to referents, distractors and videos in critical trials for the 2000 ms from referent onset, split by fluency and gesturing. Bin-by-bin analyses of the empirical logit transformed bias to the referent over the distractor indicated a main effect of disfluency emerging at 480 ms after the referent-noun onset, a main effect of gesture at 720 ms, and an interaction between the two at 840 ms.

Analysis of the 800 ms period immediately following onset of the referent-noun in audio presentation revealed that participants tended to fixate on the referent over the distractor more as this window progressed (as indicated by a main effect of time: $\beta = 8.95$, SE = 1.98, $t = 4.51$). This increasing bias was reduced following both disfluency ($\beta = -11.81$, SE = 1.10, $t = -10.75$) and adaptor gesturing ($\beta = -5.26$, SE = 1.10, $t = -4.79$). Model results are shown in Table 7.7, and Figure 7.3 shows the empirical logit transformed fixation bias towards the referent

over the distractor during the relevant window of analysis, along with the fitted values from the model.

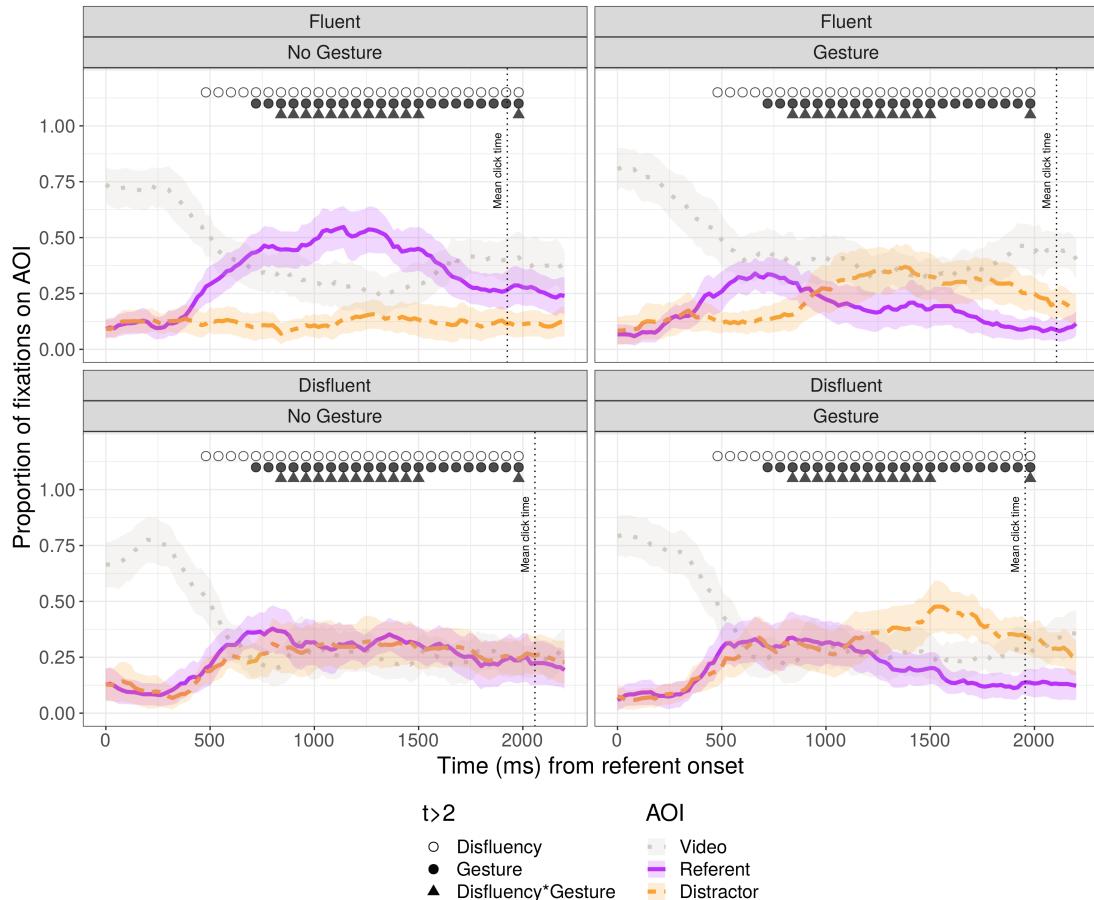


Figure 7.2: Eye-tracking results for critical trials in Experiment 7.1: Proportion of fixations to each object (referent or distractor) and the video, from 0 to 2000 ms post-referent onset, calculated out of the total sum of fixations for each 20 ms time bin. Split by presence of disfluency and adaptor gesture. Shaded areas represent 95% confidence intervals derived via bootstrapping subject data ($R=1000$). Significant effects ($t>2$) at each 60 ms bin are indicated at the top of each plot (N.B. effects compare difference in elogit-transformed proportion of fixations to referent over distractor between conditions).

Table 7.7: Model results for eye- and mouse-tracking analyses in Experiment 7.1 over the 800 ms window following referent-noun onset

	Fixations			Mouse Movements		
	β	SE	t	β	SE	t
(Intercept)	1.25	(0.26)	4.77	0.12	(0.08)	1.55
Disfluency	-1.61	(0.21)	-7.65	-0.13	(0.07)	-1.88
Gesture	-0.69	(0.19)	-3.64	-0.34	(0.10)	-3.50
Time	8.95	(1.98)	4.51	0.56	(0.25)	2.24
Disfluency \times Gesture	0.17	(0.24)	0.71	0.46	(0.09)	5.27
Disfluency \times Time	-11.81	(1.1)	-10.75	-0.82	(0.28)	-2.87
Gesture \times Time	-5.26	(1.1)	-4.79	-1.28	(0.28)	-4.49
Disfluency \times Gesture \times Time	0.24	(2.2)	0.11	2.90	(0.57)	5.09
Var(residual)	9.61			0.67		
Var(1—Participant)	1.49			0.1		
Var(Disfluency—Participant)	0.71			0.07		
Var(Gesture—Participant)	0.51			0.18		
Var(Time—Participant)	87.24			1.02		
Var(1—Referent)	0.06			0.03		
Total	19516			9992		
Participant	24			24		
Referent	20			20		

Mouse movements

Figure 7.4 shows the time course of mouse movements towards referents, distractors and videos in critical trials for the 2000 ms from referent onset, split by fluency and gesturing. Bin-by-bin analyses indicated a main effect of disfluency emerging at 720 ms after the referent-noun onset, a main effect of gesture at 240 ms, and an interaction at 720 ms.

Patterning with eye movements, mouse-tracking analysis over the same 800 ms window revealed that participants tended to move the mouse more towards the referent over the distractor as the window progressed $\beta = 0.56$, SE = 0.25, $t = 2.24$). This was reduced following both disfluency ($\beta = -0.82$, SE = 0.28,

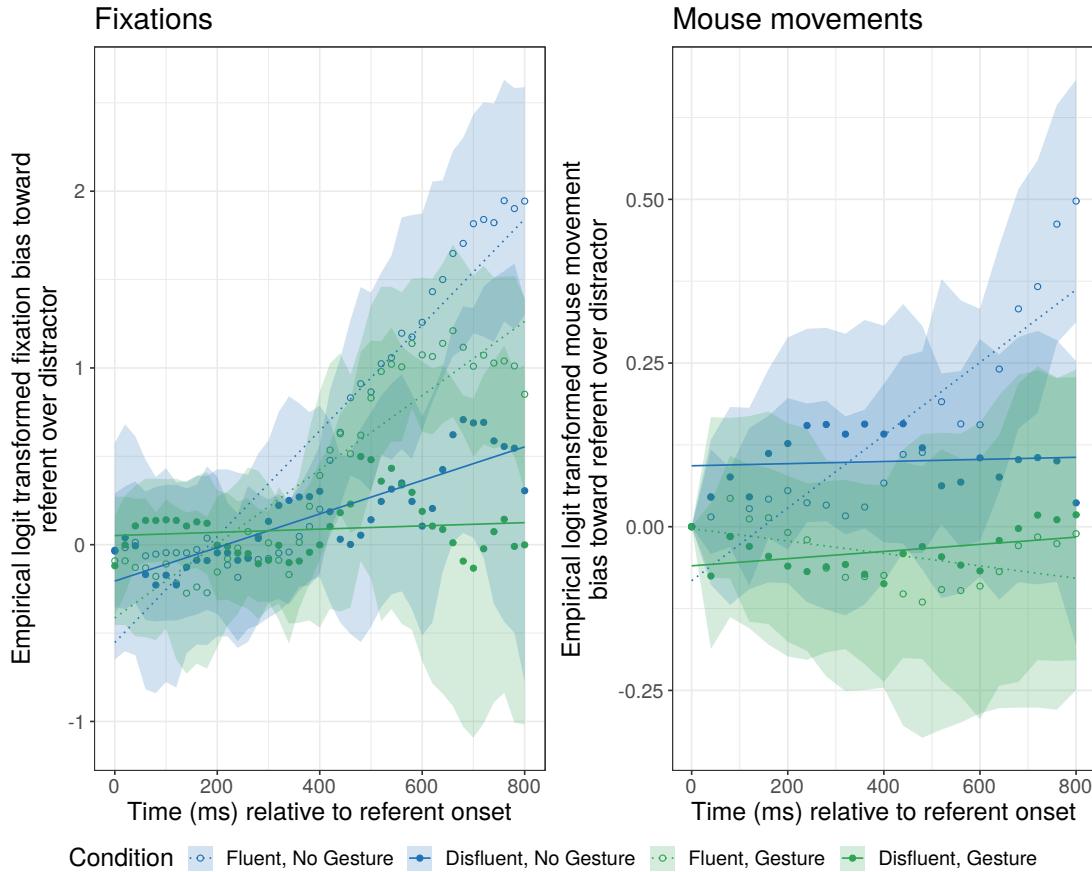


Figure 7.3: Empirical logit transformed fixation and mouse movement biases towards referent over the distractor in Experiment 7.1 for the 800 ms following referent-noun onset, by fluency and gesturing. Lines represent fitted values of the models.

$t = -2.87$) and adaptor gesturing ($\beta = -1.28$, SE = 0.28, $t = -4.49$). A three-way interaction between disfluency, gesture, and time suggests that the effect of disfluency on mouse movements over time was reduced when the video presented adaptor gesturing ($\beta = 2.90$, SE = 0.57, $t = 5.09$). Model results are shown in Table 7.7, and Figure 7.3 shows the empirical logit transformed movement bias towards the referent over the distractor during the relevant window of analysis, along with the fitted values from the model.

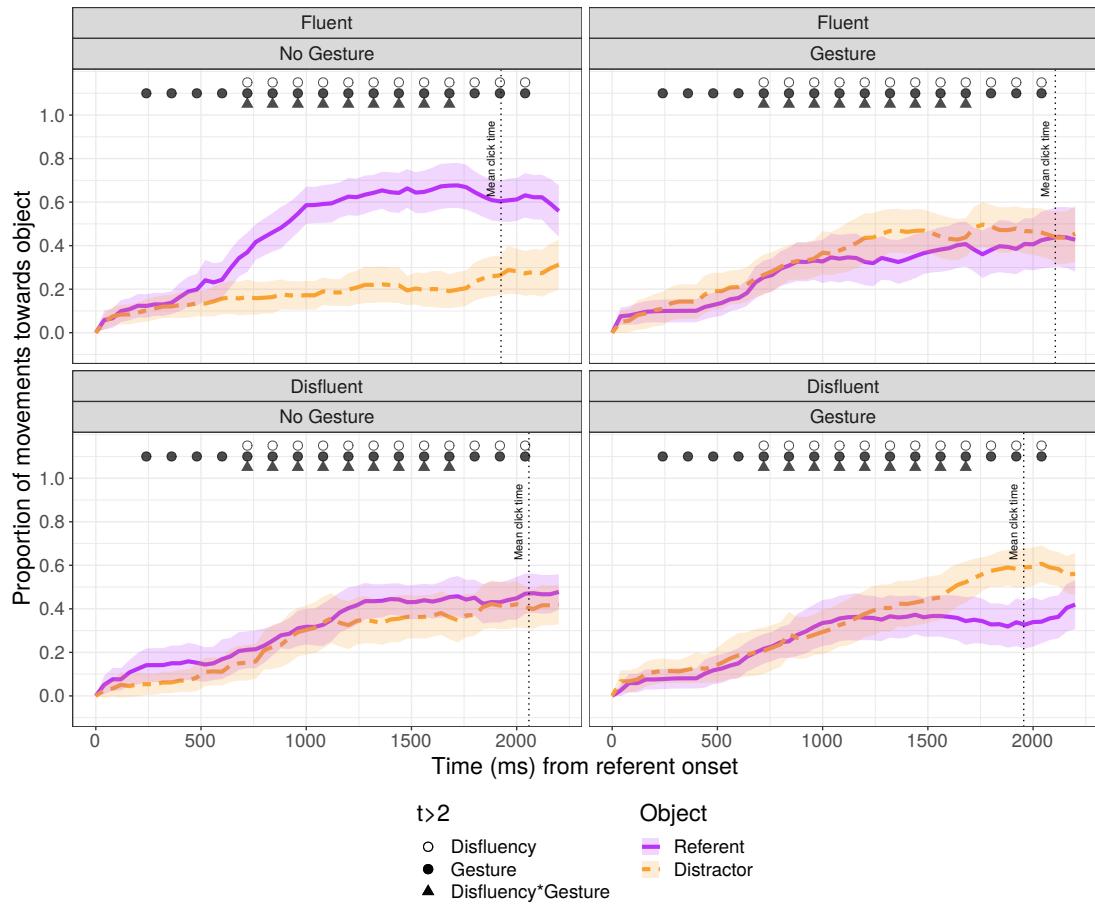


Figure 7.4: Mouse-tracking results for critical trials in Experiment 7.1: Proportion of cumulative distance travelled toward each object from 0 to 2000 ms post-referent onset. Proportions were calculated from the total cumulative distance participants moved the mouse until that time bin. Shaded areas represent 95% confidence intervals derived via bootstrapping subject data ($R=1000$). Significant effects ($t>2$) at each 120 ms bin are indicated at the top of each plot (N.B. effects compare difference in elogit-transformed proportion of mouse movements to referent over distractor between conditions).

7.2.4 Discussion

Experiment 7.1 investigated the influence of non-linguistic cues in both speech and gesture on listeners' judgements of deception. Participants saw and heard

a potentially dishonest speaker describing the location of some hidden treasure, and were tasked with guessing the true location of the treasure (thereby implicitly judging whether a given utterance was a truth or a lie). We manipulated the presentation of two different types of non-linguistic behaviours which have previously been associated with perceived deception (filled pauses in speech and adaptor gesturing) and recorded listeners' eye and mouse movements towards two possible locations of the treasure (the object named by the speaker as hiding the treasure and a distractor object).

As in previous studies investigating the impact of these cues individually on listeners' judgements of deception (Loy et al. 2017, and Chapter 6), both the presence of speech disfluency and of adaptor gesturing was associated with an increased probability of judging that the speaker was lying (more mouse clicks on the distractor object). Across the course of the experiment, participants showed a marginal tendency to click on the distractor over the referent, in contrast to previous research pointing to a prejudice towards interpreting something as true (see e.g., Barres & Johnson-Laird, 2003; McKinstry, Dale, & Spivey, 2008). This is not surprising given the distribution of non-linguistic behaviours in the experiment presented here, where only 25% of trials presented no potential cue to deception.

The present study shows that, in the absence of other evidence, people rely on potential signals in a speaker's behaviour (in both manner of spoken delivery and manner of non-verbal delivery) as markers of deception. In a context where a speaker exhibits both types of behaviour with roughly equal probability, people don't appear to latch on to one more than the other: Pairwise comparisons between conditions revealed no difference in the likelihood of participants judging the speaker as dishonest between trials in which disfluency or gesture were presented alone. Notably, however, utterances containing cues to deception in both modalities were associated with an increased likelihood of being judged as dishonest compared to utterances with only an audio cue (disfluency), but not compared to utterances

with only a visual cue (adaptor gesture). This suggests that the influence of non-linguistic behaviours on perceived deception has at least some degree of additivity.

The presence of either type of cue resulted in quicker mouse clicks on the distractor, suggesting that listeners' judgements of deception were assisted by the presence of these behaviours. Interestingly, the advantages of an available cue in terms of speed of judgement did not increase with the number of cues, but nor did it disappear, despite the potential increase in resources required in judgements based on multiple cues. One possibility is that in questioning the veracity of utterances such as those used in the present experiment, listeners rely on the first appearance of a deceptive behaviour to pre-emptively negate upcoming speech. It may be that the emergence of additional cues at later points in the utterance have little influence. This may also explain why, when adaptor gesturing was present, our fluency manipulation (occurring after the onset of gestures) did not increase the likelihood of judging an utterance as dishonest. Future research could investigate whether listeners' judgements differ when responding to, for instance, an utterance-initial disfluency with an utterance-medial occurrence of gesturing.

The influence of both disfluency and adaptor gesturing on final judgements of deception corresponded with early effects in the online measures: As in Loy et al. (2017) and Chapter 6, both potential cues to deception reduced the emerging bias to fixate, and move the mouse towards, the referent during the 800 ms post referent onset. Bin-by-bin analyses of the time course of fixations suggested that the effect of disfluency emerged at an earlier point than the effect of gesture. This is in keeping with earlier work (Experiment 6.2) which suggested that gestural cues might influence listeners' pragmatic judgements of deception along a different (and longer) time course than spoken cues. Mouse movements, however, appeared to contrast with this explanation, with an early effect of gesture evident 240 ms after referent-noun onset, almost 500 ms before an effect of disfluency. This might

be due to differences between the saccadic nature of eye movements and the continuous motion of trajectories measured in mouse movements (see, e.g., Spivey et al., 2005), or to specific sensitivities of mouse-movements to experimental design factors (see Kieslich, Schoemann, Grage, Hepp, & Scherbaum, 2019). The early effect of gesture on participants' mouse movements in the current study is even more surprising given that previous research has suggested that mouse movements tend to be launched slightly later than eye movements (e.g., Farmer, Anderson, & Spivey, 2007). We suggest that this early effect is likely due to some element of the study design—perhaps a response to on-screen movement—although it is unclear why this results in biasing on object over another.

