This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

- This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
- A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
- This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
- The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
- When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.
Imagining and anticipating another speaker’s utterances in joint language tasks

Chiara Gambi

A thesis submitted in fulfilment of requirements
for the degree of Doctor of Philosophy

to
Department of Psychology
School of Philosophy, Psychology, and Language Sciences
The University of Edinburgh

2013
I hereby declare:

(a) that this thesis is of my own composition, and

(b) that the work reported in this thesis has been carried out by myself, except where acknowledgement of the work of others is made in text, and

(c) that the work has not been submitted for any other degree or professional qualification except as specified in text.

Chiara Gambi
Abstract

There is substantial evidence that comprehenders predict language. In addition, dialogue partners seem to predict one another, as shown by well-timed turn-taking and by the fact that they can complete one another’s utterances. However, little is known about the mechanisms that (i) support the ability to form predictions of others’ utterances and (ii) allow such predictions to be integrated with representations of one’s own utterances. I propose (predictive) representations of others’ utterances are computed within a cognitive architecture that makes use of mechanisms routinely used in language production (i.e., for the representation of one’s own utterances). If this proposal is right, representing that another person is about to speak (and, possibly, representing what they are about to say) should affect the process of language production, as the two processes are based on overlapping mechanisms. I test this hypothesis in a series of novel joint language tasks. Psycholinguistic tasks (picture naming and picture description) that have traditionally been used to study individual language production are distributed across two participants, who either produce two utterances simultaneously or consecutively. In addition, solo versions of the same tasks (where only one participant speaks, while the other participant remains silent) are tested. Speech onset latencies and utterance duration measures are compared between the solo and the joint task. In a first set of experiments about simultaneous production, I show that participants take longer to name pictures when they believe that their partner is concurrently naming pictures than when they believe their partner is silent or is concurrently categorizing the pictures as being from the same or from different semantic categories. Second, I show that participants find it harder to stop speaking when they know that their partner is about to speak. These findings suggest that speakers are able to represent that another person is about to speak using some of the same mechanisms they use to produce language. However, in a third series of experiments, I show that participants do not routinely anticipate the content and timing of another person’s utterance in a way that affects concurrent production of utterances. In light of this evidence, I discuss the proposal that speakers use language production mechanisms to represent and anticipate their partner’s utterances and support coordination in dialogue.
Acknowledgements

Several people made it possible for me to complete this thesis (and to do it in time). First, I would like to thank my primary supervisor, Martin Pickering, for the continuous support he provided in the form of discussions, comments, and advice on the subject matter and on academic issues in general. I would also like to thank my secondary supervisor, Rob Hartsuiker, who was equally supportive during my stay at the University of Ghent.

Second, I am very grateful to the School of Philosophy, Psychology, and Language Studies at the University of Edinburgh for the financial support I received in the form of a Career Development Studentship, and of several Research Support Grants that allowed me to compensate participants and attend conferences including AMLaP 2011, the Joint Action Meeting 2011, and Cognitive Science 2013. In addition, I am grateful to the College of Humanities and Social Studies for granting me a Research Scholarship that covered programme fees and provided an allowance of £400 per year to support my research activities. I would also like to thank Martin Corley, who served as the School’s Director of Postgraduate Studies at the time of my admission to the PhD Programme in Psychology. Besides, I am grateful to the Experimental Psychology Society for recently awarding a Grindley Grant that allowed me to travel to Berlin to attend Cognitive Science 2013.

Third, Eddie Dubourg and Ziggy Campbell should figure high in any acknowledgements list, as they provide excellent technical support to students in PPLS. Their help with E-Prime and audio recording was certainly invaluable for me.

Fourth, I would like to thank the students I have worked with over the past three years. Their precious contributions are stated in detail at the end of each chapter, but I list them all here: Uschi Cop, Joris Van de Cavey, Sven Radcke, Kerry Goodacre, Ashild Bodal, Jamie Crowther, Alicia Harris, Regina Bieche, Isabella Malicki. Particularly, Uschi Cop and Joris Van de Cavey made significant contributions (see Chapters 2, 3, and 4). I thank them for the great team work, for sharing ups and downs of the research process, and for their friendship.

Last, but definitely not least, many of the ideas presented in this thesis would have been far less clear had I not discussed them with several brilliant people. These include all the members of the Joint Action Reading Group, and particularly Olle Blomberg, Ed Baggs, Alessia Tosi, Alex Papadopoulos, Elli Drake, Jarek Lelonkiewicz, and Ian Finlayson. But I would like to extend my thanks to all the others PhD students working on language in the Department of Psychology, and especially to Madeleine Beveridge who shared many late-work evenings with me.
# Table of contents

1. **Introduction**.......................................................................................................................... 1  
   1.1 Talking together................................................................................................................... 2  
   1.2 Acting together..................................................................................................................... 8  
      1.2.1 Synchronization ............................................................................................................ 8  
         1.2.1.1 From intrapersonal to interpersonal synchronization........................................... 8  
         1.2.1.2 Symmetric and asymmetric entrainment patterns............................................... 11  
      1.2.2 Taking turns.................................................................................................................. 13  
         1.2.2.1 Accounts of joint task effects: co-representation versus referential spatial coding........................................................................................................................................ 15  
         1.2.2.2 The neural correlates of co-representation.............................................................. 20  
         1.2.2.3 Accounts of joint task effects: co-representation versus agent conflict.................. 23  
      1.2.3 Complementary actions................................................................................................. 25  
         1.2.3.1 Complementary actions in the brain......................................................................... 27  
   1.3 Theories of acting together.................................................................................................... 30  
      1.3.1 The dynamical systems approach................................................................................... 30  
      1.3.2 The shared representations approach.......................................................................... 34  
      1.3.3 The internal simulation approach: forward models for own and others’ actions...... 38  
         1.3.3.1 Forward models........................................................................................................ 39  
         1.3.3.2 Predicting others’ actions....................................................................................... 42  
         1.3.3.3 Prediction of others’ actions and coordination in joint tasks............................... 48  
   1.4 Theories of talking together................................................................................................. 50  
      1.4.1 Interactive alignment...................................................................................................... 57  
         1.4.1.1 Sound-related parity............................................................................................... 58  
         1.4.1.2 Boundary conditions for motor involvement in speech perception...................... 60  
         1.4.1.3 Lexico-semantic and syntactic parity..................................................................... 63  
         1.4.1.4 Summary and discussion......................................................................................... 65  
      1.4.2 The integrated theory of language comprehension and language production.............. 66  
         1.4.2.1 Prediction in language comprehension.................................................................... 67  
         1.4.2.2 Prediction in language comprehension uses production mechanisms.................... 71  
         1.4.2.3 Forward models are implicated in producing, imagining, and comprehending utterances................................................................................................................................. 75  
      1.4.2.4 Forward models in dialogue...................................................................................... 77  
      1.4.3 A cognitive architecture for the coordination of utterances......................................... 79  
         1.4.3.1 Predicting one’s own utterances.............................................................................. 82  
         1.4.3.2 Predicting another’s utterances.............................................................................. 84  
      1.4.3.3 The time course of predictions.................................................................................. 87  
   1.5 Methodological rationale: comparing self- and other-representations in joint language tasks........................................................................................................................................... 90  
      1.5.1 Simultaneous production............................................................................................... 90  
      1.5.2 Consecutive productions............................................................................................. 92  
   1.6 Summary............................................................................................................................... 96  
2. **Study 1. Experiments 1-4: simultaneous speaking and imagining**.................................... 97  
   2.1 Introduction.......................................................................................................................... 97  
      2.1.1 Representing another’s task......................................................................................... 99  
      2.1.2 Representing another’s utterances.............................................................................. 101
2.2 Experiment 1....................................................................................................................................105
  2.2.1 Method....................................................................................................................................111
  2.2.2 Results....................................................................................................................................115
  2.2.3 Discussion................................................................................................................................118
2.3 Experiment 2....................................................................................................................................120
  2.3.1 Method....................................................................................................................................122
  2.3.2 Results....................................................................................................................................123
    2.3.2.1 Semantic categorization task..........................................................................................123
    2.3.2.2 Picture naming task..........................................................................................................124
  2.3.3 Discussion................................................................................................................................125
2.4 Experiment 3....................................................................................................................................127
  2.4.1 Method....................................................................................................................................130
  2.4.2 Results....................................................................................................................................130
  2.4.3 Discussion................................................................................................................................132
2.5 Experiment 4....................................................................................................................................134
  2.5.1 Method....................................................................................................................................135
  2.5.2 Results....................................................................................................................................135
  2.5.3 Discussion................................................................................................................................138
2.6 General Discussion..........................................................................................................................140
2.7 Acknowledgements..........................................................................................................................145
3. Study 2. Experiment 5: effect of anticipating another’s utterance on stopping
speech..................................................................................................................................................147
  3.1 Introduction.....................................................................................................................................147
    3.1.1 Three accounts of the relationship between prediction and production
       mechanisms.................................................................................................................................148
    3.1.2 Coordinating stopping and resuming speech........................................................................151
    3.1.3 An experimental comparison of the accounts.......................................................................153
  3.2 Method..........................................................................................................................................155
  3.3 Results..........................................................................................................................................163
    3.3.1 Response type data................................................................................................................163
    3.3.2 Initial duration........................................................................................................................166
  3.4 Discussion.......................................................................................................................................169
    3.4.1 Predicting that you are about to speak..................................................................................170
    3.4.2 Prediction and the coordination of utterances in dialogue..................................................174
  3.5 Acknowledgements........................................................................................................................175
4. Study 3. Experiments 6-8: effect of anticipating another’s utterance on time to start
speaking and time spent speaking..........................................................................................................177
  4.1 Introduction.....................................................................................................................................177
  4.2 Experiment 6....................................................................................................................................180
    4.2.1 Method....................................................................................................................................184
    4.2.2 Results....................................................................................................................................191
      4.2.2.1 Likelihood of disfluency while describing the preamble..................................................192
      4.2.2.2 Timing analyses..............................................................................................................194
    4.2.3 Discussion................................................................................................................................198
  4.3 Experiment 7....................................................................................................................................201
    4.3.1 Method....................................................................................................................................205
    4.3.2 Results....................................................................................................................................209
    4.3.3 Discussion................................................................................................................................211
  4.4 Experiment 8....................................................................................................................................214
    4.4.1 Method....................................................................................................................................218
Figures

*Figure 1-1* Simultaneous production

*Figure 1-2* Consecutive productions

*Figure 2-1* Partner’s task manipulation in Experiment 1

*Figure 2-2* Hypothesized effects of Partner’s task in Study 1 according to the co-representation, agent conflict, and the no-representation accounts

*Figure 3-1* Schematic representation of the Design and Procedure of Experiment 5

*Figure 4-1* Schematic representation of the Design of Experiment 6, with a sample item

*Figure 4-2* Schematic representation of the Design and Procedure of Experiment 7

*Figure 4-3* Schematic representation of the Design and Procedure of Experiment 8

*Figure 4-4* Distribution of the proportion of overlapping trials in the OTHER condition of Experiments 8a and 8b
Tables