The findings from this chapter inform the present thesis in several ways. Broadly, results show that listeners make use of multiple sources of non-linguistic information when they are available to them, supporting the view that communication is inherently multi-modal. Experiment 7.1 replicates previous findings that listeners perceive certain behaviours as indicators of deceptive intent, and shows that these perceptions are robust in contexts where a speaker produces different types of behaviours. This contrasts with one of our explanations for the lack of an association between trunk movements and deception in Experiment 6.1, which we suggested may have been weakened due to the variety of other cues used in the study (adaptor gestures, different static postures). It is possible that the presence of other behaviours *within* a modality interferes with the association of specific cues with lying, but not behaviours *across* modalities.

Our results confirm the speed with which manner of delivery (both of speech and of non-verbal behaviour) can influence listeners' pragmatic judgements about whether or not a speaker is being dishonest, thereby having direct consequences for comprehension. Patterning with previous research, both non-linguistic cues used in the present study were found to influence listeners' fixations and mouse movements alongside the unfolding linguistic input. While eye-tracking results

support an account of visual cues influencing deception judgements at a later point in the time-course, this was directly contrasted by the mouse movements.

When presented individually, neither one of speech disfluency nor adaptor gesturing were found to be more strongly associated with deception than the other. Future research could investigate how this may change depending on the relative distribution of particular cues: For example, an especially disfluent speaker who only gestures some of the time, or a particularly fidgety speaker who is only moderately disfluent. Additionally, the association of a given behaviour with deception may be dependent upon the availability of other explanations for that behaviour: Would the influence of a scratching gestures on deception judgements disappear if the speaker was wearing an itchy jumper? Would the effect of speech disfluency be different if it was known that the speaker had a fluency disorder (as has been shown for the disfluency-difficulty bias; Arnold et al., 2007, Experiment 2)? The next chapter explores this possibility by investigating whether the disfluency-deception bias (bias toward judging a disfluent utterance as dishonest) is attenuated when there is evidence that the speaker has been momentarily distracted (thereby providing an alternative explanation for disfluency).

Chapter 8

Competing Causes: Contextual Effects on Online Pragmatic Inferences of Deception¹

thus far, this thesis has focussed on non-linguistic behaviours as signals of speech planning difficulty (Part I), and signals of deception (Part II). Drawing on research which suggests that speech disfluencies result in listeners anticipating the more difficult to describe of two objects (see Arnold et al., 2007, 2004), Chapters 4 and 5 extended this to the gesture domain. Findings suggested that—at least for explicit predictions of upcoming referents—the presence (and duration of) of iconic gesturing biases listeners to predict more difficult-to-name shapes.

In Chapters 6 and 7, we have investigated the influence of non-linguistic behaviours

¹This chapter presents an extended version of a published paper (King, J. P., Loy, J. E., & Corley, M. (2018). Contextual Effects on Online Pragmatic Inferences of Deception. *Discourse Processes*, doi: 10.1080/0163853X.2017.1330041). There may be some superficial differences between the versions presented here and in the original paper, but content remains unchanged. Model results are detailed in Appendix B

on listeners' global interpretations of an utterance, via judgements of message truth. Two specific non-linguistic behaviours—speech disfluency and adaptor gesturing—have been shown to reliably be perceived as markers of deception, altering the time course of judgements about deception at the early stages of comprehension. Chapter 6 established that the presence of adaptor gesturing biases listeners towards interpreting an utterance as dishonest—much like the similar bias for disfluent utterances. The perception of both of these behaviours (disfluency and gesturing) as markers of deception were then found to hold in a context where the speaker produces both types of behaviour in Chapter 7. However, the previous experiments presented participants with no obvious explanation for why the speaker produced these behaviours other than being an epiphenomenon of the act of lying. Less is known about whether listeners' biases towards interpreting utterances accompanied by these non-linguistic behaviours as dishonest adjust to the relative likelihood of possible explanations of the behaviour.

Speakers may produce non-linguistic behaviours such as speech disfluency and adaptor gesturing for various reasons. As well as indicating deception, disfluency may indicate (as discussed in Part I of this thesis) difficulty in producing speech (e.g., Arnold et al., 2000; Cook et al., 2009). However, factors extraneous to the process of producing a specific utterance may also result in a speaker producing certain non-linguistic behaviours. For example, being temporarily interrupted and producing a pause, or an insect landing on a speaker's arm may lead them to produce a movement which is comparable in form to the 'nervous' gestures used in Chapter 6. Furthermore, there may be more general explanations for a behaviour: Individual speakers may have specific atypical reasons for producing behaviours such as a movement or fluency disorder.

The varied explanations for producing such behaviours differ inasmuch as they may provide a global explanation (i.e., increasing the probability of these behaviours in general), or a local one (providing an explanatory cause for a specific behaviour).

Evidence suggests that listeners are able to adjust their perception of speech disfluencies in accordance with the availability of global explanations for these behaviours. For instance, Arnold et al. (2007) found that the bias to anticipate reference to less familiar objects when presented with disfluent utterance was modulated by whether or not the listener believed the speaker to have object agnosia (likely to result in disfluency in naming both unfamiliar and familiar objects). Similarly, by using an artificial lexicon, Heller et al. (2015) found evidence suggesting that participants adapt the association between disfluency and difficult to describe objects to a given situation, with shapes with newly learned names perceived as entailing more production difficulty than unconventional unnamed shapes (thereby being perceived as more likely following disfluent speech).

A comparable sensitivity to contextual information about a given speaker has also been shown to affect the comprehension of gesture. In a study in which participants were presented with two speakers, one of whom gestured normally while the other also produced self-adaptive grooming movements, Obermeier, Kelly, and Gunter (2014) found that the benefit of an iconic gesture in disambiguating homonyms was attenuated when presented with utterances from the speaker who produced more irrelevant movements. Obermeier et al.'s findings suggest that listeners' sensitivity to speaker specific communication style influences the extent to which their gestures are integrated into comprehension, just as Arnold et al. (2007) show that knowledge about a particular speaker influences the extent to which their spoken disfluencies affect comprehension. It is worth noting that some effects of manner of delivery on comprehension appear to be robust to speaker-specific expectations. In two studies, Loy (Experiments 5.1 and 5.2, 2017) found that the association between disfluency and deception remained despite additional information that **1)** a particular speaker tends to be truthful and **2)** a particular speaker may experience more speech production difficulty (using native/non-native speakers as a proxy for language difficulty).

Whether or not listeners are sensitive to available *local* explanations for speakers' variations in manner of delivery is less well studied. Aside from one experiment (Arnold et al., 2007, Experiment 3) which found that potentially distracting sounds for a speaker (beeps and construction noises) had no effect on listeners' biases to fixate the more difficult to name of two objects following a disfluency, to our knowledge no other research has investigated whether listeners dynamically reason about what is the most likely cause of a specific behaviour during the unfolding of speech. Here, we adapt the 'treasure paradigm' used by Loy et al. (2017) to study the disfluency-deception bias, and include a plausible cause of speaker distraction (in the form of a passing car honking its horn). We focus on disfluency here due to the availability of a plausible extraneous cause (distraction), for which a comparable contextual explanation of non-verbal behaviour is less obvious (and more contrived; we refer back to the suggestion of a fly landing on a speaker's arm above!).

Experiment 8 builds on Arnold et al.'s Experiment 3, with the crucial difference that the effects of disfluency under investigation are not indicative of ephemeral predictions but of listeners' lasting interpretation of an utterance. With the disfluency-difficulty bias in Arnold et al.'s studies, misattributing disfluency to one cause over another becomes (to the listener) trivial once the speaker unambiguously names one object and not the other. The disfluency-deception bias, on the other hand, offers a context in which disfluency has direct consequences on the final interpretation of an utterance (i.e., as a true statement or a falsehood), meaning that behavioural effects associated with listeners modelling what is the likely cause of a given disfluency are likely to be easier to detect.

8.1 Experiment 8.1

8.1.1 Abstract

Where the veracity of a statement is in question, listeners tend to interpret disfluency as signalling dishonesty. Previous research in deception suggests that this results from a speaker model, linking lying to cognitive effort, and effort to disfluency. However, the disfluency-lying bias occurs very quickly: Might listeners instead simply heuristically associate disfluency with lying? To investigate this, we look at whether listeners' disfluency-lying biases are sensitive to context. Participants listened to a potentially dishonest speaker describe treasure as being behind a named object, while viewing scenes comprising the referent (the named object) and a distractor. Their task was to click on the treasure's suspected true location. In line with previous work, participants clicked on the distractor more following disfluent descriptions, and this effect corresponded to an early fixation bias, demonstrating the online nature of the pragmatic judgement. The present study, however, also manipulated the presence of an alternative, local cause of speaker disfluency: The speaker being momentarily distracted by a car-horn. When disfluency could be attributed to speaker distraction, participants initially fixated more on the referent, only later fixating on and selecting the distractor. These findings support the speaker modelling view, showing that listeners can take momentary contextual causes of disfluency into account.

8.1.2 Introduction

Everyday speech is for the most part spontaneous, and thus often disfluent, containing pauses, “um”s, “uh”s, repetitions, revisions, and mispronunciations. Excluding silent pauses, naturally occurring speech has a rate of approximately 6

to 10 disfluencies per 100 words (Bortfeld et al., 2001; Fox Tree, 1995). The disfluent nature of speech is just one of many variable aspects of *how* an utterance might be presented, and listeners must be able to cope with this variability in order to successfully understand a speaker.

Disfluencies in speech are not merely incidental. Speakers are more disfluent when utterance planning involves low-frequency words (Beattie, 1979), less-preferred syntactic structures (Cook et al., 2009), discourse-new expressions (Arnold et al., 2000), or a greater choice of expressive alternatives (Schachter et al., 1991). In this way, disfluencies provide non-linguistic ‘cues’ about the content of a speaker’s message. Research has shown that listeners can, and do, exploit these cues to make predictions about upcoming speech. For example, following a disfluency, they are more likely to predict the introduction of a new object into the discourse, as shown by visual world eye movements (Arnold et al., 2004), and less likely to have difficulty integrating an unpredictable word into its context, as indexed by a reduction in the N400 ERP component (Corley et al., 2007).

Evidence from a series of eye-tracking experiments suggests that predictions like these are sensitive to context. Arnold et al. (2007) asked participants to click on depictions of easy-to-name (ice-cream) or harder-to-name (abstract symbol) items in response to auditory instructions. When the instructions were disfluent, participants were more likely to fixate harder-to-name items before they heard the item name. Importantly, these fixation biases were modulated when participants were told that the speaker had object agnosia, and hence might be presumed to have difficulty naming easy-to-name items. The fact that a prediction that a hard-to-name item will follow a disfluency can be modulated by contextual information suggests that, on encountering a disfluency, participants are not merely making a stochastic prediction about what might be mentioned next. Instead, they may be actively modelling the speaker in order to account for the disfluency encountered and make situation-specific predictions.

However, the picture is far less clear when the cause of the disfluency is *local*, in the sense that it could be assumed to be the cause of a specific instance of disfluency, rather than of a heightened probability of disfluency in general. In Arnold et al.'s Experiment 3, for example, local causes (beeps and construction noises, assumed to distract the speaker momentarily) did not affect listeners' biases to fixate harder-to-name objects following disfluency. Moreover, several studies have shown that listeners do not seem especially sensitive to the nature of the disfluency: They have been shown to be affected by dog barks (Bailey & Ferreira, 2003) and sine waves (Corley & Hartsuiker, 2011) when they are substituted for filled-pause disfluencies. This sensitivity to non-linguistic interruptions sits poorly with the idea that the listener is modelling the speaker's production system, to anything greater than a superficial extent.

One reason that it is hard to conclude what is being modelled is that, in the studies outlined above, the effects of disfluency are ephemeral. Disfluency might affect what listeners think they are about to hear, but it has no lasting consequences at the message level: The fluent and disfluent versions of the utterances used mean the 'same thing'. For that reason, the consequences to the listener of mis-modelling the speaker are trivial, and the behavioural consequences of any such modelling relatively hard to detect. However, a parallel literature shows that in some circumstances, disfluency has pragmatic effects, in that it has direct consequences for the way a listener interprets an utterance.

For example, Brennan and Williams (1995) based their comprehension study on evidence that speakers use disfluency to manage difficulty in retrieving information (Smith & Clark, 1993). Participants were played recordings of answers to general knowledge questions which had been obtained during a production study. The answers were digitally edited and were sometimes preceded by either a silent pause or a filler. Listeners rated the answers as being less likely to be correct when the recorded answers were preceded by silence or fillers. In other words, their

interpretations of, rather than simply predictions concerning, the utterances they heard were directly affected by disfluency (see also Swerts & Krahmer, 2005). Listeners faced with disfluency had less confidence in the speaker's knowledge (a weaker "Feeling of Another's Knowing", or FOAK), and therefore revised their estimates concerning the factual correctness of what was being said.

As well as producing statements about which they have little confidence, speakers can easily utter propositions which they know to be false. This form of lying is often thought to be associated with cognitive effort. According to this view, the increased load involved in formulating and uttering a lie may lead speakers to provide verbal and non-verbal cues to deception, including disfluency (DePaulo et al., 2003; Zuckerman, Koestner, & Driver, 1981). Listeners' interpretations appear to reflect such a hypothesis: Zuckerman, Koestner, and Driver (1981) found hesitations in speech to be reliably associated with a perception of dishonesty, in both judgements made by speakers about themselves, and judgements made about another speaker.

In both FOAK and lying research, the proposed mechanism by which the interpretation of what is said is affected by disfluency is via *speaker modelling*: By reverse inference, disfluency is a symptom of cognitive difficulty, and cognitive difficulty is the consequence of limited knowledge (FOAK) or of inventing a situation (lying). In other words, to conclude that the speaker is lying requires reasoning about his or her cognitive state, in line with earlier claims by Arnold et al. (2007). However, listeners may in fact not reason in this way. Instead, they may heuristically associate certain aspects of spoken performance with uncertainty or lying, perhaps based on previous co-occurrence, or a superficial model of the speaker. This heuristic association would only be affected by very clear evidence that it wasn't relevant.

One reason for believing that the association between disfluency and lying is

heuristically calculated is evidence from Loy et al. (2017), which highlights the speed at which pragmatic interpretations are made. Loy et al.'s study was framed as a treasure hunting game. In each trial, listeners were asked to indicate which of two depicted objects they believed was concealing some treasure, by clicking on that object. Participants heard recorded utterances which indicated the location of the treasure, and which were either fluent or disfluent ("The treasure is behind [the]/[thee, uh] <referent>"). Participants were told that the speaker would be dishonest half of the time. The judgements which participants made about the speaker's honesty in each trial were implicitly measured by examining which of the two objects they clicked: Clicking on the named object corresponded to a judgement that the speaker was telling the truth, whereas a click on the other object meant that the speaker was thought to be lying. In line with previous research linking disfluency to deception (Zuckerman, Koestner, & Driver, 1981), participants were less likely to click on the named object following disfluent utterances (and instead, tended to click on the object which had not been mentioned). Importantly, eye-and mouse-tracking records showed that this effect emerged as soon as it became clear which of the two objects was being named: In other words, participants' pragmatic judgements were shown to be influenced by disfluency at the earliest detectable moment. If detailed speaker modelling is occurring, any inferences regarding the cause of a given disfluency would have to be made very fast.

Another reason for assuming that a heuristic is at play is that listeners' interpretations of disfluency may be inaccurate. Although listeners tend to associate disfluency with lying (Loy et al., 2017; Zuckerman, Koestner, & Driver, 1981), some evidence suggests that, in production, disfluency occurs more frequently during truth-telling than during deception (Arciuli, Mallard, & Villar, 2010; Arciuli, Villar, & Mallard, 2009; Benus, Enos, Hirschberg, & Shriberg, 2006). DePaulo et al. (1982) demonstrated a mismatch between disfluency as an actual and as a perceived cue to deception: The rates of filled pauses produced by speakers did

not differ during descriptions they made about people whom they liked or disliked from descriptions made when they were asked to pretend to feel the opposite way about them. However, when listening to the descriptions made by other participants, higher rates of filled pauses were associated with an interpretation that the speaker was being deceitful (see also Loy, Rohde, & Corley, 2016).

The evidence cited above suggests that, at least for the case of deception, the influence of speaker disfluency is fast, and not always accurate: Listeners appear to rely on a rule-of-thumb association between disfluency and lying. However, it is possible that any inaccuracy is actually the result of a more detailed attempt to model the possible causes of speaker disfluency. Lying is associated with cognitive effort, and cognitive effort is associated with disfluency, perhaps predicated not on experience as a listener but on introspection as a speaker: Speakers believe themselves to be more disfluent when lying (Zuckerman, Koestner, & Driver, 1981), and this belief extends to others' language production. Evidence for this conjecture would lie in whether listeners' assessments of speaker veracity were affected by specific circumstances that a heuristic would be unlikely to take into account. One such circumstance would be the availability of an alternative cause of a given disfluency, such as the speaker being momentarily distracted. If listeners are reasoning about the causes of speakers' disfluencies, and alternative causes of those disfluencies are readily available, then the association between disfluency and lying should be weakened.

In the current study, we build on Loy et al.'s (2017) treasure hunting game, using an auditory context that provides plausible causes of speaker distraction (and thus disfluency). Utterances are presented to listeners under the guise of having been recorded outdoors in a busy street, and low-level ambient noise is present behind every utterance. In the critical condition, disfluencies are immediately preceded by relatively loud noises (here, car-horns). If listeners rely upon a simple heuristic association between disfluency and deception, then the car-horn

should not influence the judgements they make about a speaker's honesty. If however listeners actively model a speaker by reasoning about the causes of specific disfluencies, the association between disfluency and deception should be weakened when the car-horn is present: Listeners might attribute disfluency to the speaker being momentarily distracted, rather than to the intention to lie.

8.1.3 Method

The experiment followed a 2 (fluent vs. disfluent) \times 2 (distraction absent vs. present) design. Participants took part in a visual world paradigm game, similar to that used by Loy et al. (2017), in which they guessed the location of some treasure based on utterances made by a potentially deceitful speaker, ostensibly recorded outdoors in a busy street. Half of the critical utterances were fluent, and half disfluent; in half of all critical cases, a car-horn was clearly audible immediately prior to the disfluency, or in the equivalent parts of fluent utterances. As in Loy et al. (2017), we measured eye and mouse movements, to study the time course of listeners' pragmatic judgements about the honesty of an utterance, as well as their final interpretation of each utterance (object clicked).