Table 2-1 Error counts (percentages) by Partner and Relatedness in Experiment 1............................116
Table 2-2 Mean voice onset times in ms (and standard deviations) by Partner and Relatedness in Experiment 1..............................................................................................................................117
Table 2-3 Error counts (percentages) by Partner and Relatedness in Experiment 2.........................124
Table 2-4 Mean voice onset times in ms (and standard deviations) by Partner and Relatedness in Experiment 2..............................................................................................................................125
Table 2-5 Error counts (percentages, out of the total number of scorable trials) by Partner and Relatedness in Experiment 3 ....................................................................................................................131
Table 2-6 Mean voice onset times in ms (and standard deviations) by Partner, Relatedness, and Size in Experiment 3..............................................................................................................................132
Table 2-7 Error counts (percentages, out of the total number of scorable trials) by Partner, Relatedness, and Size in Experiment 4..............................................................................................................................136
Table 2-8 Mean voice onset times in ms (and standard deviations) by Partner and Relatedness in Experiment 4..............................................................................................................................138
Table 3-1 Anticipated likelihood of completing the initial word in the SELF, OTHER, and NO conditions according to the separate mechanisms, shared mechanisms, and overlapping mechanisms accounts (Experiment 5)............................................................................................................................155
Table 3-2 Percentages of Completed, Interrupted, and Skipped Initial Responses by Degradation Level and Condition in Experiment 5.........................................................................................164
Table 3-3 Response Type analyses (complete model) for Experiment 5.............................................165
Table 3-4 Initial Duration by Initial Response Type and Degradation in the three conditions of Experiment 5..............................................................................................................................166
Table 3-5 Initial Duration analyses (completed responses, full model) for Experiment 5.................167
Table 3-6 Initial Duration analyses by Condition (completed responses only, full model) for Experiment 5..............................................................................................................................168
Table 4-1 Disfluency counts by Difficulty and Partner in Experiment 6...........................................193
Table 4-2 Mean Onset (seconds) and standard deviation of Onset by Difficulty and Partner in Experiment 6..............................................................................................................................194
Table 4-3 Mean Duration (seconds) and standard deviation of Onset by Difficulty and Partner in Experiment 6..............................................................................................................................195
Table 4-4 Mean Duration (seconds) and standard deviation of Gap by Difficulty and Partner in Experiment 6..............................................................................................................................197
Table 4-5 Mean Duration (seconds) and standard deviation of Continuation by Difficulty and Partner in Experiment 6..............................................................................................................................197
Table 4-6 Error counts (percentages) by Partner and Type in Experiment 7.................................210
Table 4-7 Mean onsets (and standard deviations) in milliseconds by Partner and Type in Experiment 7..................................................................................................................................................211
Table 4-8 Counts of disfluent trials (and percentages) by Partner and Difficulty in Experiment 8a (randomized) and 8b (blocked).................................................................228
Table 4-9 Mean preamble Duration (and standard deviation), in milliseconds, by Partner and Difficulty in Experiment 8a (randomized) and Experiment 8b (blocked).................................230
Table 4-10 Mean preamble Onset (and standard deviation), in milliseconds, by Partner and Difficulty in Experiment 8a (randomized) and in Experiment 8b (blocked).................................232
Table 4-11 Mean Gap (and standard deviation), in milliseconds, by Partner and Difficulty in Experiment 8a.................................................................233
Table 4-12 Mean Gap (and standard deviation), in milliseconds, by Partner and Difficulty and by Partner and Block in Experiment 8b.................................................................234
Table 4-13 Average standard deviations of Onset and Duration by Partner and Block in Experiment 8b.................................................................237
P. We're always talking Eliza.

H. Teaching Eliza.

P. Dressing Eliza.

MRS. H. What!

H. Inventing new Elizas.

H. and P., speaking together:

H. You know, she has the most extraordinary quickness of ear:
   P. I assure you, my dear Mrs. Higgins, that girl
      just like a parrot. I've tried her with every
      P. is a genius. She can play the piano quite beautifully
      H. possible sort of sound that a human being can make—
      P. We have taken her to classical concerts and to music
      H. Continental dialects, African dialects, Hottentot
      P. halls; and it's all the same to her: she plays everything
      H. clicks, things it took me years to get hold of; and
      P. she hears right off when she comes home, whether it's
      H. she picks them up like a shot, right away, as if she had
      P. Beethoven and Brahms or Lehár and Lionel Moricke; 
      H. been at it all her life.
      P. though six months ago, she'd never as much as touched a piano.

(G. B. Shaw, *Pygmalion*, Act III)

1. Introduction

In conversation, interlocutors coordinate their utterances with one another, in a
process that Clark (1996) has likened to playing a piano duet or waltzing together. This
thesis examines the cognitive mechanisms underlying the coordination of utterances. It
proposes that these mechanisms are closely related to language production mechanisms and
presents three studies that tested this hypothesis. In this chapter, we first define coordination,
and identify four distinct forms coordination can take when people “talk together” (Section
1.1). These are: synchronization of simultaneously produced utterances, repetition
(alignment), turn-taking, and completing one another’s utterances. All of these phenomena
are exemplified in (1), adapted from Lerner (1991, p. 450).
1. G: I mean I don’t (0.1s) that much. But he does. (0.4s) [1]  
S: Right [2]  
G: an’ it doesn’t matter at this point [3]  
G: I’ve waited this long I c’n (wait) [4]  
S: c’an wait another three wee:ks. [5]  

Synchronization, repetition, and joint completion are all involved in the transition between speaker G and speaker S at lines 4-5 in this extract; the transition from line 1 to line 2 illustrates an example of turn-taking where no overlap occurs, and a relatively short gap (0.4 seconds long) is present between the two speakers’ contributions.

Having distinguished between different types of coordination, we then look at three lines of research on the coordination of actions between people (Section 1.2) and at the mechanisms research in this field has identified as possible candidates for how coordination is achieved (Section 1.3). The review of studies on the coordination of actions will set the stage for a critical evaluation of theoretical proposals about the mechanisms interlocutors use to coordinate utterances (Section 1.4). In Section 1.4.3, we will present a new approach to this question, which suggests that interlocutors represent and anticipate others’ utterances using some of the mechanisms they use to produce language. Finally, in Section 1.5, we will describe the types of experimental paradigms that could be used to test the predictions of our account.

1.1 Talking together

The interactive use of language in conversation is a form of joint activity, in which individuals act together to achieve the common goal of communicative success. Clark (1996) proposed that conversation shares fundamental features with other joint activities, for example waltzing, playing a duet, or shaking hands. The most central, defining feature of all joint activities is coordination: the mutual process by which actors take into account the intentions and the (performed or to-be-performed) actions of their partners in the planning
and performance of their own actions (Clark, 1996, pp.61-62). Despite the recognition that co-actors need to coordinate both on content (the common intended goal) and on processes (“the physical and mental systems they recruit in carrying out those intentions”; Clark, 1996, p. 59) to succeed in a joint action, very little is in fact known about such processes.

This thesis is concerned with the processes used by language users to achieve coordination. In particular, the processes used in the coordination of utterances produced by two different speakers are the focus of the research presented here. Before discussing different theoretical approaches to this topic, and the available empirical evidence, it is necessary to clarify what is meant by coordination in this context. Coordination is the process by which the planning and articulation of one utterance (i.e., syllable, word, constituent, or longer stretch of language) affects the planning and articulation of a second utterance (i.e., another syllable, word, constituent, or longer stretch of language), and can in turn be affected by it. Therefore, we take coordination to be a very general process.

Indeed, because this definition focuses on utterances, rather than speakers, it admits cases in which the two coordinated utterances are produced by the same speaker (within-speaker coordination), as well as cases that involve two speakers (between-speaker coordination). Below, and throughout the thesis, we focus on between-speaker coordination (as this is the phenomenon we are trying to explain). However, we will often draw parallels with the coordination of utterances within the one speaker. In fact, this comparison will be the basis of our theoretical proposal (Section 1.4.3), and will inspire the tasks designed to test it (Chapters 2, 3, and 4).

Between-speaker coordination can take different forms. First, it can take the form of synchronization between two simultaneously produced utterances. In conversations, speakers rarely talk at the same time and, even when they do, they mostly resolve the overlap as quickly as possible (Sacks, Schegloff, & Jefferson, 1974; Schegloff, 2000). There are, however, instances in which overlapping speech appears to be the norm, rather than a problematic occurrence (e.g., choral greetings; Lerner, 2002). Interestingly, there is some
evidence that speakers who talk simultaneously coordinate their utterances. Cummins (2003, 2009) showed that when a speaker is asked to read a familiar text in time with another speaker, the gap between the two utterances is as short as 40 ms on average. Cummins argued that almost perfect synchronization in this task is the result of a mutual adjustment between the two speakers, rather than of one speaker closely following the other in a reactive manner. He also showed that one of the consequences of such mutual adjustment is a slower speech rate and reduced variability in the location of pauses in synchronized speech compared to speech produced in isolation.

Second, coordination can take the form of alignment (Pickering & Garrod, 2004). In conversation, language users continually switch between speaking and listening. A vast body of literature showed that listening to the utterances produced by another speaker affects subsequent production on several levels. Speakers tend to converge in the way they speak, for example in their speech rate (Street, 1984), as well as in many other subtle phonetic characteristics (see Gambi & Pickering, 2013a for a review). Speakers also tend to reuse words (Brennan & Clark, 1996; Garrod & Anderson, 1987), and syntactic structures (Branigan, Pickering, & Cleland, 2000; Levelt & Kelter, 1982), they have just heard from their interlocutor. These findings show that comprehended utterances can affect subsequently produced utterances. The way in which this happens is largely automatic. Pickering and Garrod (2004) argued that alignment occurs via a priming mechanism: representations that have just been built in comprehension are highly activated and therefore more likely to be used in production, given the assumption that comprehension and production share the same representations (i.e., parity of representations).

Third, coordination can take the form of smooth turn-taking. Speakers do not only avoid overlapping speech (Sacks, et al., 1974). They also avoid long silences between turns. This is reflected in the distribution of inter-turn intervals. In a recent survey of question-answer pairs extracted from natural conversations in 10 different languages, Stivers et al. (2009) showed that inter-turn intervals follow a unimodal distribution, with mode ranging
between 0 ms and 200 ms (and mean ranging from approximately 0 ms to around 500 ms), depending on the language. Note that people take 500-700 ms to start speaking in reaction to a cue, even if the material is pre-planned and only needs to be retrieved from memory (e.g., F. Ferreira, 1991). Thus, speakers must begin planning their contribution while their partner is still speaking. In addition, they must be able to time the planning and articulation of their contribution with the respect to their partner’s unfolding contribution.

Different authors have proposed different ways in which this could be achieved. The signal theory of turn taking (Duncan, 1972) emphasizes the role of cues produced by speakers, to which listeners merely react by adopting the appropriate behaviour (either starting their turn or letting the speaker continue). The oscillator model of turn-taking (M. Wilson & Wilson, 2005) assumes that the speaker’s and the listener’s production systems go through a periodic cycle, in which their readiness to speak varies between a minimum and a maximum with a period corresponding to the production of a syllable (100-150 ms); because the systems are synchronized in anti-phase, when one interlocutor is most likely to speak the other is least likely to speak. Finally, the projection theory of turn taking (Sacks, et al., 1974) proposes that a speaker’s turn “projects” its own completion, thus allowing listeners to anticipate the turn’s ending and take the floor with at most a very small gap or overlap (see Heldner & Edlund, 2010 for a critical discussion).