Materials

Visual stimuli consisting of 120 black and white line drawings, taken from Snodgrass and Vanderwart (1980), were presented to participants in pairs across 60 trials (20 experimental, 40 fillers). Each trial presented the *referent* (the object that the speaker identified as having the treasure behind for that trial), and a *distractor*, which was chosen at random without replacement from a set of 60 objects. To control for the effect of the bias towards interpreting disfluency to difficulty of description (Arnold et al., 2007), critical referents and distractors were matched

for familiarity ($F \geq 3.0$) and ease-of-naming ($H < 1.0$). Object pairings with the same phonetic onset were avoided.

Audio files were constructed such that the critical referents could be heard in four conditions varying by delivery (fluent vs. disfluent) and presence of distraction (absent vs. present). The disfluent variants were created by splicing a prolonged article followed by a filled pause (“Thee, uh”) into the fluent utterances, directly before the mention of the referent. This corresponds to the utterance-medial position used in Loy et al. (2017, Experiment 2): We considered it to be more believable that an environmental distraction might cause a disfluency once a speaker had initiated an utterance. Each referent was paired with a unique clip of ambient traffic and street noise, over which the recordings in their four variants were presented. To create a plausible cause of speaker distraction, a 520 ms car-horn sound effect was presented prior to the onset of the referent noun (1100 ms before noun onset for disfluent utterances, 600 ms for fluent utterances). All recordings, ambient noise and sound-effects were normalized and re-sampled to create 48 KHz, 16-bit, stereo Wav files. A sample set of materials is schematically represented in (2).

- (2a) **fluent, distraction absent:** The treasure is behind the <referent>.

(2b) **disfluent, distraction absent:** The treasure is behind thee, uh
<referent>.

(2c) **fluent, distraction present:** The treasure is behind the <referent>. horn

(2d) **disfluent, distraction present:** The treasure is behind thee, uh
<referent>. horn

The twenty critical referents were counterbalanced across four lists, each with 10 fluent and 10 disfluent utterances, and each containing 10 instances of the

car-horn (5 of which preceded disfluencies). Lists also contained 40 filler utterances, half of which were fluent, and half of which contained either a form of disfluency or a discourse manipulation (see Table 8.1). Additionally, 20 of these filler items (10 fluent, 10 disfluent or discourse manipulation) contained various novel noises that could be interpreted as distracting for a speaker (see Table 8.2). These filler distractions varied in position relative to the referent noun onset.

Table 8.1: Disfluencies and discourse manipulations in filler items.

Filler type	Manipulation	No. of Utterances	Example
Fluent	None	20	The treasure is behind the <referent>.
Disfluent	Prolongation	3	The treasure is behind <i>thee...</i> <referent>.
	Repetition	4	The treasure is behind <i>the-the</i> <referent>.
	Filled pause (utterance-initial)	3	<i>Umm..</i> The treasure is behind the <referent>.
Other	Discourse marker	5	<i>Okay,</i> the treasure is behind the <referent>.
	Modal	3	The treasure <i>could be</i> behind the <referent>.
	Combination	2	<i>Right,</i> the treasure <i>might be</i> behind the <referent>.

Cover story

Central to the design of the present experiment was the requirement that participants believe that the utterances were produced naturally and in a noisy

Table 8.2: Plausible causes of speaker distraction in filler items.

Noise	No. of Utterances
Vehicle horns (various)	7
Sirens (various)	3
Vehicles revving (various)	2
Car-stereo	2
Bicycle-bell	1
Bus doors opening	1
Footsteps	1
Loose drain cover	1
Man shouting	1
Dog barking	1

environment. Participants were told that the recordings were made by a participant of a previous experiment which was conducted on the side of a busy street. To reinforce the cover story, the initial explanation of the experiment included three videos which purported to be examples of the speaker producing the utterances. These were presented alongside the images that the speaker spoke about. The speaker's fluency and honesty were varied in these videos, as was the presence of a distraction that might have caused any disfluency.

To ensure that analysis could be run only on data from participants for whom the cover story held, a post-test questionnaire assessed whether participants believed that the utterances were produced outdoors, as claimed. Once they had been debriefed about the true nature of the experiment, participants were asked again whether it had occurred to them during the experiment that the recordings might not have been produced outside in the street.

Procedure

Stimuli were displayed on a 21 in. CRT monitor, placed 850 mm from an Eyelink 1000 Tower-mounted eye-tracker which tracked eye movements at 500 Hz (right eye only). Audio was presented in stereo from speakers on either side of the monitor. Mouse coordinates were sampled at 50 Hz. The experiment was presented using OpenSesame version 3.0 (Mathôt et al., 2012).

Participants were told they would see a series of pairs of objects, and that treasure was concealed behind one of the objects in each pair. For each trial, they would hear a speaker indicating the location of the treasure; but the speaker would be lying half of the time. Their task was to click on the object that they believed the treasure was behind, and thus accrue treasure over the course of the experiment.

Once participants had read the instructions, the eye-tracker was calibrated. Recalibration occurred between trials where necessary. Each trial began with a drift correction using a central fixation point, that changed from gray to red (for 500 ms) upon successful fixation. Following the red fixation point, two images (referent and distractor) were presented, horizontally to the left and right of the midpoint of the screen, and the ambient traffic audio began. Referents were presented equally often on each side. 1500 ms after the stimuli had appeared, a mouse pointer was made visible at the center of the screen, and the playback of the utterance began. Participants used the mouse to click on one of the two objects. Once this had happened, the stimuli disappeared and were replaced by a gray fixation dot, signifying the beginning of the next trial. Trials timed out after 5000 ms from utterance onset.

To maintain motivation throughout the study, participants were told that there were a number of “hidden bonus rounds” which offered more treasure. Following 25% of the filler trials, a “bonus round” message appeared before progressing to

the next trial. This informed participants that they had successfully located bonus treasure (regardless of the object chosen). Participants were also told that the top scorers would be able to enter their names on a high-score table, which was shown at the beginning of the experiment.

Participants completed five practice trials (one of which was presented as a bonus round) prior to the main experiment. Eye movements, mouse coordinates and object clicked (referent or distractor) were recorded for each experimental trial.

8.1.4 Results

Exclusion criteria

Thirty-seven participants took part in the experiment, for a planned design size of 24. Participants were recruited from the University of Edinburgh community, and participated in return for a payment of £4. Twelve participants were excluded on the basis that they indicated either in the post-test questionnaire (10) or verbally (2) that they did not believe the cover story. One further participant was excluded because they had previously taken part in similar study.

Analysis

Analysis was carried out in R version 3.3.0 (R Core Team, 2018), using the lme4 package (Bates et al., 2015). Trials in which participants did not click on either the referent or distractor (0.01%) were excluded from all analyses.

Analyses for both eye and mouse movements were conducted over a time window of 800 ms from the onset of the referent name, matching the analyses in Loy et al. (2017). This window exceeds the duration of the longest critical referent

name (776 ms) and is consistent with evidence that eye movements reflect the establishment of reference around 400-800 ms after noun onset (Eberhard, Spivey-Knowlton, Sedivy, & Tanenhaus, 1995). Eye fixation data was averaged into 20 ms bins (of 10 samples) prior to analysis. For each bin, we calculated the proportions of time spent fixating referent or the distractor, resulting in a measure of the proportions of fixations on either object over time.

The position of the mouse was sampled every 20 ms, corresponding to one bin of eye-tracking data. Using the X coordinates only, we calculated the number of screen pixels moved and the direction of movement (towards referent or distractor). The cumulative distance travelled towards each object was calculated for each bin, and divided by the total distance moved, regardless of direction. The resulting measure was the proportion of total distance travelled towards either object over time. Trials for which the total mouse distance travelled post referent-onset was less than one third of the distance from the screen center to the near edge of an object were excluded (0.03% of trials). Movements beyond the outer edge of either object were considered to be ‘overshooting’ and were not included in calculations (4% of samples).

We used an empirical logit transform to measure relative biases in eye and mouse movements (Barr, 2008). Eye movement biases were calculated from the proportions of referent to distractor fixations; mouse movement biases were calculated analogously. A value of zero in either measure indicates no bias towards either object, and positive and negative values indicate a bias towards the referent and distractor respectively. Linear mixed effects models of eye and mouse movements included fixed effects of time (Z -scored), delivery (fluent or disfluent) and speaker distraction (absent or present), and all interactions. Random intercepts and slopes for time, delivery and distraction were included by-participant and by-referent. Following Baayen (2008), we considered effects in these models to be significant where $|t| > 2$.

The object clicked (referent or distractor) was modelled using mixed effects logistic regression. This model included fixed effects of delivery (fluent or disfluent) and speaker distraction (absent or present), and all interactions, with random intercepts and slopes for delivery and distraction by-participant and by-referent.

Object click

Responses show the same overall tendency to interpret an utterance as truthful as was found by Loy et al. (2017), with 57% of trials resulting in a click on the referent and only 43% on the distractor. Table 8.3 shows the percentage of mouse-clicks on each object by condition. Analyses showed that participants were less likely to click on the referent following a disfluent utterance than a fluent one ($\beta = -2.24$; $SE = 0.67$; $p < .001$). This is in keeping with the literature (DePaulo et al., 2003; Zuckerman, Koestner, & Driver, 1981): Manner of delivery influences participants' global interpretations of the speaker's truthfulness. The presence of a plausible speaker distraction was not found to affect responses; neither was the interaction between delivery and distraction. The bias toward interpreting disfluency as a sign of dishonesty appeared to be explicit for 19 out of 24 participants, as indicated in the post-test questionnaire.

Table 8.3: Mouse clicks on each object by condition.

	Delivery	Disfluent	Disfluent	Fluent	Fluent
Speaker-distraction	Absent	Present	Absent	Present	
Distractor		62%	62%	23%	24%
Referent		38%	38%	77%	76%

Eye movements

Figures 8.1 and 8.2 show the time-courses of fixations to referents and distractors over 2000 ms from referent onset, for fluent and disfluent conditions respectively. Analyses were conducted over a time window from referent onset to 800 ms post onset. For fluent utterances, participants displayed an early fixation bias towards the referent, which increased over time ($\beta = 0.64$; $SE = 0.12$; $t = 5.44$). For disfluent utterances, the fixation bias towards the referent was greatly reduced ($\beta = -0.60$; $SE = 0.06$; $t = -10.62$), and a preference for the distractor over the referent emerged later on in the trial. When an alternative, local cause of disfluency (the car-horn) was present prior to a disfluency, the bias towards the distractor was significantly reduced ($\beta = 0.18$; $SE = 0.08$; $t = 2.20$). The presence of the car-horn was not found to have an effect on the tendency to fixate on the referent for fluent utterances ($t = -0.60$).²

Mouse movements

Figures 8.3 and 8.4 show the time-courses of proportionate mouse movements towards referents and distractors over 2000 ms from referent onset, for fluent and disfluent conditions respectively. Analyses of the window from referent onset to 800 ms post-onset patterned broadly with the eye-tracking data. When presented with a fluent utterance, participants' movements to the referent over the distractor increased over time ($\beta = 0.49$; $SE = 0.07$; $t = 7.47$), although this movement was a little slower following a car-horn ($\beta = -0.14$; $SE = 0.04$; $t = -3.59$). Without distraction, participants moved the mouse towards the distractor when the delivery was disfluent ($\beta = -0.64$; $SE = 0.04$; $t = -16.71$). When the car-horn was present

²Including participants who did not believe that the utterances were produced naturally and in a noisy environment did not change the pattern of results (disfluency-deception bias: $\beta = -0.51$; $SE = 0.05$; $t = -10.84$; effect of speaker-distraction: $\beta = 0.26$; $SE = 0.07$; $t = 3.90$).

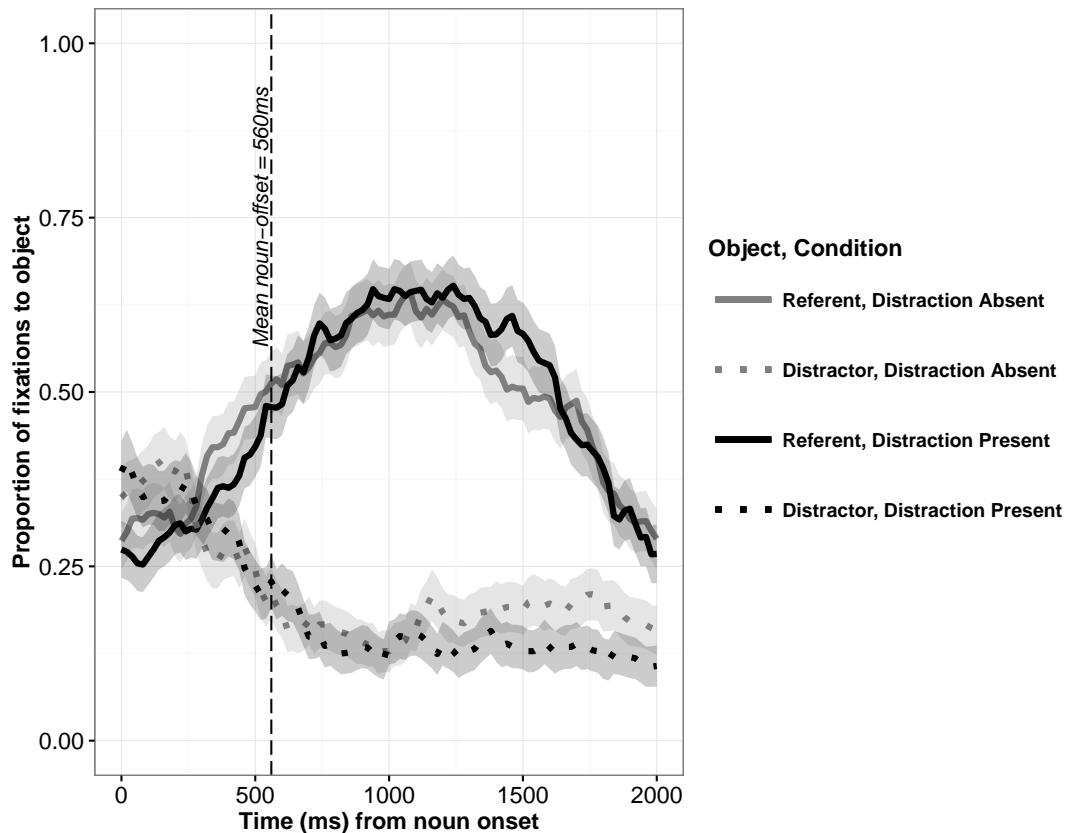


Figure 8.1: Mean proportion of fixations to either object (referent and distractor) for fluent utterances split by presence of speaker distraction, calculated out of the total sum of fixations for each 20 ms time bin from referent-onset to 2000 ms post-onset. Shaded areas represent ± 1 standard error of the mean.

prior to a disfluency, this tendency was greatly attenuated ($\beta = 0.37$; $SE = 0.05$; $t = 6.73$).

8.1.5 Discussion

Listeners' pragmatic judgements about a speaker's honesty were affected by manner of delivery. In keeping with the literature on deception perception, participants associated speaker disfluency with lying (DePaulo et al., 2003;

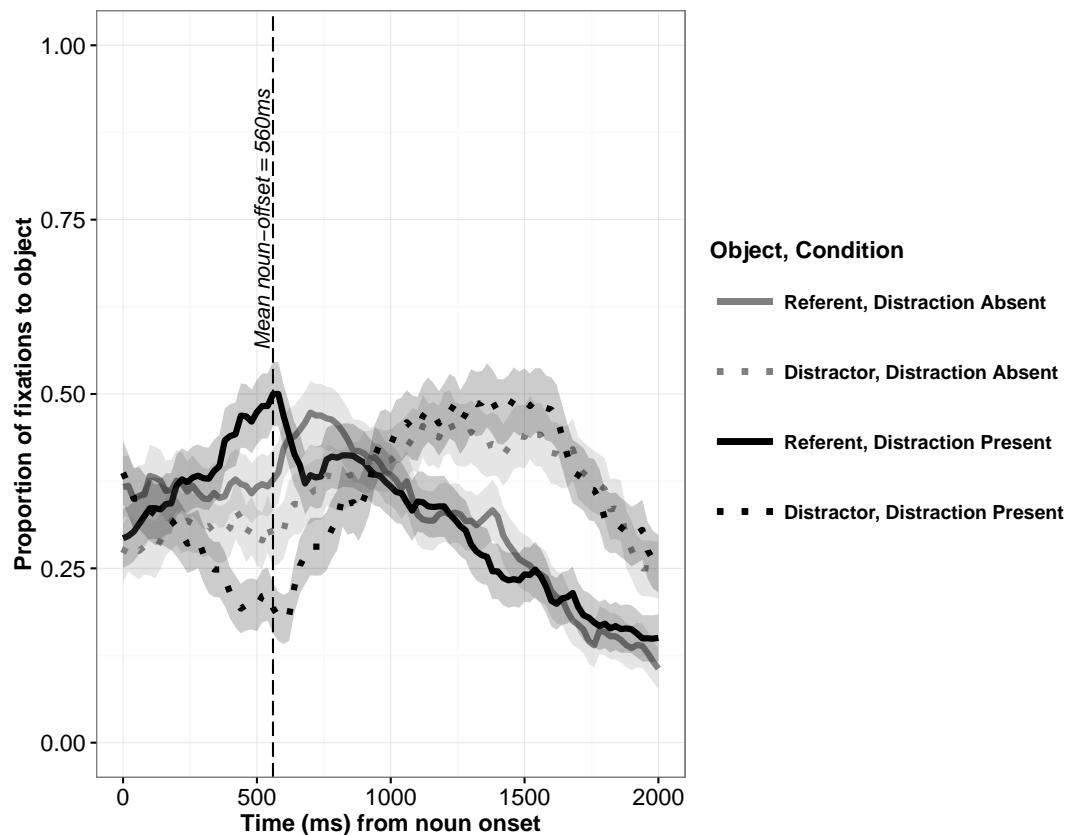


Figure 8.2: Mean proportion of fixations to either object (referent and distractor) for disfluent utterances split by presence of speaker distraction, calculated out of the total sum of fixations for each 20 ms time bin from referent-onset to 2000 ms post-onset. Shaded areas represent ± 1 standard error of the mean.

Zuckerman, Koestner, & Driver, 1981). As also shown by Loy et al. (2017), listeners made these judgements quickly. Both eye- and mouse-tracking evidence showed that biases emerged early, with listeners committed to a pragmatic interpretation of the speaker's honesty almost as quickly as the intended referent could be identified. These effects were shown to be robust against the presentation of speech in a noisy environment, in which there are potential distractions for the listener.

Importantly, listeners were not neutral with regard to the available distractions. Where a background noise (a car-horn) was a plausible cause of the speaker's

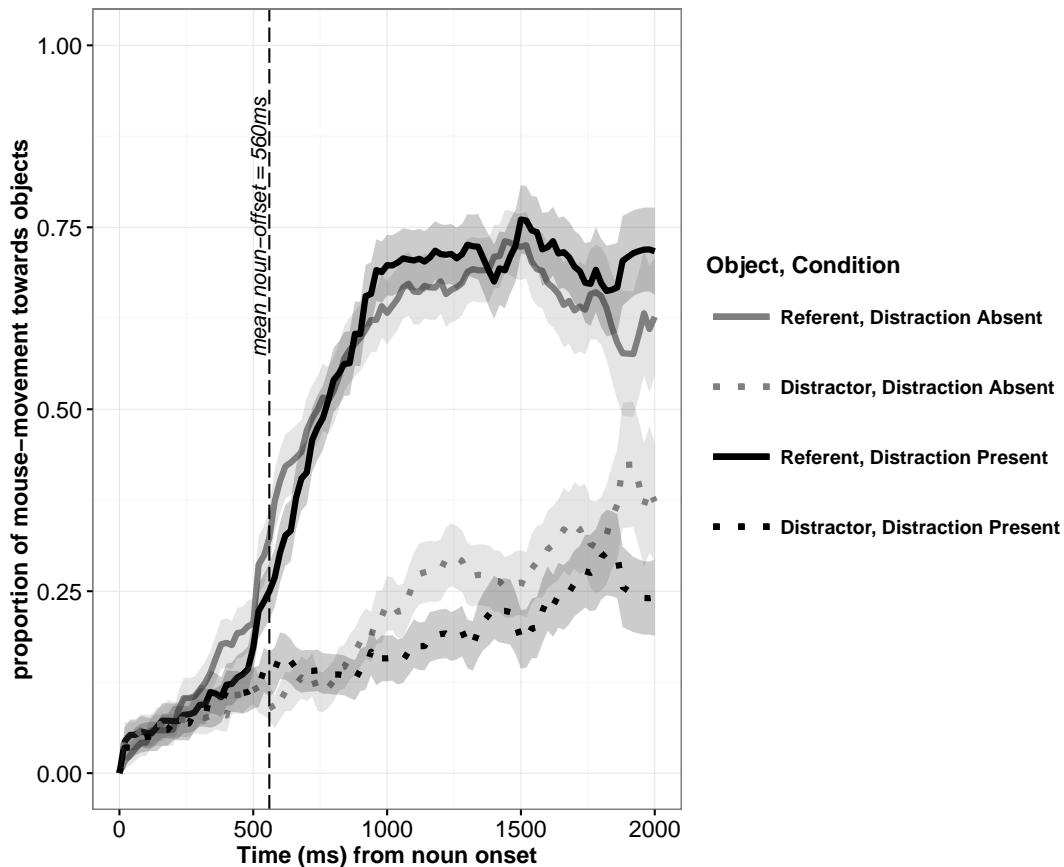


Figure 8.3: Mean proportion of cumulative distance travelled toward each object (referent or distractor) in fluent conditions split by presence of speaker distraction, from referent onset to 2000 ms post-onset. Proportions calculated out of total cumulative distance moved the mouse from referent-onset until that time bin. Shaded areas represent ± 1 standard error of the mean.

disfluency, participants showed an initial tendency both to fixate and to move the mouse pointer towards the referent, only later fixating on and eventually clicking on the distractor. Note that this finding suggests that listeners are sensitive to momentary changes to the context in which speech occurs. In this respect, it differs from Arnold et al.'s (2007) earlier finding that (constant) knowledge about the speaker can affect the ways in which listeners respond to disfluency. At face value, the finding may be taken to suggest that listeners in the present study are

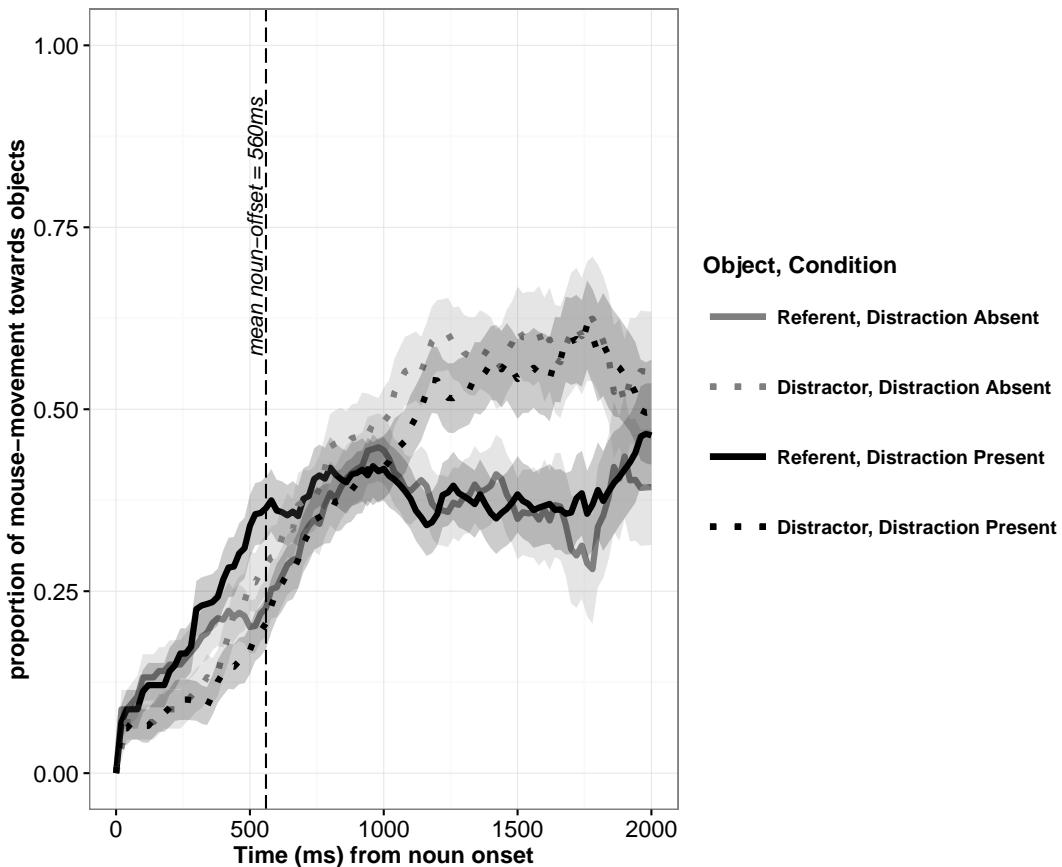


Figure 8.4: Mean proportion of cumulative distance travelled toward each object (referent or distractor) in disfluent conditions split by presence of speaker distraction, from referent onset to 2000 ms post-onset. Proportions calculated out of total cumulative distance moved the mouse from referent-onset until that time bin. Shaded areas represent ± 1 standard error of the mean.

modelling the speaker's production system in enough detail to be able to attribute a particular cause to a given disfluency.

There are, however, two potential alternative accounts of this finding. The first is that it was in fact the participants who were distracted by the car-horns, and that the findings reflect their initial lack of attention to the speaker's disfluency, rather than any attempt to model the cause of that disfluency. There is possible evidence for this in the fact that the car-horn was found to influence participants'

mouse movements during fluent utterances. With attention to any disfluency attenuated, there might be an initial bias to interpret utterances as honest. The account becomes difficult to sustain when we take the entire pattern of results into account, though, as the ‘unattended’ disfluencies clearly influence the eventual pragmatic interpretations of the speaker’s utterances.

The second alternative account is that participants’ pragmatic judgements relied on a heuristic association between disfluency and dishonesty. If the heuristic were to take into account any loud noises which preceded a disfluency, then participants might be expected to behave very much like the ones in the present experiment. This interpretation would leave us with two questions to answer, though. The first concerns the specificity of the heuristic. Might it be sensitive to car-horns, but not to dog barks, for example? Would it be contextually sensitive? Would listeners discern between the car-horn in the present experiment, which was contextually linked to the recording of the speaker, and a similarly loud car-horn which happened to sound ‘outside the testing room’? The second question is that of how the heuristic is created in the first place. One possibility is that it is trained on co-occurrences (that is, participants have previously observed car-horns to be associated with disfluency, and disfluency with lying). Unless any loud noise acted as a cause of disfluency, such a system would quickly run into a data scarcity problem (see Mitchell, Cuetos, Corley, & Brysbaert, 1995, for a similar argument concerning parsing). Another possibility is that the heuristic is based upon introspection, or the listener’s own understanding of what they would do as a speaker in given circumstances. To the extent that the latter is true, a heuristic is simply a form of speaker modelling (perhaps with differing implementational details).

The evidence therefore remains consistent with the view that listeners are able to reason dynamically about the most likely explanation of disfluency, and, as the speech unfolds, make attributions about why a particular speaker in a particular

context has been disfluent. This is in line with the speaker modelling account found in lying research, suggesting that listeners detect deception by reasoning about cues relating to the cognitive load of the speaker (DePaulo et al., 2003; Zuckerman, Koestner, & Driver, 1981). From this perspective, the findings here—that an alternative cause of disfluency modulates listeners' attributions of deception to disfluency—suggest that speaker modelling affects the early stages of comprehension.

Of note is the fact that Arnold et al. (2007, Experiment 3) did not find that distracting noises affected listeners' predictions about what speakers were likely to mention following a disfluency. There are a couple of possible reasons for this difference. Trivially, differences between experiments—the construction of utterances, for example—might simply mean that in the present experiment disfluency appeared more believably caused by distraction. Alternatively, it might be that the requirement to infer a pragmatic meaning in the lying paradigm renders listeners more likely to model the causes of disfluency. In particular, the treasure hunting game requires participants to reason about the speaker, and thus may encourage reasoning about the detail of the speaker's utterances. It may be that listeners can take context into account when it matters, but may not always do so, perhaps because for other effects of disfluency there are often no lasting consequences.

The fact that participants in this experiment are led to reason about the speaker might also go some way to explaining the overall bias to interpret disfluency as a cue to deception. Although the car-horn had a clear influence on the evaluation of disfluency, this effect was only temporary, as shown by participants' clicks on the referent or distractor objects. At the end of each utterance, listeners' interpretations were open to explicit reasoning; and initial interpretations appear to have been largely overridden. The fact that 19 out of 24 participants explicitly linked disfluency with deception in the post-test questionnaire supports this view,

and opens up the possibility that in a less game-like environment, the contribution of environmental factors to a speaker model would be larger.

The availability of an alternative, local cause of disfluency influenced the initial stages of participants' judgements about a speaker's honesty. The current study shows that, in situations which require reasoning about a speaker's honesty, listeners are sensitive to disfluencies and the context in which they occur. This sensitivity is shown simultaneously in eye movements and mouse movements, building on support for mouse-tracking as an alternative way of tracking cognitive processes (e.g., Farmer, Anderson, & Spivey, 2007). The findings are in line with suggestions in the deception literature that listeners associate disfluency with lying because of a speaker model which links lying to cognitive effort, and effort to disfluency. Moreover, they build on earlier work by Arnold et al. (2007), showing that, in cases where the pragmatic meaning of an utterance is at stake, listeners are able to take momentary contextual causes of disfluency into account. Above all, the present study emphasizes that understanding a speaker's pragmatic intentions is a contextually rich, and very fast, process.

8.2 Chapter discussion

The present chapter aimed to investigate listeners capacity for flexibly attributing a speaker's non-linguistic behaviours to contextual causes during the moment-to-moment processing of speech. Focussing on the association between speech disfluency and deception which listeners have been shown to hold during the real-time processing of speech (see Loy, 2017; Loy et al., 2017), we manipulated the presence of an alternative explanatory cause of disfluency in the form of a noise which could be perceived as potentially distracting for the speaker. Listeners' final judgements of whether or not a speaker was being deceptive revealed that

the association between disfluency and lying held despite the availability of an alternative cause for disfluency. However, following disfluencies for which an alternative explanation was present, participants displayed an initial tendency in the early stages of the comprehension process (within hundreds of milliseconds from the onset of the critical noun) to fixate on and move the mouse towards the referred to object (implicitly indicating a judgement of honesty) in comparison to disfluencies for which there was no obvious cause (other than being deceitful).

The robustness of the link between disfluent utterances and ultimate judgements of deception in Experiment 8.1 patterns with previous work from Loy. In two experiments, Loy manipulated speakers' accents (as a proxy for production difficulty, another cause of disfluency), and their perceived tendency to lie (Experiments 5.1 and 5.2, Loy 2017). Loy (2017) found that listeners developed speaker-specific expectations during a training phase, but that in the subsequent experimental phase the bias to interpret disfluent utterances as dishonest was not different depending upon which speaker produced the disfluency. These results suggests that the association between non-linguistic behaviours and deception can be so ingrained for listeners as to overwhelm other sources of information about deceptive intent and cause of disfluency.

This chapter has begun to explore some of the possible explanations of the processes by which listeners associate aspects of a speaker's manner of delivery as signals about their intention to deceive. Results suggest that effects of manner of spoken delivery on comprehension may be underpinned by flexible and rapid reasoning about the possible explanations for a specific manner in a given context. However, listeners' final interpretations of whether an utterance is true or not appear unchanged by competing explanations for a speaker's manner, with the association between disfluency and perceived deception being robust in the face of other causes of disfluency, be they specific to context (Experiment 8.1) or speaker (Experiment 5.2, Loy 2017). Comparable research is needed to investigate whether

the same is true of manner of non-verbal behaviour, but feasible contextual causes of a gesture may be harder to come by (e.g., swatting away a fly?).

Part III

Conclusions

Chapter 9

General discussion

In Chapters 3 to 8 of this thesis we reported a series of experiments examining the effects of communication of the non-linguistic behaviours produced alongside and within speech. We focussed on two contexts in which non-linguistic delivery might signal (or be perceived as signalling) information: When describing a referent in speech is more conceptually demanding, and when the speaker may be lying. We were especially interested in the time course over which non-linguistic behaviours might influence comprehension. In both contexts we used eye- and mouse-tracking paradigms in which participants responded (by selecting between two objects in a display) to recorded utterances (which in most experiments were presented in both audio and video). We manipulated the presence of different non-linguistic behaviours in the recordings, thereby allowing us to explore how those behaviours influenced listeners' comprehension alongside the unfolding of an utterance. In this chapter, we review our findings and discuss the broader implications of the results.

9.1 Signals of conceptual demand

Part I of the thesis was concerned with the extent to which non-linguistic behaviours can signal information about the upcoming message. In Experiment 3.1, we elicited descriptions which were either of easy-to-name shapes or of difficult-to-name ones. Directly measuring the relative durations of iconic gestures and spoken utterances, we found that descriptions of difficult-to-name shapes tended to involve greater durations of gesturing relative to speech. Additional analyses suggested that when describing shapes which were more difficult-to-name, the onset of gesturing tended to be earlier relative to the onset of the noun-phrase in speech.

These findings suggest that the relative durations, and possibly onsets, of speech and gesture vary with respect to the conceptual demands of describing a referent in speech. They support previous research suggesting that gesturing increases when information is more difficult to conceptualise (e.g., when a conceptualisation is not provided, Hostetter et al., 2007b). Taken together, these results suggest that the numbers, durations, and onsets of iconic gesturing signal information about upcoming speech—with more, longer, and earlier gestures signalling that the speaker may be finding a particular referent harder to conceptualise or to verbally encode (i.e., harder to package into speech).

This leads to our complementary question with respect to comprehension: Do listeners interpret these features of a speaker’s gestures as indicators of upcoming speech planning difficulty? This is a question about the extent to which the listener models the speaker’s production process, and is predicated on parallel research on another non-linguistic behaviour—speech disfluency. Previous studies have shown that increased cognitive load is associated with greater disfluency in speech (e.g., Arnold et al., 2000; Barr, 2001; Beattie, 1979), and research in comprehension has suggested that disfluency, in turn, influences listeners’ semantic predictions (Arnold et al., 2007, 2004; Barr & Seyfeddinipur, 2010; Corley et al., 2007).

Experiments 4.1 and 4.2 explored whether the same is true of non-linguistic behaviours in the gesture domain. We presented participants with the initial, ambiguous fragments of audio and video of a speaker instructing them to click on an object. Participants were tasked with clicking on whichever shape (out of an easy-to-name shape and difficult-to-name one) they thought the speaker was about to mention. Videos either showed the speaker producing gestures (iconic gestures in Experiment 4.1, adaptor gestures in Experiment 4.2) or sitting motionless. Results revealed that participants were more likely to choose the more difficult-to-name of two shapes when the partial instruction was accompanied by iconic gesturing. Additionally, this tendency was greater for fragments of iconic gesturing which were longer (and so had earlier onsets) relative to speech. These findings suggest that listeners interpret the presence of iconic gesturing (and either its relative duration or onset timing—we cannot say for sure) as a signal that a speaker is experiencing difficulty in formulating a spoken description. This effect was not found for adaptor gesturing (Experiment 4.2). Together, these results suggest that models need to distinguish qualitatively between the type of gestures that comprehenders attend to for reverse engineering speaker production difficulty, as well as quantitatively (in the relative durations of available cues).

9.1.1 The time course

Although Experiment 4.1 showed that iconic gesturing influenced listeners' explicit predictions about upcoming message content, we were unable to establish whether gesturing informed their anticipations in real-time. Analogous research into spoken delivery has found evidence that alongside the presentation of speech, disfluency biases listeners' predictions of upcoming referents (e.g., Arnold et al., 2007). In other words, even when an utterance is only temporarily ambiguous, listeners' expectations of upcoming content is influenced by fluency of speech:

When presented with an utterance of “click on thee - uh - red squiggle”, listeners are more likely to fixate the less familiar object (a squiggle as opposed to an ice cream cone) *before* the presentation of the critical noun (see Arnold et al., 2007).

In Experiment 5.1 we investigated whether it was possible to detect the influence of iconic gesturing on listeners’ eye and mouse movements across a similar, pre-disambiguation window. We presented participants with full instructions to click on an object, and tasked them with clicking on the correct shape as fast as they could (selected from a pair: one easy-to-name, one difficult-to-name). As in Experiment 4.1, the instructions were presented in both audio and video and we manipulated whether the video showed iconic gestures or no gesture. Findings were inconclusive, with the wider time course of participants’ fixations and mouse movements (along with responses in post-test questioning) suggesting that results may have been confounded by a task strategy. Furthermore, results highlighted some of the difficulties incurred when including a video component in a visual world paradigm: With presentation of gesture ongoing throughout the relevant window of analysis, effects on comprehension are made less clear due to the relative visual salience of the gestures.