Many linguistic (e.g., pitch contour) and non-linguistic (e.g., breathing) cues are reliably associated with turn-holding or turn-yielding points in a conversation. However, very few studies have systematically investigated which features of the speech signal are actually exploited by listeners to discriminate between end-of-turn and turn-holding points (see Gravano & Hirschberg, 2011; Hjalmarsson, 2011) and whether listeners are able to use such cues on-line to anticipate turn-endings (De Ruiter, Mitterer, & Enfield, 2006; Grosjean, 1996; Magyari & De Ruiter, 2008, 2012). Overall, the results of the latter set of studies suggest that listeners rely on lexical and syntactic information (more than on prosodic information) to judge when the speaker’s turn will end.
Fourth, coordination can take the form of cross-person compound contributions (Howes, Purver, Healey, Mills, & Gregoromichelaki, 2011). These are instances in which a speaker either completes (2) or continues (3) a previous contribution by another speaker.

2.  
   A: so if one person said he couldn’t invest (0.1 sec)  
   B: then I’d have to wait  
   (Lerner, 1991, p. 445)

3.  
   A: first of all they hit rain then they hit hail  
   B: then they hit snow  
   (Lerner, 1991, p. 448)

The existence of this phenomenon suggests that speakers can not only reuse parts of a previously heard utterance (as in alignment), but also build grammatically and pragmatically appropriate continuations for an utterance that was initiated by another speaker. Cross-person contributions are fairly common, accounting for 3% of all dialogue contributions in the British National Corpus (Howes, et al., 2011). They are often associated with repairs and expansions of already complete antecedents, and are more likely when the antecedent ends with an unfilled than with a filled pause (Howes, et al., 2011). However, cross-person contributions do not occur only between major linguistic constituents (sentences or clauses), but also within constituents (Helasvuo, 2004; Howes, et al., 2011), suggesting that speakers process their partner’s utterances in a highly incremental fashion. In support of this, an experiment in which dialogue contributions were artificially truncated showed that interlocutors were sensitive to lexical, syntactic, and pragmatic predictability, as these factors affected their tendency to provide continuations for their partner’s utterances (Howes, Healey, Purver, & Eshghi, 2012).

To summarize, the four phenomena described above (synchronous speech, alignment, turn taking, and compound contributions) illustrate four different ways in which speakers can be said to coordinate their utterances in natural conversations. The nature of the coordination (i.e., the way in which the utterances affect one another) is not necessarily the same in all cases. For example, in synchronous speech, fast, continuous, and reciprocal
adjustments are in order. Instead, alignment is supported by discrete priming events, in which an utterance produced by the current speaker primes the subsequent production by a similar utterance by the current listener. Although these priming events take place over relatively short time scales (Pickering & Ferreira, 2008), their effect can cumulate over the course of a conversation. Turn taking and compound contributions, in addition, might require ways of integrating representations of one’s own and another’s utterances. The next section looks at how coordination has been studied in the domain of action, and then presents three theoretical approaches that make different claims regarding the processes involved in coordination. These approaches are: the dynamical systems approach; the shared representations approach; the internal simulation approach.

Note that our list of coordination phenomena in dialogue leaves out one phenomenon that has been extensively studied: the coordination of referring expressions (Brennan, Galati, & Kuhlen, 2010; Brown-Schmidt, 2009; H. H. Clark & Wilkes-Gibbs, 1986; Garrod & Anderson, 1987). This is another way speakers coordinate their utterances in natural conversations, and several studies investigated the way in which such coordination grounds and reveals the development of shared conceptual representations. A related line of research also examined how people coordinate in the absence of a conventional communication system (i.e., a language), using non-verbal referential games (e.g., De Ruiter et al., 2010; see Galantucci & Garrod, 2011 for a review).

The studies on referential communication, however, will not be reviewed in this thesis, because our focus is on the coordination of utterances rather than on the coordination of meaning. We will, however, briefly refer to theories of dialogue (see Section 1.4) that have been informed by the results of such studies; we will do this mainly to point out that these theoretical approaches say very little about the mechanisms that allow speakers to coordinate their utterances (see Gambi & Pickering, 2011). To the contrary, the mechanisms that underlie coordination between individuals have been the focus of at least three lines of research in the domain of action coordination. Therefore, we will start by reviewing this
literature in the next section, in order to ground a discussion of the different theoretical proposals put forward in this domain (Section 1.3).

1.2 Acting together

This section reviews studies that have looked at the coordination of actions (and movements) other than speech. Some studies looked at the synchronization of movements between people and are reviewed in Section 1.2.1 (“Synchronization”). Other studies investigated whether another person’s task and her/his actions are represented when two people take turns performing unrelated tasks; these studies are reviewed in Section 1.2.2 (“Taking turns”). Finally, a few studies investigated whether people integrate representations of their own and other’s actions when performing complementary actions directed towards a common goal; these studies are reviewed in Section 1.2.3 (“Complementary actions”).

1.2.1 Synchronization

Several studies investigated under which conditions two participants come to synchronize (or entrain) their bodily movements. This line of research on interpersonal entrainment (also called “coordination”; Marsh, Richardson, & Schmidt, 2009; Riley, Richardson, Shockley, & Ramenzoni, 2011; Schmidt & Richardson, 2008; Shockley, Richardson, & Dale, 2009) is based on the idea that the entrainment of movements between people follows similar principles to the entrainment of movements (of different limbs) within one person.

1.2.1.1 From intrapersonal to interpersonal synchronization

Haken and colleagues showed that only two stable modes are observed in the synchronization of bimanual rhythmic movements (Haken, Kelso, & Bunz, 1985): in the asymmetric phase mode, the two hands are at opposite points of their movement cycle at the
same time (i.e., the relative phase angle between their movement cycles is 180 degrees); in the symmetric phase mode, instead, the two hands occupy the same point in the movement cycle at the same time (i.e., the relative phase angle between their movement cycles is 0 degrees). However, when movement frequency increases, the asymmetric phase mode becomes progressively unstable, until only the symmetric mode is observed.

Crucially, similar properties are exhibited by a system composed of two visually coupled people (i.e., when the two entrained limbs belong to different people who can see each other’s movements), as shown by Schmidt and colleagues (Schmidt, Carello, & Turvey, 1990). However, a system composed of two people appears to be less stable overall than a system composed of two limbs belonging to one and the same person: entrainment breakdowns are more likely in the former (especially for higher frequencies) and the variability in relative phase angles is also larger between than within people (Schmidt, Bienvenu, Fitzpatrick, & Amazeen, 1998).

In addition, the strength of entrainment between people appears to be affected by several variables, including differences in the preferred period of the two participants’ movement cycles (e.g., Schmidt & O'Brien, 1997), and the availability and nature of perceptual information (see Schmidt & Richardson, 2008 for a discussion). For example, one study showed that participants who conversed to solve a puzzle task while swinging handheld pendulums became entrained only when they could see each other’s movements, but not when they could listen to each other (M. J. Richardson, Marsh, & Schmidt, 2005). The importance of visual information in mediating entrainment strength was confirmed by another study in which participants sat on rocking chairs (M. J. Richardson, Marsh, Isenhower, Goodman, & Schmidt, 2007).

Interestingly, however, verbal information was also found to affect the degree to which participants became entrained in other studies. Using cross-recurrence quantification analysis to measure the similarities between participants’ postural sway trajectories over time, Shockley, Santana, and Fowler (2003) showed that people conversing with each other
to solve a puzzle task entrained their postural movements (even when they were facing away from each other) more than people who did not converse with each other (but were talking to confederates and solved the task with them). In addition, a subsequent study investigated the effect of speaking patterns on the entrainment of postural sway (Shockley, Baker, Richardson, & Fowler, 2007). Participants read words in phase or anti-phase (i.e., taking turns), at different speaking rates, while the similarity between words was also manipulated (they either read the same words, different words with the same stress pattern or different words with different stress patterns). The results suggested that the more words participants articulated and the more similar those words, the stronger the entrainment of their postural sway.

In addition to perceptual information, psychological variables appear to affect entrainment strength as well, although comparatively less work has been conducted in this area (Schmidt & Richardson, 2008). First, entrainment is observed both when participants are explicitly instructed to maintain a particular phase mode, and when they are not, but unintentional entrainment is weaker (M. J. Richardson, Marsh, Isenhower, et al., 2007; Schmidt & O'Brien, 1997; Schmidt & Richardson, 2008).

Second, in a seminal study (Schmidt, Christianson, Carello, & Baron, 1994), social competence and social dominance were shown to affect entrainment patterns. Pairs of low social competence participants as well as pairs of high social competence participants exhibited less stable dynamics than mixed social competence pairs (in which one participant had low social competence and the other has high social competence). According to Schmidt and Richardson (2008), these findings suggest that asymmetrical relationships facilitate entrainment, perhaps because one, less variable participant leads and the other, more variable participant follows (see also Bosga, Meulenbroek, & Cuijpers, 2010).

Third, positive attitudes towards the other participant and towards the interaction were associated with stronger entrainment in some studies (reviewed in Marsh, et al., 2009; Schmidt & Richardson, 2008). Further, pairs of walkers moving in phase or anti-phase were
judged by naïve observers as showing higher rapport and connectedness than non-entrained pairs (Miles, Nind, & Macrae, 2009). In addition, naïve participants entrained (i.e., synchronized in the in-phase mode) more with a confederate that arrived on time for the experiment than with a confederate who arrived late, and their ratings of rapport (collected after the experiment) were also correlated with the degree of entrainment (Miles, Griffiths, Richardson, & Macrae, 2010). However, in another study (Miles, Lumsden, Richardson, & Macrae, 2011) participants were shown to exhibit stronger in-phase entrainment with a confederate who was portrayed as belonging to a different than to the same group (with group membership being defined in terms of art preferences). Finally, young children entrained more with an adult experimenter (who was drumming at a constant beat) than with a drumming machine or a recording, suggesting that social context facilitates in-phase entrainment (Kirschner & Tomasello, 2009).

1.2.1.2 Symmetric and asymmetric entrainment patterns

Above, we mentioned (with reference to Schmidt et al., 1994) that interpersonal synchronization might sometimes exhibit an asymmetric entrainment pattern, in which one person (the follower) adjusts unilaterally to the other (the leader). It is not clear whether symmetric entrainment patterns (with continual mutual adjustments by both participants) or asymmetric entrainment patterns are more common and/or easier to maintain. In fact, a recent study of finger tapping showed that participants are worse at synchronizing their taps when only one of them can hear the other than when they can both hear each other. This suggest that a leader-follower strategy is potentially less efficient than a mutual adaptation strategy, at least when the participant acting as leader shows variable inter-tap intervals (Konvalinka, Vuust, Roepstorff, & Frith, 2010). This suggests that the degree of variability might affect which type of entrainment pattern (symmetric or asymmetric) eventually emerges.
Interestingly, a reduction of variability in action timing was shown to facilitate coordination in a study where two participants performed a Simon task alongside one another (Vesper, van der Wel, Knoblich, & Sebanz, 2011). In this task, participants responded to red and green stimuli that appeared on the left or on the right-hand side of the screen. The location of the stimuli was irrelevant. Participants responded to one colour by pressing a button with their right hand and to the other colour by pressing a button with their left hand. The basic finding in the Simon task is that responses are faster when stimulus and response location match than when they mismatch.

Crucially, this study (Experiment 1) showed that the timing of button presses was less variable when participants were asked to coordinate with the button presses produced by another person to achieve “in-phase” synchronization (i.e., responding at the same time as the other person) than when they were doing the task on their own. Note that the task tested in this experiment was different from the other studies reviewed in this section, as participants synchronized one-shot non-rhythmic movements that were goal-directed (Vesper, et al., 2011). Interestingly, the reduction in variability (which was highly correlated with the participants’ tendency to speed up their responses when they acted alongside another person) led to a reduction in asynchrony only when participants were explicitly instructed to synchronize.