Alternatively, it may be that in situations where the message will ultimately disambiguate between the two, drawing on gestures to inform fleeting predictions of upcoming content during the moment-to-moment processing of speech and gesture is needlessly demanding for the listener. The experiments of Chapter 4 indicate that listeners’ predictions of upcoming message content are sensitive to the type of gesturing (they are influenced by iconic gesturing, but not adaptor gesturing). This contrasts with studies which suggest that listeners lack sensitivity to the nature of speech disfluency (e.g., Bailey & Ferreira, 2003; Corley & Hartsuiker, 2011), suggesting that effects on comprehension may simply be responses to any form of interruption of the speech stream. It is possible that the demands on listeners’ resources required to distinguish qualitatively the type of gesturing

precludes it from influencing their on-line predictions of upcoming speech. Both of these explanations of the results of Experiment 5.1—a task effect, or a difference between modalities—are plausible, and future work is needed to better understand the influence of non-verbal behaviour on listeners' on-line predictions.

9.2 Markers of deception

Non-linguistic behaviours may also signal information about speakers' intentions. In Part II of the thesis we investigated the extent to which non-linguistic behaviours influence pragmatic understanding—listeners' inferences of what a speaker *means* by an utterance (as opposed to what they literally say).

The pragmatic effects of non-linguistic behaviours can have lasting consequences at the message level. This means that a speaker's non-linguistic behaviour may result in entirely different interpretations of an utterance, with listeners' eye and mouse movements reflecting the inferential processes leading to these interpretations. This contrasts with the influence of non-linguistic behaviours on listeners' predictions of semantic content (as in Part I), in which speech with and without gesture mean the 'same thing', and the behavioural consequences of gesture are relatively hard to detect. Specifically, we studied how non-linguistic behaviours influence listeners' judgements about whether a speaker is being honest or is telling a lie.

Although the validity of non-linguistic signals of actual deceit is less clear (see DePaulo et al., 1982; Hartwig & Bond, 2011), meta-analytic studies indicate that there are certain behaviours which listeners reliably interpret as cues to deception (Hartwig & Bond, 2011; Zuckerman, DePaulo, & Rosenthal, 1981). However, the time course during which non-linguistic behaviours influence listeners' pragmatic judgements of deception has received less attention.

We built on a study from Loy et al. (2017) which showed that listeners associate the manner of spoken delivery (specifically disfluency in speech) with perceived deception, and that this emerges from the early stages of comprehension. We investigated whether similar patterns extend to other modalities, with Experiments 6.1 and 6.2 establishing that listeners associate non-verbal delivery (specifically adaptor gesturing) with deception. We presented participants with audio and video of a potentially dishonest speaker describing the location of some hidden treasure, and manipulated the presence of non-verbal cues in the video component. Participants were tasked with guessing the true location of the treasure from two choices which indicated implicit judgements of whether a given utterance was a truth or a lie, and we recorded their eye and mouse movements throughout. Results revealed that, similar to manner of spoken delivery, manner of non-verbal delivery influences listeners' final judgements of deception, and does so alongside the lexical processing of speech.

In Chapters 7 and 8 we explored non-linguistic cues to deception further. Listeners' perceptions of speech disfluency and adaptor gestures as indicators of deceptive intent were found to be robust in contexts where a speaker was seen (and heard) to vary their delivery in both modalities. Results suggest that there may be differences in the speed and ease with which different non-linguistic behaviours affect judgements of deception, cues in the auditory modality possibly influencing the comprehension process at an earlier point than visual cues. Furthermore, results indicate that listeners may dynamically attribute a given non-linguistic behaviour to a specific cause in an on-line manner: When competing explanations for disfluency were present, the bias to interpret disfluent utterances as indicating deception was initially attenuated (although eventually sustained).

Future work could investigate whether listeners' judgements of deception depend on the relative proportions of different cues a speaker produces, or on the relative positions in an utterance that different cues occur (see, e.g., Loy et al., 2017,

for judgements of deception based on utterance-initial and utterance-medial disfluencies). Additionally, by developing a set of gestures which range in their perceived anxiety,¹ it would be possible to investigate whether listeners associate a given cue with lying based on how anxious they perceive the speaker to be.

9.3 Methodological considerations

Many of the experiments presented in this thesis used eye and mouse-tracking methodologies to study the influence of the visual behaviour of a speaker on listeners' comprehension. There are comparatively few studies which have attempted this (although see Saryazdi & Chambers, 2017; Silverman et al., 2010), and the experiments presented here shed light on some important aspects which may be of value to future work.

Primarily, the current thesis shows that it is possible to embed a video in the visual world paradigm and discern behavioural consequences of the comprehension process alongside unfolding audiovisual input. Notably, Experiment 7.1 found an effect of disfluency in listeners' eye-movements emerging at a similar point (roughly 500 ms after referent-noun onset) as was present in Loy et al. (2017), Experiment 2, which used the same audio recordings, but had no video.

However, results suggest that there may be limits as to how feasible this approach is for studying comprehension of iconic gestures, which may attract more visual attention due to the relevance of a particular gesture's trajectory and shape. In Experiment 5.1, results indicated that on-going iconic gesturing delayed listeners from fixating on—and moving the mouse towards—objects in the display.

Additionally, experiments in this thesis have provided a novel attempt at controlling

¹In Experiment 6.2 we used those gestures with the highest ratings of perceived anxiety

the amount of information presented in gesture stimuli by the construction of a gestural point-of-disambiguation (i.e., before which gestures were temporarily ambiguous). In creating this sort of stimuli it is difficult to maintain a natural style to the movements. We emphasise the need for comprehensive debrief questionnaires to ensure that data from participants who suspect the artificial nature of the stimuli can be excluded. Furthermore, the comparative stiffness and precision of these gestures (relative to, for example, those elicited in the production study) means that the generalisability of Experiments 4.1 and 5.1 is questionable. It may be that the visual world paradigm is more appropriate for studying non-iconic gestures, where the exact shape and trajectory matter less, and listeners are perhaps more able to take up non-verbal information via peripheral vision, rather than closely attending to the gesture.

Lastly, it is worth noting the discrepancies between listeners' eye- and mouse-movements in several of the experiments presented here. Some of these differences (Experiments 4.1, 4.2 and 5.1) we suggested may in part be explained by the relative visual salience of objects in the display, something to which eye movements are perhaps more sensitive than mouse movements. The results of Experiment 7.1 are harder to explain, and further research is needed to fully understand the various factors of study design affecting mouse trajectories—for instance, whether on-screen movement (e.g., in a video) influences participants' movements of the mouse.

9.4 Conclusions

When is meaning recovered from non-linguistic signals? Part I of this thesis goes some way to showing that iconic gestures, like speech disfluencies, can influence listeners' semantic predictions of upcoming message content by signalling

information that an object is more difficult to describe in speech. Increases in the rates and durations of iconic gesturing are associated with increased conceptual demand (Chapter 3), and listeners are, in certain contexts, sensitive to this association (Chapter 4). However, whether this meaning is recovered during listeners' real-time comprehension (like previous research suggests it is for speech disfluency, see Arnold et al. 2007) is less clear (Chapter 5). It may be that while interruptions of speech tend to reliably reflect the speech production process, a speaker's motor actions are simply more varied, making it harder for listeners to discern what type of movement is being produced in time to inform their predictions of upcoming semantic content.

In contrast, meaning recovered from non-linguistic behaviours also has the potential to change the overall interpretation of a message (e.g., indicating deceit, Part II). In these cases, listeners can make use of multiple sources of non-linguistic information when presented in both the auditory and visual modalities to infer a contextually relevant interpretation. Crucially, this information is shown to have direct consequences during real-time comprehension (Chapters 6 and 7), with the suggestion that listeners may dynamically reason about the causes of a given non-linguistic behaviour (Chapter 8).

Taken together, results from the studies presented here suggest that listeners' sensitivity to a speaker's non-verbal behaviour might depend on how consequential it would be to miss or ignore this source of non-linguistic information. In other words, the uptake of non-verbal information may be optional rather than obligatory. Experiments presented here found non-verbal information to influence listeners' comprehension when this information *mattered*—e.g., when it was the only available cue to what the speaker might be describing, or whether they might be being deceitful. In contrast, when listeners could simply wait for the lexical item (i.e., in Experiment 5.1), they did not appear to be influenced by the visible gesturing of the speaker. It should be noted that in Experiment 5.1 there were still potential

consequences of attending to non-verbal information in terms of response time, however, this account aligns with previous research finding that listeners are more likely to fixate a gesture when they expect it to be non-redundant (see Yeo & Alibali, 2017).

Overall, this thesis shows how both the spoken and non-verbal delivery of an utterance can have an important influence on comprehension. The recovery of meaning is not simply the process of combining the literal meanings of individual words, but involves the integration of contextual information from multiple channels to shape comprehension, highlighting the fact that communication is fundamentally multi-modal.

Appendix A

Model results for Experiments 6.1 and 6.2

Table A.1: Model results for clicks to referent over distractor in critical trials in Experiment 6.1

	β	SE	P
(Intercept)	0.55	(0.21)	.008
Trunk Movement Cue	-0.56	(0.32)	.08
Var(1—Participant)	0.35		
Var(Trunk Movement Cue—Participant)	0.85		
Cov(1 × Trunk Movement Cue—Participant)	-0.55		
Var(1—Referent)	0.04		
Var(Trunk Movement Cue—Referent)	0.29		
Cov(1 × Trunk Movement Cue—Referent)	-0.10		
Total	399		
Participant	20		
Referent	20		

Table A.2: Model results for times taken to click the mouse in critical trials in Experiment 6.1

	β	SE	t
(Intercept)	7.45	(0.05)	158.53
Trunk Movement Cue	-0.06	(0.03)	-1.68
Clicked Distractor	0.05	(0.03)	1.68
Var(residual)	0.08		
Var(1—Participant)	0.03		
Var(Trunk Movement Cue—Participant)	0.01		
Var(Clicked Distractor—Participant)	0.00		
Var(1—Referent)	0.00		
Var(Trunk Movement Cue—Referent)	0.00		
Total	399		
Participant	20		
Referent	20		

Table A.3: Model results for eye- and mouse-tracking analyses of critical trials Experiment 6.1 over the 800 ms window following referent-noun onset

	Fixations			Mouse Movements		
	β	SE	t	β	SE	t
(Intercept)	-0.38	(0.32)	-1.21	-0.09	(0.06)	-1.50
Time (Z-Scored)	2.23	(0.67)	3.34	0.30	(0.13)	2.41
Trunk Movement Cue	-0.18	(0.33)	-0.54	-0.05	(0.09)	-0.58
Trunk Movement Cue \times Time	0.34	(0.22)	1.54	0.03	(0.08)	0.37
Var(residual)	11.45			0.62		
Var(1—Participant)	0.91			0.04		
Var(Trunk Movement	0.91			0.07		
Cue—Participant)						
Var(Time—Participant)	4.57			0.17		
Var(1—Referent)	0.97		0.02			
Var(Trunk Movement	1.00			0.08		
Cue—Referent)						
Var(Time—Referent)	3.84			0.09		
Total	16359			7978		
Participant	20			20		
ref	20			20		

Table A.4: Model results for clicks to referent over distractor in filler trials in Experiment 6.1

	β	SE	p
(Intercept)	0.63	(0.16)	<.001
Adaptor Gesture	-1.03	(0.33)	.002
Different Posture	-0.73	(0.31)	.02
Var(1—Referent)	0.08		
Var(1—Participant)	0.20		
Var(Adaptor Gesture—Participant)	1.41		
Var(Different Posture—Participant)	1.17		
Total	797		
Referent	40		
Participant	20		

Table A.5: Model results for times taken to click the mouse in filler trials in Experiment 6.1

	β	SE	t
(Intercept)	7.46	(0.05)	160.36
Adaptor Gesture	-0.04	(0.04)	-1.04
Different Posture	0.01	(0.03)	0.27
Clicked Distractor	0.03	(0.03)	1.15
Var(residual)	0.08		
Var(1—Referent)	0.01		
Var(Adaptor Gesture—Referent)	0.01		
Var(Different Posture—Referent)	0.00		
Var(1—Participant)	0.03		
Var(Adaptor Gesture—Participant)	0.01		
Var(Different Posture—Participant)	0.00		
Var(Clicked Distractor—Participant)	0.01		
Total	797		
Referent	40		
Participant	20		

Table A.6: Model results for eye- and mouse-tracking analyses of filler trials in Experiment 6.1 over the 1100 ms window following referent-noun onset

	Fixations			Mouse Movements		
	β	SE	t	β	SE	t
(Intercept)	0.18	(0.24)	0.75	-0.10	(0.05)	-2.04
Time (Z-Scored)	1.05	(0.34)	3.13	0.40	(0.08)	4.75
Different Posture	0.24	(0.29)	0.84	0.02	(0.09)	0.28
Adaptor Gesture	-0.10	(0.32)	-0.31	0.13	(0.10)	1.25
Different Posture \times Time	-0.96	(0.13)	-7.42	-0.22	(0.05)	-4.58
Adaptor Gesture \times Time	-0.58	(0.13)	-4.47	-0.36	(0.05)	-7.50
Var(residual)	12.54			0.75		
Var(1—Referent)	1.02			0.06		
Var(Time—Referent)	2.29			0.15		
Var(Different Posture—Referent)	1.14			0.12		
Var(Adaptor Gesture—Referent)	1.52			0.16		
Var(1—Participant)	0.58			0.01		
Var(Time—Participant)	1.01			0.05		
Var(Different Posture—Participant)	0.94			0.07		
Var(Adaptor Gesture—Participant)	1.15			0.11		
Total	44632			21332		
Referent	40			40		
Participant	20			20		

Table A.7: Model results for clicks to referent over distractor in Experiment 6.2

	β	SE	p
(Intercept)	1.53	(0.23)	<.001
Adaptor Gesture	-2.78	(0.38)	<.001
Var(1—Participant)	0.27		
Var(Adaptor Gesture—Participant)	1.01		
Var(1—Referent)	0.00		
Var(Adaptor Gesture—Referent)	0.33		
Total	397		
Participant	20		
Referent	20		

Table A.8: Model results for times taken to click the mouse in Experiment 6.2

	β	SE	t
(Intercept)	7.43	(0.05)	159.94
Adaptor Gesture	-0.03	(0.04)	-0.66
Clicked Distractor	0.08	(0.04)	1.93
Var(residual)	0.10		
Var(1—Participant)	0.03		
Var(Adaptor Gesture—Participant)	0.00		
Var(Clicked Distractor—Participant)	0.00		
Var(1—Referent)	0.00		
Var(Adaptor Gesture—Referent)	0.00		
Total	397		
Participant	20		
Referent	20		

Table A.9: Model results for eye- and mouse-tracking analyses in Experiment 6.2 over the 800 ms window following referent-noun onset

	Fixations			Mouse Movements		
	β	SE	t	β	SE	t
(Intercept)	-0.65	(0.31)	-2.10	-0.15	(0.08)	-1.84
Time	2.99	(0.67)	4.48	0.61	(0.10)	5.88
Adaptor Gesture	0.74	(0.32)	2.33	0.22	(0.14)	1.58
Adaptor Gesture \times Time	-2.94	(0.21)	-14.27	-0.78	(0.08)	-9.27
Var(residual)	9.59			0.86		
Var(1—Participant)	0.86			0.06		
Var(Adaptor Gesture—Participant)	0.89			0.17		
Var(Time—Participant)	4.11			0.12		
Var(1—Referent)	0.94			0.05		
Var(Adaptor Gesture—Referent)	0.96			0.17		
Var(Time—Referent)	4.38			0.02		
Total	16277			8304		
Participant	20			20		
Referent	20			20		

Appendix B

Model results for Experiment 8.1

Table B.1: Model results for clicks to referent over distractor in Experiment 8.1

	β	SE	p
(Intercept)	1.58	(0.39)	<.001
Disfluency	-2.24	(0.67)	<.001
Distraction	-0.07	(0.45)	.87
Disfluency \times Distraction	0.12	(0.57)	.83
Var(1—Participant)	1.63		
Var(Disfluency—Participant)	6.92		
Var(Distraction—Participant)	0.47		
Var(Disfluency \times Distraction—Participant)	0.03		
Var(1—Referent)	0.06		
Var(Disfluency—Referent)	0.20		
Var(Distraction—Referent)	0.14		
Var(Disfluency \times Distraction—Referent)	0.54		
Total	476		
Participant	24		
Referent	20		

Table B.2: Model results for eye- and mouse-tracking analyses in Experiment 8.1 over the 800 ms window following referent-noun onset

	Fixations			Mouse Movements		
	β	SE	t	β	SE	t
(Intercept)	-0.27	(0.15)	-1.84	-0.20	(0.07)	-2.81
Time	0.64	(0.12)	5.44	0.49	(0.07)	7.47
Disfluency	0.37	(0.14)	2.69	0.36	(0.10)	3.57
Distraction	-0.09	(0.12)	-0.77	0.05	(0.09)	0.54
Time \times Disfluency	-0.60	(0.06)	-10.62	-0.64	(0.04)	-16.71
Time \times Distraction	0.03	(0.06)	0.60	-0.14	(0.04)	-3.59
Disfluency \times Distraction	0.03	(0.08)	0.37	-0.11	(0.05)	-1.95
Time \times Disfluency \times Distraction	0.18	(0.08)	2.19	0.37	(0.05)	6.73
Distraction						
Var(Residual)	1.79			0.80		
Var(1—Participant)	0.20			0.03		
Var(Time—Participant)	0.14			0.03		
Var(Disfluency—Participant)	0.22			0.07		
Var(Distraction—Participant)	0.19			0.09		
Var(1—Referent)	0.24			0.06		
Var(Time—Referent)	0.13			0.05		
Var(Disfluency—Referent)	0.13			0.11		
Var(Distraction—Referent)	0.08			0.06		
Total	18564			17856		
Participant	24			24		
Referent	20			20		

Appendix C

Replication of Kelly et al. (2010)

C.1 Experiment C.1

Experiment C.1 presents a replication of Kelly et al.'s (2010) investigation into the comprehension of semantic mismatches in speech and gesture. Participants are tasked with responding as quickly and as accurately as possible, indicating whether either modality in a gesture-word pair (spoken verb alongside pantomime gesture) matched with a previously seen action (someone dialling a phone). A further manipulation varies the level of this semantic incongruity, with content mismatches either weak (*type* to *dial*) or strong (*knock* to *dial*). Measuring participants' responses (collected via keypress) and response times, the original study (Kelly et al., 2010) found that, in comparison to congruent speech and gesture, participants were both slower and produced more errors when either modality mismatched the action. Additionally, the rates of incorrect responses in Kelly et al. (2010) increased in line with the level of semantic incongruity of a modality.

C.1.1 Method

Twenty-three self-reported native English speakers were recruited from the University of Edinburgh community, and took part in the experiment in return for £4. Consent was obtained in accordance with the University of Edinburgh's Psychology Research Ethics Committee guidelines (ref number: 163-1516/1) Sixteen participants were right-handed, and seven were left-handed.

The stimuli were identical to those used in Kelly et al. (2010), and comprised a 1000 ms video of an action being performed followed by a black screen for 500 ms, followed by a 1000 ms video of a pantomime gesture accompanied by a recording of a spoken word. There were sixteen different experimental items (videos of actions). Each of these was presented in nine trials across the experiment — in five different experimental conditions, and in four filler trials — resulting in a total of 144 trials (80 experimental, 64 fillers). The experimental conditions consisted of a baseline condition (where both speech and gesture matched the action in the video) and four conditions where one modality (speech or gesture) mismatched either weakly or strongly with the action in the video, and the other modality matched the action. Examples of stimuli can be see in the sample timeline of procedure of two trials in Figure C.1. In filler trials, neither speech nor gesture matched the action seen in the video.

The experiment was presented using OpenSesame version 2.9 (Mathôt et al., 2012). Stimuli were displayed on a 21 in. CRT monitor with a resolution of 1024×768 , placed 850 mm the edge of the table. Audio was presented in stereo, and sampled at 48000 Hz. Videos were played at 30 frames per second, and measured 720×480 , positioned centrally on a black screen.

Participants were tasked with responding via keypress whether they thought that *either* speech or gesture matched the action in the video. Participants were

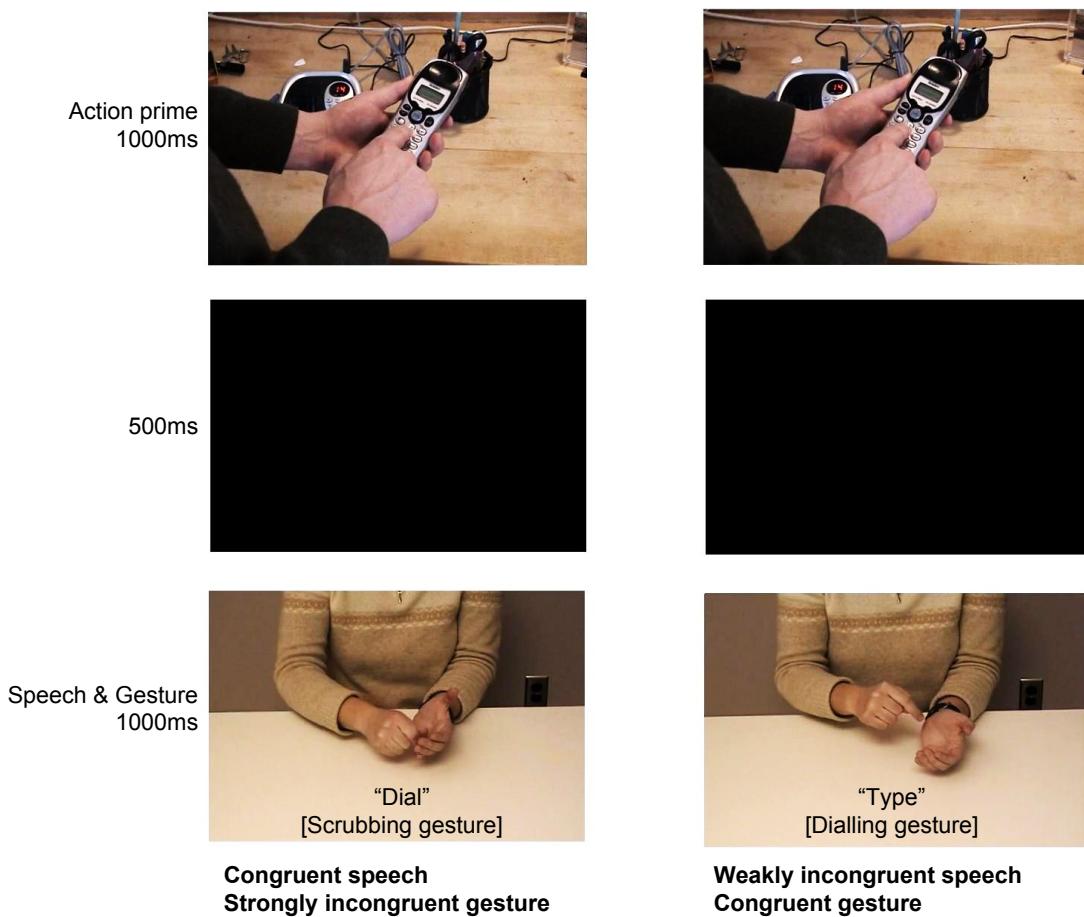


Figure C.1: Example procedure for two trials in Experiment C.1

asked to place their index fingers of each hand on the response keys ('m' and 'z'). Proceeding between trials required participants to press the space bar, and they were advised to do so without moving their index fingers (i.e. by using a thumb). Response keys and instructions were dependent upon handedness, such that affirmative responses were on the participants dominant side. Instructions encouraged participants to respond quickly and accurately.

Following the instructions, participants completed six practice trials. These comprised two in which speech and gesture both mismatched the action in the video, and four in which one modality (two speech, two gesture) strongly mismatched the action, and the other modality matched.

C.1.2 Analysis

Of the 1840 experimental trials, 21 were excluded due to either no keypress or an invalid response. A further 29 trials resulted in response times greater than three standard deviations above the mean and were also excluded, leaving a total of 1790 trials.

Following Kelly et al. (2010), two analyses were conducted, the first between congruent and incongruent trials, and the second comparing the modality and strength of semantic incongruence between the four incongruent conditions. Log transformed reaction times were modelled using mixed effects linear models. Incorrect keypresses were modelled using mixed effects logistic regression. In the first analysis, models included fixed effects of speech-gesture congruence (congruent vs. incongruent, deviation coded), and random intercepts and effects of congruence both by-item and by-participant. In the second, models included fixed effects of the modality in which the incongruence was presented (speech vs. gesture, deviation coded), the strength of the incongruence (weak vs. strong, deviation coded) and their interaction. By-participant and by-item random intercepts and effects of modality, strength and their interaction were included. We considered effects in these models to be significant where $|t| > 2$ (see Baayen, 2008).

C.1.3 Results

Figures C.2 and C.3 show the reaction times and error rates respectively, split by each of the experimental conditions.

Successfully replicating the effects found in Kelly et al. (2010), relative to matching speech-gesture pairs, trials in which one modality presented semantically incongruent material resulted in slower responses ($\beta = 0.09$, SE = 0.01, $t = 7.43$),

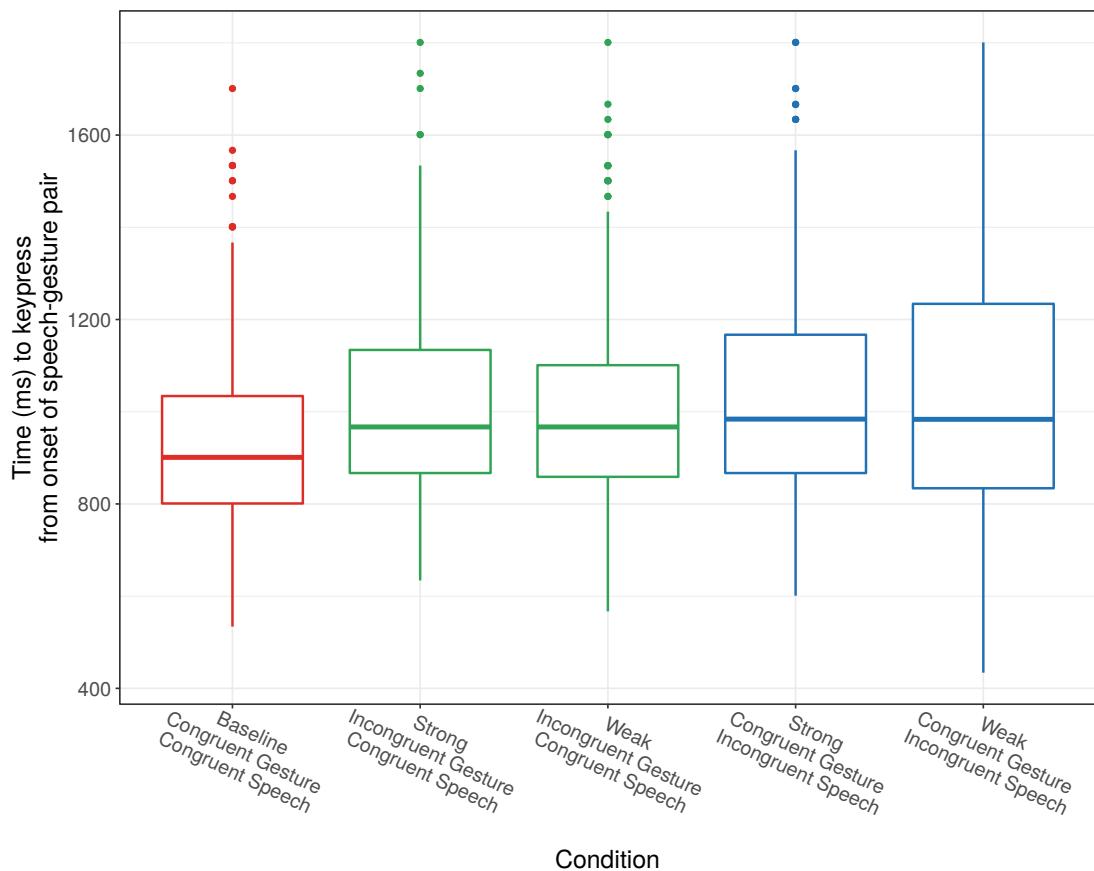


Figure C.2: times measured from the onset of the target speech-gesture pair, split by experimental condition.

and participants also produced more errors in these trials ($\beta = -1.65$, $SE = 0.39$, $p < 0.001$).

Consistent with Kelly et al. (2010), participants' reaction times were not influenced by the level of semantic incongruence presented, nor was there an interaction effect between the level of incongruence and the modality in which it was presented. In contrast to Kelly et al. (2010), in which there was a main effect of the target modality on reaction times (congruent speech with incongruent gestures elicited faster responses than congruent gestures with incongruent speech), the present study found no effect of which modality was (in)congruent on participants' reaction times.

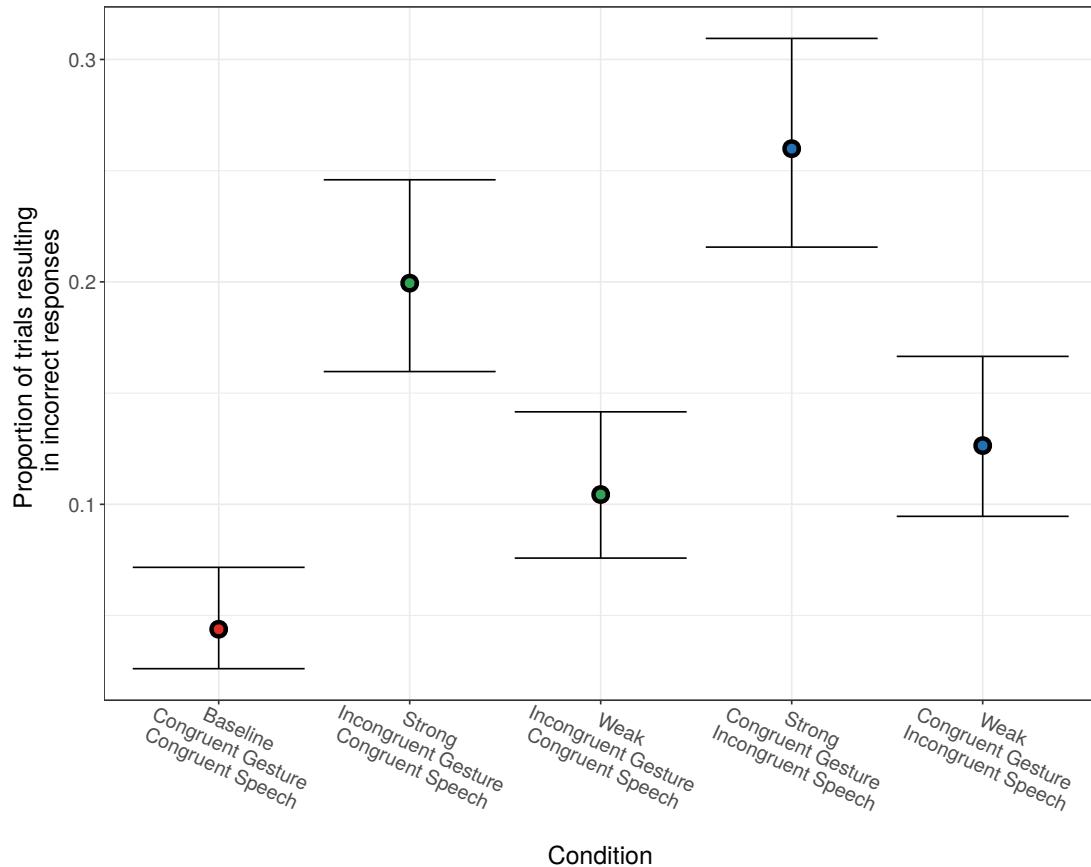


Figure C.3: Proportion of trials in each experimental condition which elicited an incorrect response.

As in Kelly et al. (2010), the rate of incorrect responses was greater when the incongruent modality was strongly incongruent than when it was only weakly incongruent ($\beta = 1.11$, $SE = 0.26$, $p < 0.001$), and there was no interaction of this effect with the modality in which the incongruity was presented. There was, however, a significant main effect of the modality, with trials presenting incongruity in speech eliciting a greater number of errors than those presenting incongruity in gesture ($\beta = 0.81$, $SE = 0.35$, $p = 0.02$).

C.1.4 Discussion

Experiment C.1 successfully replicates the main findings from Kelly et al. (2010): when tasked with determining whether either speech or gesture matched a previously seen action, relative to when both modalities matched, when one modality presented incongruent information participants were slower to respond and more prone to do so incorrectly. Furthermore, the likelihood of responding incorrectly was also dependent upon the strength of the semantic incongruity (i.e. whether it was weakly incongruent as “dial” is to “type” vs. strongly incongruent as “dial” is to “stir”), and this effect did not change according to which modality the incongruent material was presented in.

The results presented here differed from those of Kelly et al. on two fronts: Whereas the original study found that participants were slower to respond when faced with incongruent speech (and congruent gesture) than incongruent gesture (and congruent speech), we found no such effect. We did, however, find a similar effect of modality on error rates, with a greater number of errors when incongruity was presented in speech than when it was presented in gesture, but this was not found in the original study. It is possible that these discrepancies reflect differences in how participants in each study prioritised the speed and accuracy of their responses (See Figures C.2 and C.3 compared to Kelly et al. 2010, Figure. 2).

Both of these findings are somewhat unexpected given that congruent gestures were physically similar to the actions seen in the videos, whereas congruent speech was related only through linguistic convention (as noted by Kelly et al. 2010). Taken together, they suggest that listeners may process the relationship of speech to the previous action faster (Kelly et al., 2010) than that of gesture, and more accurately (Experiment C.1). One possible reason for this is that in comparison to discrete words and phonemes, gestures occur in a continuous physical space, often

leading to misinterpretation (see Feyereisen et al., 1988; Hadar & Pinchas-Zamir, 2004; Krauss et al., 1991).

Here, we replicate the main findings from Kelly et al. (2010), supporting the view that the semantic content of speech and gesture bidirectionally interact during comprehension. Although Experiment C.1 and the original study found differences in how semantic incongruity influenced error rates and reaction times depending on the modality in which it was presented, their respective findings are not incompatible. Taken together, results point towards possible differences in how listeners process speech and gesture, with spoken content being more quickly and more accurately interpreted than gestures, even when those gestures are comparatively transparent in meaning (in relation other types of gesturing).

References

- Akehurst, L., Kohnken, G., Vrij, A., & Bull, R. (1996). Lay Persons' and Police Officers' Beliefs Regarding Deceptive Behaviour. *Applied Cognitive Psychology*, 10, 461–471.
- Alibali, M. W. (2005). Gesture in Spatial Cognition: Expressing, Communicating, and Thinking About Spatial Information. *Spatial Cognition & Computation*, 5(4), 307–331.
- Alibali, M. W., Evans, J. L., Hostetter, A. B., Ryan, K., & Mainela-Arnold, E. (2009). Gesture–speech integration in narrative: Are children less redundant than adults? *Gesture*, 9(3), 290–311.
- Alibali, M. W., Heath, D. C., & Myers, H. J. (2001). Effects of Visibility between Speaker and Listener on Gesture Production: Some Gestures Are Meant to Be Seen. *Journal of Memory and Language*, 44(2), 169–188.
- Alibali, M. W., Kita, S., & Young, A. J. (2000). Gesture and the process of speech production: We think, therefore we gesture. *Language and Cognitive Processes*, 15(6), 593–613.
- Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the Time Course of Spoken Word Recognition Using Eye Movements: Evidence for Continuous Mapping Models. *Journal of Memory and Language*, 38(4), 419–439.
- Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition*, 73(3), 247–264.