A more recent study replicated this finding in different tasks, and showed that this reduction in variability occurs more the less information participants have access to about the actions of their interaction partner (Vesper, Schmitz, Sebanz, & Knoblich, 2013). The authors interpreted this finding as suggesting that variability reduction could be a strategy used by participants to coordinate when they cannot monitor or anticipate their partners’ actions. Conversely, a study of joint tapping in musicians (Pecenka & Keller, 2011) showed that, when participants had access to their partner’s performance (though only in the form of continuous auditory, but not visual, feedback), pairs of participants with high predictive abilities (i.e., with a tendency to anticipate, rather than track the beat) achieved more
accurate and less variable entrainment than pairs with low predictive abilities (with mixed pairs achieving intermediate performance).

### 1.2.2 Taking turns

Another line of research investigated action coordination using paradigms in which two participants (let’s call them A and B) take turns performing actions (usually, button presses) in response to shared (auditory or visual) stimuli, according to pre-specified task instructions (Knoblich, Butterfill, & Sebanz, 2011). The instructions are delivered to both participants, who therefore have knowledge of their partner’s as well of their own task. A task in this context is defined as a series of stimulus-response mappings: these include a specification of GO stimuli (stimuli that require a response from a given participant) and NO-GO stimuli (stimuli that do not require a response from a given participant, but might require one from the participant’s partner), and might also include a specification of which response (e.g., right or left button press) should be given to each type of GO stimulus.

Participants in these studies do not interact with one another. Rather, they act alongside one another. In most studies, there is no overlap between the GO stimulus set for participant A and the GO stimulus set for participant B. Therefore, A and B never act concurrently; rather, they take turns. Since the stimuli are presented in randomized sequences, the participants do not necessarily alternate (i.e., they do not necessarily follow an ABABAB sequence). Some studies also required participants to act concurrently on some of the trials (De Bruijn, Mars, Bekkering, & Coles, 2012; De Bruijn, Miedl, & Bekkering, 2008, 2011; Sebanz, Knoblich, & Prinz, 2005; Wenke et al., 2011).

In the first study of this kind, Sebanz, Knoblich, and Prinz (2003) asked participants to perform a spatial compatibility task either individually or jointly. In their task, stimuli were photographs of a human hand, pointing either to the right or to the left of the screen. In the photographs, the index finger showed a ring that was either red or green in colour. Participants had to respond to the ring colour, while ignoring the pointing direction. In one
condition, participants performed the task alone and responded to both colours (e.g., they responded to green stimuli with a right button press and to red stimuli with a left button press). Let us call this the *solo* condition.

The solo condition is equivalent to the classic Simon task, which we already introduced in Section 1.2.1. This task leads to S-R (Stimulus-Response) compatibility effects: Despite the fact that the location of the stimulus is completely irrelevant, respondents are slower when the spatial features of the stimulus and the response mismatch (i.e., when they respond to left stimuli with a right button press or, vice versa, when they respond to right stimuli with a left button press), presumably because the spatial features associated with the stimulus are automatically activated.

In another condition of Sebanz et al.’s (2003) study, participants also performed the task alone, but this time they only responded to one colour with one hand. In other words, they performed a go/no-go version of the solo task. The authors referred to this as the *individual go/no-go* condition. Importantly, spatial SR compatibility effects are greatly reduced (e.g., Sebanz, et al., 2005), or absent (e.g., Sebanz, et al., 2003) in this go/no-go version. In a third and final condition, participants were tested in pairs. They were seated on either sides of a computer screen, one next to the other. Each participant again responded to only one colour (with one hand), but this time the other participant responded to the other colour, so that the task was distributed between two people, each performing complementary halves of it (i.e., each performing complementary go/no-go tasks). Let us call this the *joint* task.

Crucially, Sebanz et al. (2003) found a compatibility effect in the joint task (which they therefore termed a joint compatibility effect). This finding was subsequently replicated in several studies, and it is often referred to as the Social Simon Effect (SSE) or the Joint Simon Effect (JSE) (Colzato, De Bruijn, & Hommel, 2012; Constantini & Ferri, 2013; Dittrich, Rothe, & Klauer, 2012; Ferraro, Iani, Mariani, Milanese, & Rubichi, 2011; Guagnano, Rusconi, & Umiltà, 2010; Hommel, Colzato, & van den Wildenberg, 2009;
Moreover, joint effects occur with different tasks, including: an auditory version of the standard Simon task (Dolk et al., 2011; Dolk, Hommel, Prinz, & Liepelt, in press; Dolk, Liepelt, Prinz, & Fiehler, 2013; Ruys & Aarts, 2010; Vlainic, Liepelt, Colzato, Prinz, & Hommel, 2010), the SNARC task (Atmaca, Sebanz, Prinz, & Knoblich, 2008), and the Flanker task (Atmaca, Sebanz, & Knoblich, 2011). Joint tasks have also been used to investigate how another’s perspective (Böckler, Knoblich, & Sebanz, 2011) and focus of attention (as defined in the Navon task; Böckler, Knoblich, & Sebanz, 2012; Böckler & Sebanz, 2012) can affect the processing of jointly attended stimuli.

1.2.2.1 Accounts of joint task effects: co-representation versus referential spatial coding

Despite this vast body of evidence, it is still unclear what mechanisms underlie the observed joint effects. Below, we describe three competing accounts that have been put forward in the literature, and briefly review evidence supporting each of them (against the others). The aim of this review is to show that, although some of the findings can be explained without any reference to the joint nature of these effects (i.e., the JSE is not social at all; Dolk et al. 2011), several studies support the idea that joint task effects do indeed reflect the workings of a mechanism that might support the coordination of actions between people.
The latter position is primarily reflected in the leading account of joint task effects, which is known as the *co-representation* account (Knoblich, et al., 2011; Sebanz, et al., 2005). According to this account, joint effects arise because participants in joint tasks automatically represent their partner’s task (set of S-R mappings) as well as their own. Therefore, properties of the stimulus that evoke the partner’s response activate that response. If the activated response is incompatible with the participant’s own response, the result is interference and the participant’s response is slowed down. For example, when the participant sits on the right-hand side of the screen in a joint Simon task, a stimulus pointing to the left will evoke the partner’s response. Since the partner is sitting on the left, his response is coded as a “left” response. The participant’s own response, instead, is coded as a “right” response, and interference occurs between these competing spatial codes.

Crucially, the co-representation account of joint effects assumes that interference is caused by conflict at the level of response (action) selection, which arises because co-actors represent one another’s responses. Such co-representation could be one of the mechanisms underlying coordination in joint actions (see Section 1.3.2). However, according to the *referential spatial coding* account, there is no need to assume that participants represent their partner’s responses. Proponents of this account suggested that the JSE is in fact due to the participants using the presence of a partner to define a spatial frame of reference. In other words, it is the mere presence of a salient stimulus (e.g., another actor) in the environment that causes participants to code their own responses as “right” or “left”.

This account is supported by different sources of evidence. For example, one study manipulated the saliency of the right-left (horizontal) spatial dimension by varying whether the participants responded to horizontally (high saliency) or vertically (low saliency) arranged stimuli by pressing buttons that were also horizontally or vertically arranged (Dittrich, et al., 2012). Interestingly, they found a JSE only in the high saliency condition, suggesting that referential coding might indeed underlie the JSE. Further evidence in support of this conclusion comes from the fact that the JSE is absent when the partner sits outside of
the participant’s peripersonal space (i.e., outwith arm-reach; Guagnano, et al., 2010).

However, note that a recent study failed to replicate this finding (Welsh, et al., 2013).

Some of the most compelling evidence in favour of the referential spatial coding account comes from Dolk and colleagues. Dolk et al. (2011) combined a joint (auditory) Simon task with the rubber-hand illusion. In their version of the rubber-hand illusion, the participant’s resting hand received tactile stimulation while the partner’s corresponding hand was also being stroked, either in-synch or out-of-synch (with respect to the strokes delivered to the participants’ hand). Synchronous stroking is usually associated with an illusionary sense of ownership of the stroked hand. Therefore, synchronous stroking should induce participants to incorporate the partner’s hand into a representation of their own body schema, and increase the magnitude of the JSE.

Instead, they found that synchronous stroking reduced the magnitude of the JSE compared to asynchronous stroking (Experiment 1). In addition, they showed that the JSE was present also when the participants were tested individually, but the presence of the “stroking” device used to induce the illusion could still provide a frame for referential coding (Experiment 3). Furthermore, a recent series of experiments (Dolk, et al., in press) demonstrates that any kind of salient auditory event that attracts attention can induce referential coding, and a JSE, irrespective of whether it is produced by a moving agent, a moving object (a Chinese cat), or a non-moving object (clock, metronome).

But importantly, there is also evidence in favour of the co-representation account (and against the referential spatial coding account). First, one study (Welsh, 2009) asked participants to respond with their hands crossed, so that the participant sitting on the right would press a button situated on the left, in front of her partner (and vice versa for the partner), with either her right or left hand (see also Liepelt, et al., 2013). Responses were slower when the location of the responding hand was not compatible with the location of the stimulus (regardless of the anatomical position of the hand used to respond). In other words, the participant did not code her partner’s responses as “left” because her partner was sitting
on the left-hand side of the screen. Rather, the right participant coded her partner’s responses as “right” when her partner’s responding hand was situated on the right side of the screen, thus indicating that the spatial code of the partner’s response was represented and not simply the location of the partner in space with respect to the participant. (The absence of an anatomical response hand effect also suggests that the JSE is not due to a faster perception-action link between ipsilateral effectors and visual fields.)

In addition, there are independent pieces of evidence supporting the co-representation account (see also Welsh, et al., 2013). One source of evidence comes from studies that found joint interference effects when partners were seated in the same room but could not perceive each other (Sebanz, et al., 2005; Vlainic, et al., 2010), or when participants sat in a room on their own but were led to believe another person was performing the task with them. The latter finding occurred both when the participant obtained feedback while carrying out the task (i.e., a stimulus on the participant's screen signalled when the believed partner produced a response; see Ruys & Aarts, 2010; Tsai, et al., 2008; but see Welsh, et al., 2007 for a failure to find the JSE in a similar condition) and when no feedback was available (Atmaca, et al., 2011). This set of findings is difficult to explain in terms of referential coding, unless one assumes that the mere belief that another person is taking part in the task in a different room is sufficient to induce referential coding. But such belief, at least in the absence of any feedback (Atmaca, et al., 2011; see also Chapter 2 in this thesis) can hardly be considered salient.

Another source of evidence for the co-representation account is provided by the fact that the presence of joint effects is modulated by the participants’ beliefs and attitudes towards their partners. For example, the JSE is not present when participants believe their partner is not a human agent, but a computer (Tsai, et al., 2008), or when they take turns with a virtual model of a wooden hand (Tsai & Brass, 2007); however, if the animacy of the wooden hand is emphasized (by asking the participants to watch a video of Pinocchio, and to identify with the main character), then a JSE effect can be induced for the wooden hand as
well (Mueller, Brass, et al., 2011; see Tsai, Knoblich, et al., 2011, Experiment 1, for a related finding).

Further, the nature of the relationship between participants also affects the presence of joint task effects. One study found that participants who were paired with a friendly and cooperative confederate showed the JSE, whereas participants who were paired with an intimidating and competitive confederate did not (Hommel, et al., 2009). A subsequent study showed that joint effects are more pronounced in situations that induce interdependence between the participants in a pair, that is when each individual’s success in pursuing a personal goal depends on the actions of the other member of the pair (Ruys & Aarts, 2010). In addition, participants who were better at decoding intentions from pictures of other people’s eyes showed a joint effect even in the condition where their success did not depend on their partner’s actions (Ruys & Aarts, 2010). Further, patients with acquired deficits on ToM (Theory of Mind) abilities also show no JSE, at least when they are not explicitly asked to direct attention to what their partner is doing (Humphreys & Bedford, 2011).

Interestingly, putting participants in direct competition with their partner reduces the magnitude of the JSE (Iani, et al., 2011), and priming their independence from the social context (by having them circle first-person singular pronouns in a text) also reduces the JSE (Colzato, et al., 2012). Overall, these findings strongly suggest that the emergence of joint task effects is associated to the participants’ ability to attend to their partners’ actions and, furthermore, to perceive their own actions as somehow dependent on their partners’ actions.

These tendencies might be more pronounced when interacting with a partner who belongs to the same group. However, group membership (i.e., performing a joint task with a member of the same “cognitive style” group versus a member of a different group) did not have any effect in Iani et al.’s (2011) study. This partially contrasts with the findings of another study that manipulated group membership more effectively. White participants from the Netherlands showed a joint compatibility effect only when they interacted with a virtual partner from the same group (i.e., with a white hand), but not when they interacted with a
virtual partner from a different group (i.e., with a black hand; Mueller, Kuhn, et al., 2011). Nevertheless, this difference was eliminated when white participants read a passage about an out-group character, and they were asked to identify with him before performing the joint Simon task (Mueller, Kuhn, et al., 2011). Finally, making the participant feel ostracized (in a virtual ball tossing game) also seems to eliminate the JSE (Constantini & Ferri, 2013).

Overall, these findings are more consistent with the co-representation account than with the referential spatial coding account of joint task effects as they show that the effect is sometimes absent when another participant (or virtual co-actor) is present. What appears to be crucial, then, is not the presence of a salient entity that can function as a landmark, but rather the fact that this entity is perceived as an intentional agent whose responses can be represented in a way that is similar to how the participant’s own responses are represented. This could be linked to the participants perceiving the co-actor as sufficiently similar to herself, or to the extent to which her attitude towards her co-actor is positive.

1.2.2.2 The neural correlates of co-representation

Further evidence for the co-representation account is provided by ERP studies that show increased response inhibition demands on NO-GO trials in joint tasks than in individual go/no-go tasks. This strongly suggests that participants in joint tasks represent their partner’s actions on NO-GO trials, and need to apply a higher level of inhibition (indexed by an enhanced no-go P3 component) to avoid responding overtly when it is their partner’s turn to respond (Sebanz, Knoblich, et al., 2006). Tsai, et al. (2006) independently replicated this finding, and showed that the nogo-P3 amplitude was larger in the joint condition than in an individual condition with a co-present but inactive participant. In addition, they found that the amplitude of the nogo-P3 was modulated by stimulus compatibility (larger on incompatible trials) only in the joint but not in the individual condition. The latter finding was also replicated by Tsai, et al. (2008), who showed that it
occurred only when the participants believed they were interacting with another participant, but not when they believed they were interacting with a computer.

In addition, in Sebanz et al.’s (2007) fMRI study, NO-GO trials in the joint condition (compared to an individual condition in which the partner was present but not active) showed increased activity in the SMA (Supplementary Motor Area), which is implicated in the execution of motor responses (see e.g., Mostofsky & Simmonds, 2008). Moreover, de Bruijn, Miedl, and Bekkering (2008) showed that no-go P3 amplitudes were reduced when a co-actor in a competitive task was responding compared to when the co-actor was not responding (so both actors had to inhibit a response), although this was the case only for slow responders (i.e., participants who were on average slower than their partner). Finally, de Bruijn, Miedl, and Bekkering (2011) analyzed erroneous responses on NO-GO trials and showed that the error-related negativity (Ne/ERN) was reduced when the trial required a response from the co-actor (but not from the participant) than when it required no response from both the co-actor and the participant. This further suggests that participants represent that their partner is responding on NO-GO trials.

However, it is not clear to what extent the specific content of the partner’s response is represented on NO-GO trials. This issue has been investigated by looking at the amplitude of the LRP (Lateralized Readiness Potential), which reflects the preparation of motor responses (and is specific to the side of the body the participant is preparing to move). Tsai et al. (2006) reported a modulation of the LRP (in the 100-200 ms time window) on NO-GO trials (more negative in the joint condition), but only for compatible trials. Instead, Tsai et al. (2008) reported that in the same early time-window the LRP was more positive-going when the participants believed they were interacting with another human agent than when the participants believed they were interacting with a computer, but with no effect of compatibility. However, in a later time window (200-400 ms), the amplitude of the LRP was significantly affected by compatibility only when participants believed they were interacting with another human agent.
More recently, Holländer, Jung, and Prinz (2011) showed that a reliable LRP (more negative than baseline) was observed on NO-GO trials only in a joint but not in an individual go-no task; interestingly, such LRP was always observed (in the non-responding participant) in the hand closer to the responding partner, irrespective of whether the partner used the right or the left hand to respond. According to the authors, this suggests that co-representation uses an egocentric frame of reference. In addition, it might suggest that representations of others’ responses are not effector-specific. Therefore, it is possible that co-actors represent whether their partner is responding, but not necessarily the specific content of their response (see below).

Another potential issue for the co-representation account is that the neural correlates of acting in joint tasks are not very well understood when it comes to GO trials. Sebanz et al. (2006) reported that the amplitude of the go P3 was affected by compatibility only in the individual, but not in the joint condition (and the effects of condition were present only in the later portion of the component, around 400-600 ms). On the contrary, Tsai et al. (Tsai, et al., 2006; Tsai, et al., 2008) reported that compatibility affected the amplitude on the go P3 more in the joint than in the individual (or acting with a computer) condition, but only at specific electrode sites (mainly Pz).

In addition, larger P3 amplitudes were observed on incompatible than compatible trials, which contradicts previous findings using the solo Simon task (see Tsai, et al., 2008, p. 2022 for discussion). Finally, the fMRI study conducted by Sebanz et al. (2007) showed no increased activation of areas involved in cognitive control (e.g., ACC, anterior cingulate cortex) in the joint compared to the individual task. The latter would have been expected if participants need to manage conflict between their own and their partner’s response in the joint task.
1.2.2.3 Accounts of joint task effects: co-representation versus agent conflict

Given the above-mentioned issues with the co-representation account, a third account of joint compatibility effects has been recently proposed. According to the agent-conflict account (Wenke, et al., 2011), joint compatibility effects are due to agent conflict rather than to response conflict. In other words, it is the fact that one’s partner might (potentially) be responding on the current trial that matters, not the nature of one’s partner’s response. The account predicts that double-response trials (i.e., trials that evoke a response from both co-actors) should show longer latencies than single-response trials (i.e., trials that evoke a response only from one co-actor). This should occur irrespective of whether a compatible response or an incompatible response is evoked.

Accordingly, Sebanz et al. (2005) found that participants took longer to respond on trials on which their partner acted as well (performing a different task), than on trials on which their partner did not act. Importantly, this occurred also on compatible trials (i.e., when the stimulus was pointing in the participant’s direction). The authors interpreted this as evidence that participants co-represented their partner’s task (e.g., “respond to colour” vs. “respond to direction” of stimulus), as well as their response (e.g., “press the right” vs. “press the left” button), and that conflict between co-activated task-representations then causes interference at the level of response selection. However, this finding can also be interpreted as suggesting that it does not matter whether the stimulus evokes a spatially compatible or incompatible code, as long as it requires a response from the other participant (i.e., it signals that it is the other’s turn to respond).

In addition, Philipp and Prinz’s (2010) results also speak in favour of the agent-conflict account. They reported joint agent-face compatibility in an experiment in which speakers had to respond to coloured shapes by uttering their own or their partner’s name. Together with the target shape, they were shown a picture of their own or their partner’s face (as task irrelevant distractors). Responses were faster when participants saw their own face
than when they saw their partner’s face, regardless of which name they used. Participants in this study appear to have interpreted the pictured face as a cue to whose turn it was to respond (see Wenke et al., 2011, p. 165).

However, responding at the same time as one’s partner does not always lead to agent-conflict. In one study, pairs of people (one participant and a confederate) watched either one or two actors perform finger-lifting movements. The number of movements in the stimuli either matched or mismatched the number of movements the pairs were asked to perform. When the participants were watching a single actor, no compatibility effect was observed. However, a compatibility effect was found when they were watching two actors. They were faster to respond with a single (individual) movement when only one of the two actors on screen executed a movement than when both did and, conversely, they were faster to respond with two movements (i.e., to respond at the same time as their partner) when both actors on screen executed a movement than when only one did. Crucially, they were not overall slower when they responded at the same time as their partner than when they responded on their own (Tsai, Sebanz, & Knoblich, 2011).

To summarize, many studies investigated whether participants take into account their partners’ responses in tasks where they do not interact but simply take turns responding (sometimes responding at the same time). They showed that they do, provided their partners are intentional agents, they are encouraged to attend to their partners’ (real or imagined) responses, and they perceive their own and their partners’ responses to be dependent on each other. However, it is not clear whether participants indeed represent their partners’ responses in a similar way to their own responses (thus leading to increased response conflict and inhibition demands), or whether they represent whether it is their partner’s turn to respond, but not what their response would be. We will return to this issue in Chapter 2, in which we will present the results of four experiments that investigated this issue in a joint picture naming task.
1.2.3 Complementary actions

Complementary actions, as defined here, are actions that complement each other in the sense that they are directed towards a joint goal. Importantly, this goal is not simple synchrony (either in-phase or anti-phase), which distinguishes complementary actions from cases in which two actions or movements are intentionally synchronized (see Section 1.2.1). Also importantly, this goal is defined in terms of a distal event (i.e., something that occurs in the world), and is therefore separate from the (inherent) goal of coordinating smoothly (e.g., taking turns), which distinguishes it from all coordinated actions reviewed in Section 1.2.2.

For example, in one study (Knoblich & Jordan, 2003), pairs of participants attempted to keep a circle aligned with a moving dot on a computer screen. In each pair, one participant could only accelerate the tracker to the right, while the other could only accelerate it to the left. To succeed at this task, good timing is essential. For example, the right participant should anticipate when the left participant is going to start accelerating in order to avoid overlap and interference. Knoblich and Jordan (2003) found that performance in the joint task improved with practice, and eventually became as good as when an individual participant was controlling the tracker’s velocity in both directions. They concluded that the two co-actors had learned to predict each other’s button presses and could plan their own button presses accordingly in advance.

Interestingly, another study in which participants moved a rotating device so as to align a mark with a target showed that pairs doing the task jointly were actually faster than individuals, perhaps thanks to the (in this case spontaneous) development of a strategy in which one participant mainly accelerated and the other one mainly decelerated (Reed et al., 2006). Yet more studies implemented conceptually similar tasks, in which pairs of participants were asked to coordinate their actions in order to: move a ball through a jointly held wooden labyrinth (Valdesolo, Ouyang, & DeSteno, 2010); move a pole between two targets (van der Wel, Knoblich, & Sebanz, 2011); keep one participant’s index finger within a circular target held by the other participant (Ramenzoni, Riley, Shockley, & Baker, 2012);
lift planks of increasing length from a conveyor belt (M. J. Richardson, Marsh, & Baron, 2007); land at the same time after taking a jump (Vesper, Van der Wel, Knoblich, & Sebanz, 2013); build a ball track together (Vesper, Soutschek, & Schubö, 2009); lift and balance a virtual bar (Bosga & Meulenbroek, 2007); transfer a cylinder from one location to another (Meulenbroek, Bosga, Hulstijn, & Miedl, 2007); pass a jug to a confederate (Ray & Welsh, 2011).