- Arciuli, J., Mallard, D., & Villar, G. (2010). “{U}m, {I} can tell you’re lying”: Linguistic markers of deception versus truth-telling in speech. *Applied Psycholinguistics*, 31, 397–411.
- Arciuli, J., Villar, G., & Mallard, D. (2009). Lies, lies and more lies. In *Proceedings of the 31st annual conference of the cognitive science society* (pp. 2329–2334).
- Arnold, J. E., Hudson Kam, C., & Tanenhaus, M. K. (2007). If you say thee uh you are describing something hard: The on-line attribution of disfluency during reference comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33, 914–930.
- Arnold, J. E., Losongco, A., Wasow, T., & Ginstrom, R. (2000). Heaviness {vs.\@} Newness: The Effects of Structural Complexity and Discourse Status on Constituent Ordering. *Language*, 28–55.
- Arnold, J. E., Tanenhaus, M. K., Altmann, R. J., & Fagnano, M. (2004). The Old and Thee, uh, New. *Psychological Science*, 15, 578–582.
- Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge University Press.
- Babad, E., Bernieri, F., & Rosenthal, R. (1989). Nonverbal communication and leakage in the behavior of biased and unbiased teachers. *Journal of Personality and Social Psychology*, 56(1), 89.
- Bailey, K. G. D., & Ferreira, F. (2003). Disfluencies Affect the Parsing of Garden-Path Sentences. *Journal of Memory and Language*, 49, 183–200.
- Bailey, K. G. D., & Ferreira, F. (2007). The processing of filled pause disfluencies in the visual world. In R. Van Gompel, M. Fischer, W. Murray, & R. Hill (Eds.), *Eye movements: A window on mind and brain* (pp. 487–502). Elsevier.
- Bangerter, A. (2004). Using pointing and describing to achieve joint focus of attention in dialogue. *Psychological Science*, 15(6), 415–419.
- Barbieri, F., Buonocore, A., Volta, R. D., & Gentilucci, M. (2009). How symbolic gestures and words interact with each other. *Brain and Language*, 110(1), 1–11.

- Barr, D. J. (2001). Trouble in mind: Paralinguistic indices of effort and uncertainty in communication. In I. Guaitella, S. Santi, & C. Cavé (Eds.), *Oralité et gestualité: Interactions et comportements multimodaux dans la communication* (pp. 597–600). L’Harmattan.
- Barr, D. J. (2008). Analyzing visual world’ eyetracking data using multilevel logistic regression. *Journal of Memory and Language*, 59(4), 457–474.
- Barr, D. J., & Seyfeddinipur, M. (2010). The role of fillers in listener attributions for speaker disfluency. *Language and Cognitive Processes*, 25(4), 441–455.
- Barres, P. E., & Johnson-Laird, P. (2003). *On imagining what is true (and what is false)* (Vol. 9) (No. 1).
- Bates, D., Mächler, M., Bolker, B. M., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1215–1225.
- Bavelas, J. B., Chovil, N., Lawrie, D. A., & Wade, A. (1992). Interactive gestures. *Discourse Processes*, 15(4), 469–489.
- Bavelas, J. B., Gerwing, J., Sutton, C., & Prevost, D. (2008). Gesturing on the telephone: Independent effects of dialogue and visibility. *Journal of Memory and Language*, 58(2), 495–520.
- Beattie, G. (1979). Planning units in spontaneous speech: Some evidence from hesitation in speech and speaker gaze direction in conversation. *Linguistics*, 17, 61–78.
- Beattie, G., & Shovelton, H. (1999a). Do iconic handgestures really contribute anything to the semantic information conveyed by speech? An experimental investigation. *Semiotica*, 123, 1–30.
- Beattie, G., & Shovelton, H. (1999b). Mapping the range of information contained in the iconic hand gestures that accompany spontaneous speech. *Journal of Language and Social Psychology*, 18(4), 438–462.
- Benus, S., Enos, F., Hirschberg, J., & Shriberg, E. (2006). Pauses in Deceptive Speech. In *Proceedings of isca 3rd international conference on speech prosody*.

- Bergmann, K., Aksu, V., & Kopp, S. (2011). The relation of speech and gestures: temporal synchrony follows semantic synchrony. In *Proceedings of the 2nd workshop on gesture and speech in interaction (gespin 2011)* (pp. 1–6).
- Bernardis, P., & Gentilucci, M. (2006). Speech and gesture share the same communication system. *Neuropsychologia*, 44(2), 178–190.
- Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc*, 10(3), 214–234.
- Boomer, D. S. (1965). Hesitation and grammatical encoding. *Language and Speech*, 8(3), 148–158.
- Borovsky, A., Elman, J. L., & Fernald, A. (2012). Knowing a lot for one's age: Vocabulary skill and not age is associated with anticipatory incremental sentence interpretation in children and adults. *Journal of Experimental Child Psychology*, 112(4), 417–436.
- Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., & Brennan, S. E. (2001). Disfluency Rates in Conversation: Effects of Age, Relationship, Topic, Role, and Gender. *Language and Speech*, 44, 123–147.
- Brennan, S. E., & Williams, M. (1995). The Feeling of Another's Knowing: Prosody and Filled Pauses as Cues to Listeners about the Metacognitive States of Speakers. *Journal of Memory and Language*, 34, 383–398.
- Broaders, S. C., & Goldin-Meadow, S. (2010). Truth is at hand: How gesture adds information during investigative interviews. *Psychological Science*, 21(5), 623–628.
- Buller, D. B., & Burgoon, J. K. (2006). Interpersonal Deception Theory. *Communication Theory*, 6(3), 203–242.
- Burgoon, J. K., & Koper, R. J. (1984). Nonverbal and relational communication associated with reticence. *Human Communication Research*, 10(4), 601–626.
- Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., ... Narayanan, S. (2004). Analysis of emotion recognition using facial

- expressions, speech and multimodal information. In *Proceedings of the 6th international conference on multimodal interfaces - icmi '04* (pp. 205–211). New York, New York, USA: ACM Press.
- Butterworth, B., & Beattie, G. (1978). Gesture and silence as indicators of planning in speech. In R. Campbell (Ed.), *Recent advances in the psychology of language* (pp. 347–360). Springer.
- Carpenter, P. A., & Just, M. A. (1975). Sentence comprehension: a psycholinguistic processing model of verification. *Psychological review*, 82(1), 45.
- Cassell, J., Nakano, Y. I., Bickmore, T. W., Sidner, C. L., & Rich, C. (2001). Non-verbal cues for discourse structure. In *Proceedings of the 39th annual meeting on association for computational linguistics - acl '01* (pp. 114–123). Morristown, NJ, USA: Association for Computational Linguistics.
- Chawla, P., & Krauss, R. M. (1994). Gesture and Speech in Spontaneous and Rehearsed Narratives. *Journal of Experimental Social Psychology*, 30(6), 580–601.
- Chen, L., Harper, M., & Quek, F. (2002). Gesture patterns during speech repairs. In *Proceedings of the 4th ieee international conference on multimodal interfaces* (p. 155).
- Chieffi, S., Secchi, C., & Gentilucci, M. (2009). Deictic word and gesture production: Their interaction. *Behavioural Brain Research*, 203(2), 200–206.
- Christenfeld, N., Schachter, S., & Bilous, F. (1991). Filled pauses and gestures: It's not coincidence. *Journal of Psycholinguistic Research*, 20(1), 1–10.
- Chu, M., & Kita, S. (2016). Co-thought and co-speech gestures are generated by the same action generation process. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(2), 257.
- Chui, K. (2005). Temporal patterning of speech and iconic gestures in conversational discourse. *Journal of Pragmatics*, 37(6), 871–887.
- Church, R. B., Kelly, S. D., & Holcombe, D. (2014). Temporal synchrony between speech, action and gesture during language production. *Language, Cognition*

- and Neuroscience*, 29(3), 345–354.
- Clark, H. H. (1996). *Using language*. Cambridge University Press: Cambridge.
- Clark, H. H., & Wasow, T. (1998). Repeating Words in Spontaneous Speech. *Cognitive Psychology*, 37(3), 201–242.
- Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, 22(1), 1–39.
- Cohen, A. A. (1977). The Communicative Functions of Hand Illustrators. *Journal of Communication*, 27(4), 54–63.
- Cohen, D., Beattie, G., & Shovelton, H. (2010). Nonverbal indicators of deception: How iconic gestures reveal thoughts that cannot be suppressed. *Semiotica*, 2010(182), 133–174.
- Cole, P. M. (1986). Children's Spontaneous Control of Facial Expression. *Child Development*, 57(6), 1309.
- Cook, S. W., Jaeger, T. F., & Tanenhaus, M. K. (2009). Producing Less Preferred Structures: More Gestures, Less Fluency. In *Proceedings of the 31st annual conference of the cognitive science society* (pp. 62–67).
- Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken Language. *Cognitive Psychology*, 107(1), 84–107.
- Corley, M., & Hartsuiker, R. J. (2011). Why um helps auditory word recognition: The temporal delay hypothesis. *PLoS ONE*, 6, e19792.
- Corley, M., MacGregor, L. J., & Donaldson, D. I. (2007). It's the way that you, er, say it: Hesitations in speech affect language comprehension. *Cognition*, 105, 658–668.
- Dahan, D., & Tanenhaus, M. K. (2004). Continuous Mapping From Sound to Meaning in Spoken-Language Comprehension: Immediate Effects of Verb-Based Thematic Constraints. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(2), 498–513.
- Dahan, D., & Tanenhaus, M. K. (2005). Looking at the rope when looking for the snake: Conceptually mediated eye movements during spoken-word

- recognition. *Psychonomic Bulletin and Review*, 12(3), 453–459.
- Dahan, D., Tanenhaus, M. K., & Chambers, C. G. (2002). Accent and reference resolution in spoken-language comprehension. *Journal of Memory and Language*, 47(2), 292–314.
- Daly, S. (1978). Behavioural correlates of social anxiety. *British Journal of Social and Clinical Psychology*, 17(2), 117–120.
- De Ruiter, J. P. (n.d.). The production of gesture and speech. In D. McNeill (Ed.), *Language and gesture* (pp. 284–311). Cambridge: Cambridge University Press.
- De Ruiter, J. P. (1998a). Gesture and speech production. *Unpublished doctoral dissertation. Radboud University Nijmegen, Nijmegen*.
- De Ruiter, J. P. (1998b). *Gesture and speech production* (Unpublished doctoral dissertation). Radboud University Nijmegen Nijmegen.
- De Ruiter, J. P. (2006). Can gesticulation help aphasic people speak, or rather, communicate? *Advances in Speech Language Pathology*, 8(2), 124–127.
- De Ruiter, J. P., Bangerter, A., & Dings, P. (2012). The Interplay Between Gesture and Speech in the Production of Referring Expressions: Investigating the Tradeoff Hypothesis. *Topics in Cognitive Science*, 4(2), 232–248.
- DePaulo, B. M. (1985). Deceiving and detecting deceit. In B. R. Schlenker (Ed.), *The self and social life* (pp. 323–370). New York: McGraw-Hill.
- DePaulo, B. M., & Kashy, D. A. (1998). Everyday lies in close and casual relationships. *Journal of Personality and Social Psychology*, 74(1), 63–79.
- DePaulo, B. M., Kashy, D. A., Kirkendol, S. E., Wyer, M. M., & Epstein, J. A. (1996). Lying in everyday life. *Journal of Personality and Social Psychology*, 70(5), 979–995.
- DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to Deception. *Psychological Bulletin*, 129, 74–118.
- DePaulo, B. M., Rosenthal, R., Rosenkrantz, J., & Green, C. R. (1982). Actual and Perceived Cues to Deception: A Closer Look at Speech. *Basic and*

- Applied Social Psychology*, 3(4), 291–312.
- der Sluis, I., & Krahmer, E. (2007). Generating multimodal references. *Discourse Processes*, 44(3), 145–174.
- Driskell, J. E., & Radtke, P. H. (2003). The Effect of Gesture on Speech Production and Comprehension. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 45(3), 445–454.
- Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C., & Tanenhaus, M. K. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research*, 24, 409–436.
- Eisenstein, J., & Davis, R. (2004). Visual and linguistic information in gesture classification. In *Proceedings of the 6th international conference on multimodal interfaces* (pp. 113–120).
- Ekman, P. (1992). An Argument for Basic Emotions. *Cognition and Emotion*, 6(3-4), 169–200.
- Ekman, P., & Friesen, W. V. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, 1(1), 49–98.
- Ekman, P., & Friesen, W. V. (1975). *Unmasking the face: A guide to recognizing emotions from facial clues*. Prentice-Hall.
- Ekman, P., O'Sullivan, M., Friesen, W. V., & Scherer, K. R. (1991). Invited article: Face, voice, and body in detecting deceit. *Journal of Nonverbal Behavior*, 15(2), 125–135.
- Esposito, A., McCullough, K. E., & Quek, F. (2001). Disfluencies in gesture: Gestural correlates to filled and unfilled speech pauses. *IEEE International Workshop on Cues in Communication “Cues 2001”*, On CD—Rom.
- Farmer, T. A., Anderson, S., & Spivey, M. J. (2007). Gradiency and Visual Context in Syntactic Garden Paths. *Journal of Memory and Language*, 57, 570–595.
- Farmer, T. A., Cargill, S. A., Hindy, N. C., Dale, R., & Spivey, M. J. (2007).

- Tracking the continuity of language comprehension: Computer mouse trajectories suggest parallel syntactic processing. *Cognitive Science*, 31(5), 889–909.
- Feyereisen, P., Van de Wiele, M., & Dubois, F. (1988). The meaning of gestures: What can be understood without speech? *Cahiers de Psychologie Cognitive*, 8(1), 3–25.
- Finlayson, S., Forrest, V., Lickley, R., Beck, J. M., & Margaret, Q. (2003). Effects of the restriction of hand gestures on disfluency. *Proceedings of DiSS'03, Disfluency in Spontaneous Speech Workshop*(September), 21–24.
- Fox Tree, J. E. (1995). The Effects of False Starts and Repetitions on the Processing of Subsequent Words in Spontaneous Speech. *Journal of Memory and Language*, 34, 709–738.
- Galati, A. (2014). Speakers adapt gestures to addressees' knowledge: Implications for models of co-speech gesture. *Language, Cognition and Neuroscience*, 29(4), 435–451.
- Ganis, G., Kutas, M., & Sereno, M. I. (1996). The Search for "Common Sense": An Electrophysiological Study of the Comprehension of Words and Pictures in Reading. *Journal of Cognitive Neuroscience*, 8(2), 89–106.
- Gentilucci, M., Dalla Volta, R., & Gianelli, C. (2008). When the hands speak. *Journal of Physiology Paris*, 102(1-3), 21–30.
- Genz, A., & Bretz, F. (1999). Numerical computation of multivariate t-probabilities with application to power calculation of multiple contrasts. *Journal of Statistical Computation and Simulation*, 63(4), 103–117.
- Gerwing, J., & Allison, M. (2011). The flexible semantic integration of gestures and words: Comparing face-to-face and telephone dialogues. *Gesture*, 11(3), 308–329.
- Gibbs Jr, R. W., & Colston, H. L. (2012). *Interpreting figurative meaning*. Cambridge University Press.
- Ginzburg, J. (2012). *The interactive stance*. Oxford University Press.

- Glenberg, A. M., Schroeder, J. L., & Robertson, D. A. (1998). Averting the gaze disengages the environment and facilitates remembering. *Memory and Cognition*, 26(4), 651–658.
- Goldin-Meadow, S., Nusbaum, H., Kelly, S. D., & Wagner, S. (2001). Explaining math: gesturing lightens the load. *Psychological science: A Journal of the American Psychological Society / APS*, 12(6), 516–522.
- Goldin-Meadow, S., & Sandhofer, C. M. (1999). Gestures convey substantive information about a child's thoughts to ordinary listeners. *Developmental Science*, 2(1), 67–74.
- Gregersen, T. (2005). Nonverbal cues: Clues to the detection of foreign language anxiety. *Foreign Language Annals*, 38(3), 388–400.
- Grice, H. P. (1975). Logic and conversation. *Syntax and Semantics*, 3, 41–68.
- Grodner, D. J., Klein, N. M., Carbury, K. M., & Tanenhaus, M. K. (2010). "Some," and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. *Cognition*, 116(1), 42–55.
- Gullberg, M., & Holmqvist, K. (2006). What speakers do and what addressees look at: Visual attention to gestures in human interaction live and on video. *Pragmatics & Cognition*, 14(1), 53–82.
- Gullberg, M., & Kita, S. (2009). Attention to speech-accompanying gestures: Eye movements and information uptake. *Journal of Nonverbal Behavior*, 33(4), 251–277.
- Habets, B., Kita, S., Shao, Z., Özyürek, A., & Hagoort, P. (2011). The role of synchrony and ambiguity in speech-gesture integration during comprehension. *Journal of Cognitive Neuroscience*, 23(8), 1845–1854.
- Hadar, U. (1989). Two Types of Gesture and Their Role in Speech Production. *Journal of Language and Social Psychology*, 8(3-4), 221–228.
- Hadar, U., & Butterworth, B. (1997). Iconic gestures, imagery, and word retrieval in speech. *Semiotica*, 115(1-2), 147–172.
- Hadar, U., & Pinchas-Zamir, L. (2004). The semantic specificity of gesture:

- Implications for gesture classification and function. *Journal of Language and Social Psychology*, 23(2), 204–214.
- Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of Word Meaning and World Knowledge in Language Comprehension. *Science*, 304(5669), 438–441.
- Hans, A., & Hans, E. (2015). Kinesics, Haptics and Proxemics: Aspects of Non-Verbal Communication. *Journal Of Humanities And Social Science Ver*, 20(2), 47–52.
- Hart, C. L., Fillmore, D. G., & Griffith, J. D. (2009). Current research in social psychology. *Current Research in Social Psychology*, 14(9), 134–142.
- Hartwig, M., & Bond, C. F. (2011). Why do lie-catchers fail? A lens model meta-analysis of human lie judgments. *Psychological Bulletin*, 137(4), 643–659.
- Heller, D., Arnold, J. E., Klein, N. M., & Tanenhaus, M. K. (2015). Inferring Difficulty: Flexibility in the Real-time Processing of Disfluency. *Language and Speech*, 58(2), 190–203.
- Hoetjes, M., Koolen, R., Goudbeek, M., Krahmer, E., & Swerts, M. (2015). Reduction in gesture during the production of repeated references. *Journal of Memory and Language*, 79-80, 1–17.
- Hoetjes, M., Krahmer, E., & Swerts, M. (2014). Does our speech change when we cannot gesture? *Speech Communication*, 57, 257–267.
- Holle, H., & Gunter, T. C. (2007). The Role of Iconic Gestures in Speech Disambiguation: ERP Evidence. *Journal of Cognitive Neuroscience*, 19(7), 1175–1192.
- Holle, H., Obermeier, C., Schmidt-Kassow, M., Friederici, A. D., Ward, J., & Gunter, T. C. (2012). Gesture facilitates the syntactic analysis of speech. *Frontiers in Psychology*, 3(MAR), 1–12.
- Holler, J., Kendrick, K. H., & Levinson, S. C. (2017). Processing language in face-to-face conversation: Questions with gestures get faster responses. *Psychonomic Bulletin and Review*, 1–9.