All of the above-mentioned studies imply to some extent that participants performing complementary actions integrate their partner’s actions into the planning and execution of their own actions. However, they also reached different conclusions with regard to the mechanisms that allowed participants to perform such integration. Some authors claim the major determinant of joint task performance is perceptual sensitivity to the partner’s movements (Valdesolo, et al., 2010). Others suggest that participants might use sensory channels existing between them to convey information that is useful for coordination, in a form of “implicit” communication (Reed, et al., 2006; van der Wel, et al., 2011).

Other researchers suggested that integration in complementary tasks takes a deeper form. Participants in a dyad might come to “act as one” in the sense that the properties exhibited by their coordinated actions are similar to the properties exhibited by the coordination of complementary actions within an individual (M. J. Richardson, Marsh, & Baron, 2007; Vesper, Van der Wel, et al., 2013). One possibility is that in both cases participants make use of prediction mechanisms that allow them to anticipate the likely consequences of both their own and their partner’s actions (Knoblich & Jordan, 2003; Ray & Welsh, 2011; Vesper, Van der Wel, et al., 2013). But it is also possible that members of a dyad become integrated into a higher order unit (a synergy) that constraints their movements (Bosga & Meulenbroek, 2007; Ramenzoni, et al., 2012; Riley, et al., 2011). We return to these issues in Section 1.3, when we discuss different theories of acting together. In the next section we review neurophysiological studies of complementary actions.
1.2.3.1 Complementary actions in the brain

Kourtis et al. (Kourtis, Knoblich, & Sebanz, 2013; Kourtis, Sebanz, & Knoblich, 2010, 2013) compared action observation and anticipation when the action was performed by a confederate who did not interact with the participant and when it was performed by a confederate (or a real participant) who interacted with the participant, and performed complementary giving and receiving actions on some trials. They found that two EEG correlates of motor preparation (increase in the late Contingent Negative Variation, or CNV, and decrease in the beta rhythm, between 15 and 25 Hz) were more pronounced when participants anticipated an action performed by their interaction partner than when they anticipated an action performed by an individual that did not take part in the interaction (Kourtis, et al., 2010). In addition, in a subsequent study, increased CNV amplitude was only observed when the participants performed the role of receiver, and not when they performed the role of giver, perhaps because participants actively anticipated their partner’s action only in the former case (Kourtis, Knoblich, et al., 2013).

Importantly, the authors also compared EEG activation when the participants were preparing to perform an individual action with activation when the participants were preparing to perform a joint action (in which they took the receiver or giver role). They found that the P3a and (what they termed) medial P3b components were more pronounced when planning joint than individual action, suggesting increased attentional (P3a) and perhaps task representation (medial P3b) demands in the joint action condition (Kourtis, Sebanz, et al., 2013). They found no difference in the amplitude of the late CNV; interestingly, however, this component peaked roughly at the same time in all conditions, possibly suggesting that even when the participants were preparing a receiving action they first anticipated their partner’s giving action (note that receiving actions had on average a much later onset than giving actions in this experiment).

A related set of findings implicates the Mirror Neuron System (MNS), which is involved in action observation and imitation (Rizzolatti & Craighero, 2004), in the
preparation of complementary actions as well (Newman-Norlund, van Schie, van Zuijlen, & Bekkering, 2007). Moreover, parts of the MNS were found to be more active when participants performed a virtual lifting and balancing task (Bosga & Meulenbroek, 2007) together with another participant (who sat outside the fMRI scanner) than when they performed it separately (Newman-Norlund, Bosga, Meulenbroek, & Bekkering, 2008). In addition, activation in certain areas of the MNS in the right hemisphere was found to be greater when participants performed a complementary (each controlling one end of the bar) than a simultaneous joint task (each controlling both ends). However, others found evidence for only a limited role for the MNS in complementary joint actions (Kokal, Gazzola, & Keysers, 2009).

Behaviourally, the advantage for imitative over non-imitative actions (e.g., Brass, Bekkering, & Prinz, 2001) can be reversed by asking participants to respond with a grasping action that is complementary to an observed grasping action (Van Schie, van Waterschoot, & Bekkering, 2008), suggesting that observed actions can prime complementary actions as well as matched, imitative actions (see also Ocampo & Kritikos, 2010; Poljac, Van Schie, & Bekkering, 2009). Similarly, when participants take turns in a virtual game, the trajectories of their ideomotor movements while they observe the other playing reflect their own rather than the others’ intentions, when their own goals do not match the co-player’s goals (Häberle, Schütz-Bosbach, Laboissière, & Prinz, 2008).

In addition, Motor Evoked Potentials (MEPs) recorded during observation of action requests showed evidence that participants were preparing the appropriate complementary action, and that such preparation activated their own motor system so that the appropriate muscles of the observer’s hand showed increased MEPs (Sartori, Cavallo, Bucchioni, & Castiello, 2011, 2012). Furthermore, when participants watched a video clip in which a model first performed an action that did not require any complementary action from the participant and then began an action request, MEPs recorded from the observer’s hand showed a temporal pattern in which imitative activation was quickly followed by
complementary activation at the very start of the action request (Sartori, Bucchioni, & Castiello, 2013).

The results reviewed so far, however, raise the question of how people who take part in a joint action manage to represent both their own and their partner’s actions when such actions take place (quasi-) simultaneously. This issue has been directly investigated in a study (Novembre, Ticini, Schutz-Bosbach, & Keller, 2012) that compared MEP amplitude recorded from the left arm of pianists who were performing the right-hand part of a piano piece, which they previously practised in its entirety (i.e., with both hands playing their respective parts simultaneously). Crucially, pianists performed the same piece on their own, or in the presence of an experimenter who pretended to perform the left piece. MEPs measured from the left arm were significantly larger when playing with another than when playing on their own (even in a mute session), suggesting that a representation of the left part of the piece was inhibited in solo performance but not in joint performance.

Interestingly, the complexity of the left-hand accompaniment to a melody played with the right-hand affects performance with the right hand in a similar way when the accompaniment is played by the same participant and when it is played by another participant, even though the effect is stronger in the solo than in the joint condition. Similarly to joint finger tapping (see Section 1.2.1), duet partners exhibit reciprocal mutual adaptation, especially when the two partners’ preferred performance rates are more similar (Loehr & Palmer, 2011).

Importantly, when auditory feedback is altered to simulate errors in duet performance, EEG recordings show that the amplitude of the Feedback Related Negativity (FRN) is comparable for the participants’ own errors and for errors produced by their partners. However, the amplitude of the P300 is larger for own errors. In addition, the P300 is more pronounced for errors that affect the joint outcome of the participants and their partners’ performance, compared to errors that affect only the participants’ own performance. Overall, these results suggest that in complementary joint activities partners
monitor each other’s performance, and the joint outcome of their individual actions, but in a way that distinguishes between self-produced and other-produced actions (Loehr, Kourtis, Vesper, Sebanz, & Knoblich, 2013).

### 1.3 Theories of acting together

As anticipated (see Section 1.1), we will present three theoretical approaches that make different claims regarding the processes involved in action coordination. These approaches are: the dynamical systems approach; the shared representations approach; the internal simulation approach.

#### 1.3.1 The dynamical systems approach

The dynamical systems approach to cognition (e.g., Bressler & Kelso, 2001; Kelso, 1995; Warren, 2006) shares with other theoretical approaches (e.g., De Jaegher, Di Paolo, & Gallagher, 2010; Schilbach et al., 2013) an emphasis on processes that do not reside in individual minds/brains. This means that the explanation of observed behavioural patterns of interpersonal coordination is thought to reside in features of the system comprised of two interacting individuals more than in the features of the individuals themselves considered separately.

For example, synchronization, or entrainment, of rhythmic movements (reviewed in Section 1.2.1) emerges because the limbs of two different individuals self-organize to form an informationally coupled system. This happens in a way that is similar to how two moving limbs of a single individual self-organize into a mechanically (as well as informationally) coupled system. Coupled systems (or synergies, or coordinative structures) are composed of multiple interacting parts whose collective dynamics (i.e., the way their behaviour evolves over time as a function of some control parameter, such as frequency of oscillations) can be described with a lower number of dimensions than would be required to describe the
dynamics of their separate components. This feature of coupled systems is referred to as dimensional compression, and it emerges because of constraints that limit the degrees of freedom of the system (Riley, et al., 2011). For example, in the case of synchronization tasks, dimensional compression is reflected by the fact that the dynamics of the system can be characterized in terms of the relative phase angle between the two oscillating limbs (this is called an order parameter; e.g., Schmidt, et al., 1998).

Another feature of coupled systems is reciprocal compensation (Riley, et al., 2011). This captures the fact that the component elements of a coupled system are mutually adaptive. One part of the system will therefore change in response to changes in the other part (i.e., they are dependent on each other) and, furthermore, it will change so as to compensate for deviations that a change in the other part would introduce in the relevant order parameter. For example, in synchronization tasks, if one limb is longer than the other, then the longer limb will have to accelerate (i.e., complete a cycle is a shorter amount of time) and the shorter limb will have to decelerate in order to maintain the relative phase angle fixed at, say, 0 or 180 degrees (see Black, Riley, & McCord, 2007).

The latter example also serves to illustrate the point that characteristics of the individual components of a coupled system can affect its collective behaviour, though proponents of the dynamic systems approach emphasize that they do so mainly by changing the relationship between the components (e.g., Schmidt & Richardson, 2008). In addition, the collective behaviour of the system is thought to be influenced by the nature and amount of information that can be exchanged between its component parts, as illustrated by several studies (see Section 1.2.1) that manipulated, for example, whether synchronizing participants could see each other or not (e.g., M. J. Richardson, et al., 2005). Furthermore, a crucial feature of the coupling of component parts in a synergy is its functional or task specificity (Warren, 2006). This means that each task is characterized by a different set of parameters and by different relations between those parameters which can be discovered only by
empirically studying the behaviour of the system under a range of different conditions (e.g., different values of the control parameters, or different levels of task difficulty).

If these principles indeed apply to the coordination of movements between people, as well as within people (Riley, et al., 2011), interpersonal coordination might not be fully and comprehensively explained by a limited set of mechanisms that are encoded inside the minds/brains of individual participants. Therefore, the dynamical systems approach argues against theories of acting together that are representational, that is against theories that assume that coordination is achieved (at least in part) by individuals representing the actions of others in their own mind/brain. More generally, the dynamic systems approach also argues against theories that study coordination by comparing how the behaviour of individual participants varies when it occurs in isolation as opposed to being performed alongside (Section 1.2.2) or in interaction with (Section 1.2.3) another person.

However, this approach has at least two limitations. First, its proponents have mainly studied one form of coordination between people, that is, the entrainment of rhythmic movements that are analogous across people. For this type of coordination, they could capitalize on an existing model of coordination within people, the HKB model introduced in Section 1.2.1 (Haken, et al., 1985). But this is clearly not sufficient to explain how coordination occurs for non-rhythmic movements. More specifically, this account faces two challenges. One is how to explain the coordination of goal-directed, “one-shot” movements, which might be analogous between people but do not exhibit rhythmic properties. The other is to explain the coordination of complementary movements, which might exhibit rhythmic properties but are by definition not analogous between people.