- Holler, J., Shovelton, H., & Beattie, G. (2009). Do Iconic Hand Gestures Really Contribute to the Communication of Semantic Information in a Face-to-Face Context? *Journal of Nonverbal Behavior*, 33(2), 73–88.
- Holler, J., & Stevens, R. (2007). The effect of common ground on how speakers use gesture and speech to represent size information. *Journal of Language and Social Psychology*, 26(1), 4–27.
- Holler, J., & Wilkin, K. (2011). Co-speech gesture mimicry in the process of collaborative referring during face-to-face dialogue. *Journal of Nonverbal Behavior*, 35(2), 133–153.
- Hostetter, A. B. (2011). When do gestures communicate? A meta-analysis. *Psychological Bulletin*, 137(2), 297–315.
- Hostetter, A. B., Alibali, M. W., & Kita, S. (2007a). Does sitting on your hands make you bite your tongue? The effects of gesture prohibition on speech during motor descriptions. *Proceedings of the Cognitive Science Society*, 29(29).
- Hostetter, A. B., Alibali, M. W., & Kita, S. (2007b). I see it in my hands' eye: Representational gestures reflect conceptual demands. *Language and Cognitive Processes*, 22(3), 313–336.
- Hostetter, A. B., Alibali, M. W., & Schrager, S. M. (2011). If you don't already know, I'm certainly not going to show you! In G. Stam & M. Ishino (Eds.), *Integrating gestures: The interdisciplinary nature of gesture* (Vol. 4). Amsterdam, Netherlands: John Benjamins Publishing Company.
- Huang, Y. T., & Snedeker, J. (2009). Online interpretation of scalar quantifiers: Insight into the semantics-pragmatics interface. *Cognitive Psychology*, 58(3), 376–415.
- Huettig, F., & Altmann, G. T. M. (2005). Word meaning and the control of eye fixation: Semantic competitor effects and the visual world paradigm. *Cognition*, 96(1), 23–32.
- Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world

- paradigm to study language processing: A review and critical evaluation. *Acta Psychologica*, 137(2), 151–171.
- Igualada, A., Esteve-gibert, N., & Prieto, P. (2017). Child Beat gestures improve word recall in 3- to 5-year-old children. *Journal of Experimental Child Psychology*, 156, 99–112.
- Ito, A., Pickering, M. J., & Corley, M. (2018). Investigating the time-course of phonological prediction in native and non-native speakers of English: A visual world eye-tracking study. *Journal of Memory and Language*, 98, 1–11.
- Jacobs, N., & Garnham, A. (2007). The role of conversational hand gestures in a narrative task. *Journal of Memory and Language*, 56(2), 291–303.
- Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, 49(1), 133–156.
- Kaup, B., Ludtke, J., & Zwaan, R. A. (2007). The experiential view of language comprehension: How is negation represented? In F. Schmalhofer & C. A. Perfetti (Eds.), *Higher level language processes in the brain: Inference and comprehension processes* (pp. 255–288). Psychology Press.
- Kelly, S. D. (2001). Broadening the units of analysis in communication: Speech and nonverbal behaviours in pragmatic comprehension. *Journal of Child Language*, 28(2), 325–349.
- Kelly, S. D., Barr, D. J., Church, R. B., & Lynch, K. (1999). Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory. *Journal of Memory and Language*, 40(4), 577–592.
- Kelly, S. D., & Church, R. B. (1998). A comparison between children's and adults' ability to detect conceptual information conveyed through representational gestures. *Child Development*, 69(1), 85–93.
- Kelly, S. D., & Goldsmith, L. H. (2004). Gesture and right hemisphere involvement in evaluating lecture material. *Gesture*, 4(1), 25–42.

- Kelly, S. D., Kravitz, C., & Hopkins, M. (2004). Neural correlates of bimodal speech and gesture comprehension. *Brain and Language*, 89(1), 253–260.
- Kelly, S. D., Özyürek, A., & Maris, E. (2010). Two Sides of the Same Coin. *Psychological Science*, 21(2), 260–267.
- Keltner, D. (1995). Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame. *Journal of Personality and Social Psychology*, 68(3), 441–454.
- Keltner, D., & Harker, L. (1998). The forms and functions of the nonverbal signal of shame. In P. Gilbert & B. Andrews (Eds.), *Series in affective science. shame: Interpersonal behavior, psychopathology, and culture* (pp. 78–98). New York, NY, US: Oxford University Press.
- Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press.
- Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking Perspective in Conversation: The Role of Mutual Knowledge in Comprehension. *Psychological Science*, 11(1), 32–38.
- Kieslich, P. J., Schoemann, M., Grage, T., Hepp, J., & Scherbaum, S. (2019). Design factors in mouse-tracking: What makes a difference? *Behavior Research Methods*, 1–25.
- King, J. P. J., Loy, J. E., & Corley, M. (2018). Contextual Effects on Online Pragmatic Inferences of Deception. *Discourse Processes*, 55(2), 123–135. Retrieved from <https://www.tandfonline.com/doi/full/10.1080/0163853X.2017.1330041> doi: 10.1080/0163853X.2017.1330041
- Kita, S. (2000). How representational gestures help speaking. In D. McNeill (Ed.), *Language and gesture* (Vol. 1, pp. 162–185). Cambridge University Press.
- Kita, S. (2014). Production of speech-accompanying gestures. In V. Ferreira, M. Goldrick, & M. Miozzo (Eds.), *The oxford handbook of language production* (Vol. 48, p. 89). Oxford University Press.
- Kita, S., & Özyürek, A. (2003). What does cross-linguistic variation in semantic

- coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language*, 48(1), 16–32.
- Kjelgaard, M. M., & Speer, S. R. (1999). Prosodic Facilitation and Interference in the Resolution of Temporary Syntactic Closure Ambiguity. *Journal of Memory and Language*, 40(2), 153–194.
- Krahmer, E., & Swerts, M. (2007). The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language*, 57(3), 396–414.
- Krauss, R. M., Chen, Y., & Gotfexnum, R. F. (2000). Lexical gestures and lexical access: a process model. *Language and Gesture*, 2, 261.
- Krauss, R. M., Dushay, R. a., Chen, Y., & Rauscher, F. H. (1995). *The Communicative Value of Conversational Hand Gesture* (Vol. 31) (No. 6).
- Krauss, R. M., Morrel-Samuels, P., & Colasante, C. (1991). Do Conversational Hand Gestures Communicate? *Journal of Personality and Social Psychology*, 61(5), 743–754.
- Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, 31(1), 32–59.
- Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the n400 component of the event-related brain potential (erp). *Annual Review of Psychology*, 62, 621–647.
- Kutas, M., & Hillyard, S. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science*, 207(4427), 203–205.
- Kutas, M., Neville, H. J., & Holcomb, P. J. (1987). A preliminary comparison of the N400 response to semantic anomalies during reading, listening and signing. *Electroencephalography and Clinical Neurophysiology. Supplement*, 39, 325–330.
- Louwerse, M. M., & Bangerter, A. (2010). Effects of Ambiguous Gestures and

- Language on the Time Course of Reference Resolution. *Cognitive Science*, 34(8), 1517–1529.
- Loy, J. E. (2017). Effects of manner of delivery in on-line pragmatic inferences. *Unpublished doctoral dissertation. University of Edinburghm, Edinburgh.*
- Loy, J. E., Rohde, H., & Corley, M. (2016). Lying, in a Manner of Speaking. In J. Barnes, A. Brugos, S. Shattuck-Hufnagel, & N. Veilleux (Eds.), *Proceedings of speech prosody 8* (pp. 984–988). Boston, MA.
- Loy, J. E., Rohde, H., & Corley, M. (2017). Effects of Disfluency in Online Interpretation of Deception. *Cognitive Science*, 41, 1434–1456.
- Lucero, C., Zaharchuk, H., & Casasanto, D. (2014). Beat gestures facilitate speech production. In *Proceedings of the 36th annual conference of the cognitive science society* (Vol. 36).
- Lyons, J. (1977). *Semantics* (Vol. 2). Cambridge University Press.
- Maricchiolo, F., Gnisci, A., Bonaiuto, M., & Ficca, G. (2009). Effects of different types of hand gestures in persuasive speech on receivers' evaluations. *Language and Cognitive Processes*, 24(2), 239–266.
- Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition*, 25(1-2), 71–102.
- Masson-Carro, I., Goudbeek, M., & Krahmer, E. (2015). Can you handle this? The impact of object affordances on how co-speech gestures are produced. *Language, Cognition and Neuroscience*, 3798(3), 430–440.
- Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, 44(2), 314–324.
- Mayberry, R. I., & Jaques, J. (2000). Gesture production during stuttered speech: Insights into the nature of gesture-speech integration. *Language and Gesture*, 2(199), 15.
- McKinstry, C., Dale, R., & Spivey, M. J. (2008). Action dynamics reveal parallel competition in decision making. *Psychological Science*, 19(1), 22–24.

- McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago press.
- McNeill, D. (2000). *Language and gesture* (Vol. 2). Cambridge University Press.
- McNeill, D. (2005). *Gesture and Thought*. University of Chicago press.
- McPherson, W. B., & Holcomb, P. J. (1999). An electrophysiological investigation of semantic priming with pictures of real objects. *Psychophysiology*, 36(1), 53–65.
- Mehrabian, A., & Ferris, S. R. (1967). Inference of attitudes from nonverbal communication in two channels. *Journal of Consulting Psychology*, 31(3), 248.
- Melinger, A., & Kita, S. (2007). Conceptualisation load triggers gesture production. *Language and Cognitive Processes*, 22(4), 473–500.
- Melinger, A., & Levelt, W. J. M. (2004). Gesture and the communicative intention of the speaker. *Gesture*, 4(2), 119–141.
- Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language*, 59(4), 475–494.
- Mitchell, D. C., Cuetos, F., Corley, M., & Brysbaert, M. (1995). Exposure-based models of human parsing: Evidence for the use of coarse-grained (nonlexical) statistical records. *Journal of Psycholinguistic Research*, 24(6), 469–488.
- Mol, L., Krahmer, E., Maes, A., & Swerts, M. (2011). Seeing and being seen: The effects on gesture production. *Journal of Computer-Mediated Communication*, 17(1), 77–100.
- Moreno, K. (2011). Sincere Facial Expression; Sarcastic Facial Expression. In *Clippix etc.* Florida Center for Instructional Technology. Retrieved from <https://etc.usf.edu/clippix>
- Morrel-Samuels, P., & Krauss, R. M. (1992). Word Familiarity Predicts Temporal Asynchrony of Hand Gestures and Speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(3), 615–622.

- Morsella, E., & Krauss, R. M. (2004). The Role of Gestures in Spatial Working Memory and Speech. *American Journal of Psychology*, 117(3), 411–424.
- Moss, H. E., & Marslen-Wilson, W. D. (1993). Access to word meanings during spoken language comprehension: effects of sentential semantic context. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19(6), 1254–1276.
- Nieuwland, M. S., & Van Berkum, J. J. A. (2006). When Peanuts Fall in Love: N400 Evidence for the Power of Discourse. *Journal of Cognitive Neuroscience*, 18(7), 1098–1111.
- Obermeier, C., Kelly, S. D., & Gunter, T. C. (2014). A speaker's gesture style can affect language comprehension: ERP evidence from gesture-speech integration. *Social Cognitive and Affective Neuroscience*, 10(9), 1236–1243.
- Obermeier, C., Kelly, S. D., & Gunter, T. C. (2015). A speaker's gesture style can affect language comprehension: ERP evidence from gesture-speech integration. *Social Cognitive and Affective Neuroscience*, 10(9), 1236–1243.
- Oviatt, S. (1995). Predicting spoken disfluencies during human-computer interaction. *Computer Speech and Language*, 9(1), 19–36.
- Özyürek, A., Willems, R. M., Kita, S., & Hagoort, P. (2007). On-line Integration of Semantic Information from Speech and Gesture: Insights from Event-related Brain Potentials. *Journal of Cognitive Neuroscience*, 19(4), 605–616.
- Picheny, M. A., Durlach, N. I., & Braida, L. D. (1986). Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. *Journal of Speech, Language, and Hearing Research*, 29(4), 434–446.
- R Core Team. (2018). R: A Language and Environment for Statistical Computing [Computer software manual]. Vienna, Austria. Retrieved from <https://www.r-project.org/>
- Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological methodology*, 25, 111–164.
- Rauscher, F. H., Krauss, R. M., & Chen, Y. (1996). Gesture, Speech, and Lexical

- Access: The Role of Lexical Movements in Speech Production. *Psychological Science*, 7(4), 226–231.
- Ravizza, S. (2003). Movement and lexical access: do noniconic gestures aid in retrieval? *Psychonomic Bulletin & Review*, 10(3), 610–615.
- Rayner, K. (1998). Eye movements in Reading and Information Processing: 20 Years of Research. *Psychological Bulletin*, 124(3), 372–422.
- Roberts, C. (2012). Information structure: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, 5, 1–69.
- Rommers, J., Meyer, A. S., Praamstra, P., & Huettig, F. (2013). The contents of predictions in sentence comprehension: Activation of the shape of objects before they are referred to. *Neuropsychologia*, 51(3), 437–447.
- Rozin, P., & Cohen, A. B. (2003). High Frequency of Facial Expressions Corresponding to Confusion, Concentration, and Worry in an Analysis of Naturally Occurring Facial Expressions of Americans. *Emotion*, 3(1), 68–75.
- Saryazdi, R., & Chambers, C. G. (2017). Attentional factors in listeners' uptake of gesture cues during speech processing. *Interspeech, 2017-Augus*, 869–873.
- Schachter, S., Christenfeld, N., Ravina, B., & Bilous, F. (1991). Speech disfluency and the structure of knowledge. *Journal of Personality and Social Psychology*, 60, 362–367.
- Schachter, S., Rauscher, F. H., Crone, K. T., & Christenfeld, N. (1994). The vocabularies of academia. *Psychological Science*, 5(1), 37–41.
- Schnadt, M. J., & Corley, M. (2006). The influence of lexical, conceptual and planning based factors on disfluency production. In *Proceedings of the 28th annual meeting of the cognitive science society* (Vol. 28).
- Searle, J. R. (1969). *Speech Acts* (Vol. 46). Cambridge: Cambridge University Press.
- Shriberg, E. (1996). Disfluencies in switchboard. In *Proceedings of international conference on spoken language processing* (Vol. 96, pp. 11–14).

- Shriberg, E. (2001). To errr' is human: Ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, 31(1), 153–169.
- Silverman, L. B., Bennetto, L., Campana, E., & Tanenhaus, M. K. (2010). Speech-and-gesture integration in high functioning autism. *Cognition*, 115(3), 380–393.
- Smith, V. L., & Clark, H. H. (1993). On the Course of Answering Questions. *Journal of Memory and Language*, 32, 25–38.
- Snodgrass, J., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, 6(2), 174–215.
- So, W. C., Kita, S., & Goldin-Meadow, S. (2009). Using the hands to identify who does what to whom: Gesture and speech go hand-in-hand. *Cognitive Science*, 33(1), 115–125.
- Spivey, M. J., & Geng, J. J. (2001). Oculomotor mechanisms activated by imagery and memory: Eye movements to absent objects. *Psychological Research*, 65(4), 235–241.
- Spivey, M. J., Grosjean, M., & Knoblich, G. (2005). From The Cover: Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences*, 102(29), 10393–10398.
- Sporer, S. L., & Schwandt, B. (2006). Paraverbal indicators of deception: a meta-analytic synthesis. *Applied Cognitive Psychology*, 20(4), 421–446.
- Swerts, M. (1998). Filled pauses as markers of discourse structure. *Journal of Pragmatics*, 30(4), 485–496.
- Swerts, M., & Krahmer, E. (2005). Audiovisual prosody and feeling of knowing. *Journal of Memory and Language*, 53, 81–94.
- Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217), 1632–1634.

- Tavakoli, H. R., Ahmed, F., Borji, A., & Laaksonen, J. (2017). Saliency Revisited: Analysis of Mouse Movements Versus Fixations. In *2017 ieee conference on computer vision and pattern recognition (cvpr)* (pp. 6354–6362). IEEE.
- Tian, Y., & Breheny, R. (2016). Dynamic pragmatic view of negation processing. In P. Larrivée & C. Lee (Eds.), *Negation and polarity: Experimental perspectives* (pp. 21–43). Springer.
- Van Berkum, J. J. A., Hagoort, P., & Brown, C. M. (1999). Semantic Integration in Sentences and Discourse: Evidence from the N400. *Journal of Cognitive Neuroscience*, 11(6), 657–671.
- Van Berkum, J. J. A., van den Brink, D., Tesink, C. M. J. Y., Kos, M., & Hagoort, P. (2008). The Neural Integration of Speaker and Message. *Journal of Cognitive Neuroscience*, 20(4), 580–591.
- Vrij, A. (1995). Behavioral Correlates of Deception in a Simulated Police Interview. *The Journal of Psychology*, 129(1), 15–28.
- Vrij, A., Kneller, W., & Mann, S. (2000). The effect of informing liars about Criteria-Based Content Analysis on their ability to deceive CBCA-raters. *Legal and Criminological Psychology*, 5(1), 57–70.
- Vrij, A., & Semin, G. R. (1996). Lie experts' beliefs about nonverbal indicators of deception. *Journal of Nonverbal Behavior*, 20(1), 65–80.
- Vrij, A., Semin, G. R., & Bull, R. (1996). Insight Into Behavior Displayed During Deception. *Human Communication Research*, 22(4), 544–562.
- Wagner, P., Malisz, Z., & Kopp, S. (2014). Gesture and speech in interaction: An overview. *Speech Communication*, 57, 209–232.
- Watanabe, M. (2002). Fillers as indicators of discourse segment boundaries in Japanese monologues. In *Speech prosody 2002, international conference*.
- Wesp, R., Hesse, J., Keutmann, D., & Wheaton, K. (2001). Gestures maintain spatial imagery. *American Journal of Psychology*, 114(4), 591–600.
- Wu, Y. C., & Coulson, S. (2005). Meaningful gestures: Electrophysiological indices of iconic gesture comprehension. *Psychophysiology*, 42(6), 654–667.

- Yeo, A., & Alibali, M. W. (2017). Evidence for overt visual attention to hand gestures as a function of redundancy and speech disfluency. In *Cogsci: The annual meeting of the cognitive science society*.
- Zuckerman, M., DePaulo, B. M., & Rosenthal, R. (1981). Verbal and Nonverbal Communication of Deception. *Advances in Experimental Social Psychology*, 14, 1–59.
- Zuckerman, M., Koestner, R., & Driver, R. (1981). Beliefs about cues associated with deception. *Journal of Nonverbal Behavior*, 6(2), 105–114.