We are aware of only two studies that have taken up the second challenge. Richardson et al. (2007) asked individual participants to lift planks off a conveyor belt using either one or two hands, and either alone or together with another participant. The latter task requires the coordination of complementary movements (Sebanz, Bekkering, et al., 2006). The authors were interested in the spontaneous transition from solo to joint lifting. They
showed that such transition is governed by similar dynamic principles as the transition from one-handed to two-handed lifting within an individual (with the relevant parameter being the combined arm length of the two participants, instead of the arm length of the individual). However, they did not investigate the processes that sustain coordination of complementary movements, while it happens.

On the contrary, Ramenzoni et al. (2012) asked one participant to hold a circular target while another participant had to keep his or her index finger inside the target, without touching its borders. They investigated directly whether the participants coordinate their bodily movements in a way that reflects the creation of a synergy between them. They concluded that they indeed exhibited dimensional compression. They acknowledged that their study could not investigate whether reciprocal compensation between the two participants took place. The study on joint tapping described in Section 1.2.1 (Konvalinka, et al., 2010) does show the emergence of a process of mutual adaptation between tappers, but note that the task used in that study requires the synchronization of rhythmic (though variable) movements.

There is also a second limitation to the dynamic systems approach. Even if one considers only the entrainment of rhythmic movements, the evidence (see Section 1.2.1) indicates that the behavioural dynamics of an interpersonal synergy is affected by intentions, attitudes, and other social variables. It has been proposed to treat such variables as additional control parameters (Riley, et al., 2011; Schmidt & Richardson, 2008), that is, in the same way as purely physical properties of the coupled system (like frequency of oscillation), but it is not clear how the approach can be extended in this way without such notions as control parameter and synergy losing in specificity. Another possibility is to treat intentions, attitudes, and goals as separate explanatory constructs that can indirectly affect the creation and maintenance of synergies (e.g., Tollefsen & Dale, 2012). However, if this is the case, one could also assume that such constructs are somehow represented inside individual
minds/brains, which seems to be incompatible with the dynamic systems approach to cognition.

1.3.2 The shared representations approach

According to the shared representations approach, when people act together they can form representations of their partner’s actions, as well as forming representations of their own actions (Sebanz, Bekkering, et al., 2006; Sebanz & Knoblich, 2009). Furthermore, they represent their own and their partner’s joint goal as well as representing their own and their partner’s individual (sub-) goals (Loehr, et al., 2013; Pezzulo & Dindo, 2011; Tsai, Sebanz, et al., 2011). Importantly, proponents of this approach do not claim that shared representations are necessary for any form of coordinated action to take place, nor that they are sufficient (Knoblich, et al., 2011; Vesper, Butterfill, Knoblich, & Sebanz, 2010).

This account of how coordination is achieved is based on the notion of parity between action and perception, that is, the idea that representations underlying actions use the same “code” as representations underlying perception. This idea, also known as “common coding”, is central to the Theory of Event Coding, or TEC (Hommel, Müßeler, Aschersleben, & Prinz, 2001; Prinz, 1997). According to TEC, both planned actions and perceived stimuli are represented in terms of event codes, that is, as distal events taking place in the external world. This explains how perception can affect action (e.g., Brass, et al., 2001; Brass, Bekkering, Wohlschläger, & Prinz, 2000; see Heyes, 2011 for a review) and also how action can affect perception (e.g., Müßeler, Steininger, & Wühr, 2001).

Parity of action and perception is supported by the extensive literature on activation of the MNS (Rizzolatti & Craighero, 2004) during action observation (Molenberghs, Cunnington, & Mattingley, 2012), and on unconscious behavioural mimicry (Chartrand & Lakin, 2013). Interestingly, common neural substrates have been identified for the planning and perception of communicative actions (Noordzij et al., 2009), and one study linked moment-by-moment activity in the MNS of a participant trying to guess the meaning of
gestures with activity in the MNS of the participant who produced those gestures (Schippers, Roebroeck, Renken, Nanetti, & Keysers, 2010). In addition, perceiving actions modulates the excitability of muscles, in the observer, that would be used to perform that action (e.g., Fadiga, Craighero, & Olivier, 2005). However, note that the evidence for somatotopic activation of motor areas during action observation is somewhat mixed (e.g., Lorey et al., 2013).

Importantly, the evidence for close links between action and perception, and particularly for the involvement of motor areas in action observation is usually interpreted as supporting the internal simulation approach (Gallese & Goldman, 1998; M. Wilson & Knoblich, 2005), which will be discussed in Section 1.3.3. However, it can also be considered as evidence for the more general claim that action and perception share a common representational code. But in addition, the studies that show motor activation during the observation of actions suggest that this common code uses the “language” of action (possibly including effector-specific representations), and are therefore not fully compatible with the idea that both perceived stimuli and planned actions are coded in terms of their distal properties (Hommel, et al., 2001).

In any case, if perception and action share a common code, then participants acting together in joint tasks could represent the observed actions of their partners using the same representational format they use for their own to-be-executed actions. This could explain joint compatibility effects (see Section 1.2.2), but only if one makes the further assumption that participants can form representations of their partner’s actions also when they cannot directly observe those actions (Sebanz, Bekkering, et al., 2006). This is analogous to the idea that internal simulations of others’ actions can be run when others’ actions are only imagined (see Section 1.3.3 below). In the studies that reported joint compatibility effects, participants must have represented their partner’s actions based on their knowledge of their partner’s task. This knowledge was perhaps reinforced by observing their partner perform the task repeatedly (on other trials), but we know from studies that put participants in separate rooms
that this is not necessary to induce participants to form a representation of their partner’s actions (Atmaca, et al., 2011).

Therefore, it appears that the tendency to represent another person’s actions is very strong in joint tasks, as participants do it even if it is not beneficial to their performance, when they do not interact with their partner, and even when this partner is only imagined. However, the extent to which another’s actions are represented depends also on the identity of the other and the social nature of the relationship between people (see Section 1.2.2). The shared representations account can accommodate this evidence by assuming that the extent to which others’ actions are represented (or monitored) depends on context (Knoblich, et al., 2011). This assumption is supported by neuropsychological evidence (Kourtis, Knoblich, et al., 2013; Kourtis, et al., 2010; Tsai, et al., 2008).

Important, proponents of the shared representations approach also acknowledge that there might be cases in which coordination between two people acting together is achieved without them forming shared representations (Vesper, et al., 2010); such cases could include the synchronization of rhythmic movements (see Section 1.2.1), especially when it occurs unconsciously and without explicit instruction to the participants to synchronize (cf. Knoblich, et al., 2011; Tollefsen & Dale, 2012).

In addition, Vesper et al. (2010) proposed that shared representations could be accompanied (or indeed replaced by) other mechanisms able to support coordination. They termed this collection of mechanisms coordination smoothers. These include: reducing the variability of one’s own movements (Vesper, Schmitz, et al., 2013; Vesper, et al., 2011), or restricting one’s own performance otherwise (e.g., by limiting our actions in space and/or time); using conventional or unconventional signals to communicate with one’s partner (Pezzulo & Dindo, 2011; van der Wel, et al., 2011); intentionally or unintentionally synchronizing with one’s partner (which increases similarity between the coordinating participants); exploiting constraints related to the nature of the task, especially when the task involves the manipulation of objects (pp. 1001-1002).
However, it is not yet clear how these different mechanisms are supposed to interact with one another, and with the formation of shared representations. It appears likely that different types of coordination (see Section 1.2) are supported by different types of mechanisms, but few studies have directly investigated the relationship between them (though see Knoblich, et al., 2011 for a discussion of indirect evidence). In one of these studies, co-actors that rocked chairs in synchrony were faster at jointly moving a ball through a labyrinth (Valdesolo, et al., 2010). Therefore, entrainment with another actor can enhance performance on a subsequent, unrelated joint task requiring the coordination of complementary actions. Interestingly, entrained actors felt more similar to each other and more connected, but these feelings did not predict performance. Instead, enhancement appeared to be mediated by increased perceptual sensitivity to each other’s actions (Valdesolo et al., 2010).

In addition, proponents of the shared representations account have emphasized the importance of another process: prediction (Knoblich, et al., 2011; Sebanz & Knoblich, 2009; Vesper, et al., 2010). Importantly, they acknowledge that such process is needed to account for the coordination of actions in time, which is evident especially when people perform complementary actions in close temporal succession or simultaneously (see Section 1.2.3). Regarding the mechanism used to perform prediction, however, these authors refer to the internal simulation approach, so we will discuss this point in more detail in the next section (1.3.3).

To conclude, the shared representations approach seems better equipped than the dynamic systems approach to account for the ample range of findings on action coordination, at least when the representation (and monitoring) of others’ actions (and goals) is augmented with additional mechanisms including coordination smoothers and prediction mechanisms. It is important to remember that the shared representations approach has been challenged, as discussed at length in Section 1.2.2, and that some joint compatibility effects might be caused, at least in part, by a mechanism different from the representation of others’ actions.
However, the additional evidence coming from recent studies of complementary action coordination (Kourtis, et al., 2010; Loehr, et al., 2013) provides strong support for a version of the shared representations approach that includes a mechanism used to form predictive representations of others’ actions.

### 1.3.3 The internal simulation approach: forward models for own and others’ actions

Many researchers agree that motor involvement in action observation (see Section 1.3.2) can aid action understanding and recognition (Blakemore & Decety, 2001; Buccino, Binkofski, & Riggio, 2004; Rizzolatti, Fogassi, & Gallese, 2001). In addition, it was proposed that the mechanism that underlies motor activation during the observation of others’ actions is an internal simulation mechanism (Gallese & Goldman, 1998). This means that an observer understands and recognizes the action generated by another agent by “putting herself in the agent’s shoes”, adopting his perspective, and using the neural mechanisms that underlie the planning and execution of internally-generated actions to run a simulation of the observed, externally-generated actions. This hypothesis was also extended (Jeannerod, 2001) to suggest that the very same neural mechanism that are involved in the planning and execution of overt actions, and in the observation of actions, are also recruited when we intend to act, or imagine performing an action (i.e., when actions are not overtly executed).

Importantly, part of such proposals is the idea that an internal simulation mechanism could serve the function of anticipating (or predicting) others’ actions (and indeed one’s own actions), as well as understanding them (Frith & Frith, 2006; Gallese & Goldman, 1998; Jeannerod, 2001; Kilner, Friston, & Frith, 2007; Prinz, 2006). Interestingly, other authors (Bar, 2007; A. Clark, 2013; Friston, 2005; Hesslow, 2002, 2012) proposed that anticipation or prediction is the default mode of operation of the mind/brain, not only in the domain of action (or action perception and imagery), but in other domains of cognition as well (from
perception to thinking). Specifically, some proposed that predictions are based on associative, memory-related mechanisms (Bar, 2007; Hesslow, 2002); the predictive coding framework developed by Friston and collaborators (Friston, 2005; Kilner, et al., 2007), instead, argues that predictions are computed by a generative Bayesian model that is a hierarchical model of the (probability) distribution of the causes of sensory input.

A detailed review of all the accounts of prediction in the mind/brain is beyond the scope of this thesis. Instead, we will focus our attention on one mechanism, forward models, that was originally proposed as a basis for prediction in motor and action control (Wolpert, 1997), and learning (Jordan & Rumelhart, 1992; Wolpert, Diedrichsen, & Flanagan, 2011). This mechanism has also been implicated in motor and perceptual imagery (Grush, 2004), implicit thought (M. Ito, 2008), and, crucially for the purposes of this thesis, action observation and joint action (M. Wilson & Knoblich, 2005; Wolpert, et al., 2011; Wolpert, Doya, & Kawato, 2003; Wolpert & Flanagan, 2001).

1.3.3.1 Forward models

A forward model is an internal model of the motor system. Wolpert (1997) distinguished between a forward dynamic model and a forward output model. Given the current state (e.g., the current position of the arm or hand in Wolpert’s example) of the motor system, a forward dynamic model maps from a copy of the motor command sent to the motor system (called an efference copy) to the (predicted) next state of the motor system (e.g., the position of the arm at some point in the future). A forward output model, instead, takes the predicted next state of the motor system as input and computes a prediction of the sensory consequences associated with that state (e.g., the proprioceptive feeling of the arm being in the predicted position, and the visual feedback due to observing the arm in the predicted position).

Some authors have proposed that forward models are computed within the cerebellum (Wolpert, Miall, & Kawato, 1998). This hypothesis has been confirmed by
neurophysiological evidence. For example, Blakemore, Frith, and Wolpert (2001) manipulated the delay between execution of a movement and tactile feedback, and showed that activation in the right cerebellum was positively correlated with the amount of delay. More recently, Knolle et al. (2012) showed that patients with lesions in the cerebellum do not show N100 suppression in auditory cortex in response to self-produced vs. externally-produced sounds. This finding strongly suggests that the cerebellum is implicated in the computation of motor-to-auditory predictions.

The use of forward models in motor control has many potential advantages (Miall & Wolpert, 1996; Webb, 2004; Wolpert, 1997; Wolpert & Flanagan, 2001). First, as suggested by the study by Knolle et al. (2012), since a forward model anticipates the sensory consequences of executing a motor command, its output can be used to cancel sensory reafference (i.e., the sensory consequences of self-generated motion). This function is thought to be crucial in allowing a moving organism to factor out the effects of its own movements when interpreting sensory input, and in distinguishing between internally-caused and externally-caused movements.

Second, a forward model can be used in combination with sensory feedback to estimate the current state of the motor system. This cannot be known directly by the controller (i.e., the central nervous system), which has to derive it from proprioceptive, visual, or other kind of sensory input. However, sensory input is unreliable in two ways: it is subject to delays and to errors (noise). Both of these problems can be overcome by combining actual sensory input with predicted sensory input (based on forward models) to obtain more reliable estimates. The mechanism that adjusts the relative contributions of actual vs. predicted inputs is known as a Kalman filter (Wolpert, 1997).

Third, a forward model can be used to implement an internal feedback loop for the fast correction of errors (which guarantees stability) and, similarly, it can be used to implement a predictive control system, in which errors can be detected and corrected even before they occur. These two cases are conceptually similar in that forward-model
predictions replace actual feedback in the monitoring process. However, in the case of an internal feedback loop, these predictions are computed after a motor command has already been executed and the resulting error signal is used to select an appropriate motor command at the next time stamp to correct for the detected error. Instead, in the case of predictive control, predictions are computed before executing a motor command. These are predictions of the state the motor system would be in after executing a given command; if the desired state is known in advance, then the most appropriate motor command to reach that state can be selected and eventual selection errors can be corrected (Miall & Wolpert, 1996).

Fourth, a forward model can be used to guide sensorimotor learning of suitable inverse models (Jordan & Rumelhart, 1992). An inverse model is also an internal model of the motor system. However, the mapping provided by an inverse model goes from desired outputs (defined as states of the motor system and the associated sensory feedback) to motor commands. As such, inverse models implement the process of motor planning, by which a goal is translated into selection of a motor command. Learning inverse models is difficult because the same goal can be achieved by a potentially infinite number of different motor commands (the one-to-many mapping problem), and vice versa, the same motor command might achieve different goals in different contexts.

Wolpert and Kawato (1998) proposed a model in which multiple forward models are active in parallel and generate predictions that are weighted on the basis of their correspondence to actual sensory feedback. Each forward model is paired with an inverse model, and prediction errors therefore also shift the contribution of these inverse models to the next motor command. Importantly, each forward-inverse pair is tailored to a particular context. Once such contextually-specific models have been learned, identification of a suitable context (through some rapidly available sensory information) could trigger anticipatory selection of the most suitable forward-inverse pair. A hierarchical version of this model was also proposed to account for the generation of sequences of elementary movements (Haruno, Wolpert, & Kawato, 2003).
Fifth, a forward model can be run “off-line”, or decoupled from the motor system (Grush, 2004; Webb, 2004; M. Wilson & Knoblich, 2005). This might serve a variety of functions, from mental practice (Miall & Wolpert, 1996; Wolpert, 1997), to imagery (Grush, 2004), understanding of observed actions (M. Wilson & Wilson, 2005; Wolpert & Flanagan, 2001), observational learning (Wolpert, et al., 2011) and reconstruction of noisy or incomplete perceptual input (Grush, 2004; M. Wilson & Knoblich, 2005).

A detailed review of the evidence that forward models of the motor system do indeed serve all of these functions is beyond the scope of this thesis. Crucially, forward model predictions and the derived prediction error signals could also be used to imitate co-actors, learn their internal models (Wolpert, et al., 2003), potentially complement the observed actions of co-actors with quick and appropriate responses (M. Wilson & Knoblich, 2005, p. 468), and even learn joint internal models for interdependent actions (Braun, Ortega, & Wolpert, 2011; see Wolpert, et al., 2011 for discussion).

We will now discuss evidence for the idea that people can predict others’ actions. While doing so, we will consider whether the available evidence supports the notion that such predictions are (i) computed using mechanisms that are also used in action planning and execution generally, and (ii) computed using forward models of the motor system specifically. Our review will not be exhaustive; we refer to Wilson and Knoblich (2005) and to a recent special issue on the topic (Springer, Hamilton, & Cross, 2012) for further evidence. Finally, we will consider whether there is evidence for the role of action prediction as a mechanism supporting action coordination.

### 1.3.3.2 Predicting others’ actions

First, anticipating the to-be-performed actions of others activates brain areas that are involved in action control. Ramnani and Miall (2004) trained participants to associate cues with the preparation of actions by themselves, a human partner, or a computer. Importantly, these cues were associated with either unspecific preparation (as the actual action to be
performed was indicated by a subsequent go signal) or specific preparation (as the cue indicated which action would have to be performed later on). After training, participants performed their part of the task inside an MRI scanner, and were led to believe that their partner would perform his part of the task in a separate room (while in fact the feedback they received was generated by a computer).

Anticipation of actions performed by the partner (versus the computer) activated brain areas involved in mental state attribution, irrespective of whether the cue was unspecific or specific. Anticipation of the specific actions performed by the partner also activated areas involved in action control. However, these areas did not overlap with action control areas activated during preparation of self-generated actions. Therefore, Ramnani and Miall’s (2004) study is only partially consistent with the internal simulation account of prediction.

Instead, another study provided direct evidence for the internal simulation account by showing that the readiness potential, that indexes the preparation of motor responses, is present from about 500ms prior the observation of a predictable hand action (Kilner, Vargaa, Duval, Blakemore, & Sirigu, 2004). Interestingly, the readiness potential was always contralateral to the observed hand anatomical position, regardless of whether the movement was presented from an egocentric or an allocentric perspective (in which, therefore, the spatial position of the observed hand does not match its anatomical position), which provides further support for a simulation account.

A TMS study recently provided some indication that part of the premotor cortex might be causally involved in action prediction. Participants watched video clips of everyday actions that were occasionally occluded. When the occlusion was removed, the actions continued in a way that was coherent or incoherent (time-wise) with respect to the duration of the occlusion. Repetitive TMS applied to premotor cortex at the start of the occlusion (but not 300 ms later) disrupted their ability to make such judgments compared to when TMS was
applied to a control site, but it must be noted that this effect was small and not supported by
a fully significant interaction between TMS site and TMS onset (Stadler et al., 2012).

Furthermore, it is not clear to what extent simulation during action prediction
resembles action execution. For example, one fMRI study (Lamm, Fischer, & Decety, 2007)
showed no somatotopic activation of motor areas in a task where participants had to judge
whether an object was within the reach of the hand or foot of an agent. The object and agent
were shown as static pictures. Behavioural results indicated a tendency to overestimate the
agent’s reaching range in a way that is similar to how people overestimate their own
reaching range. This suggests that participants’ judgments were based on motor predictions
(or imagery), but the lack of somatotopic activation speaks against a detailed simulation
within their own motor system.

In contrast, however, another study showed that when participants are asked to
anticipate the end state of an unfolding action, their performance is impaired when their own
body posture is incongruent with the end state. Importantly, incongruence between predicted
and own body posture led to higher activation of a brain region (left intraparietal sulcus) that
is also involved in producing state estimates of the position of one’s own limbs
(Zimmermann, Toni, & de Lange, 2013).

This finding adds to behavioural results indicating that the action capabilities of the
observer’s body affect her ability to predict an unfolding action. For example, it is known
that gaze consistently anticipates hand movements in manipulation tasks, and that the same
predictive relationship holds during action observation as during action execution (Flanagan
& Johansson, 2003). Importantly, this predictive relationship is disrupted when the observer
is not allowed to move her hands (Ambrosini, Sinigaglia, & Costantini, 2012), or when her
hand is concurrently holding an object using a grip shape that is not compatible with the
observed grip shape (Costantini, Ambrosini, & Sinigaglia, 2012). Moreover, predictive eye
movements during observation of hand actions are disrupted when TMS is applied to the
motor hand (but not leg) area (Elsner, D’Ausilio, Gredebäck, Falck-Ytter, & Fadiga, 2013).
Further evidence that motor plans learned for the control of one’s own actions are used to support prediction during action observation comes from studies that compared observation of self-generated vs. other-generated actions. If observers internally simulate the actions they observe using their own motor system, then they should be better at predicting how their own actions unfold over time compared to actions performed by others. This is because simulation of one’s own actions should be more accurate than simulation of actions performed by others.

Indeed, people are better at predicting a movement trajectory (e.g., in dart-throwing, handwriting, or table tennis) when watching a video of themselves versus others (Bischoff et al., 2012; Knoblich & Flach, 2001; Knoblich, Seigerschmidt, Flach, & Prinz, 2002). However, this could be due to better memory for self-generated than other-generated actions, or to greater perceptual exposure to self-generated than other-generated actions (and their consequences). Contrary to this alternative explanation, though, basketball players are better at predicting the success of throws than experts with comparable visual experience (e.g., sports journalists) but lower relevant motor experience, and their prediction abilities might be specifically linked to increased motor activation during observation (Aglioti, Cesari, Romani, & Urgesi, 2008). Moreover, the onset of infants’ ability to predict the goal of an action coincides with the onset of their ability to perform the same action during development (Kanakogi & Itakura, 2011).

The studies reviewed so far focused on the role of simulation in predictions of movement kinematics or movement end states. Therefore, they predominantly studied what predictions. Participants were expected to anticipate what the observed agent would do next. However, when predictions are just as important as what predictions in everyday life, and especially in joint action: it is important not only to anticipate what another is about to do, but also when, or for how long, he is about to do it (Sebanz & Knoblich, 2009).

Graf et al. (2007) showed participants point-light displays of unfolding human actions. The dynamic displays were occluded, and then the participants saw a static frame
after a variable time lag; the frame always displayed the same action, but at a later point in time, which was also varied. The authors manipulated whether the lag between occlusion and presentation of the frame corresponded to the actual time elapsed during execution of the action till the point at which the static frame was taken. Participants were instructed to judge whether the frame was a continuation of the dynamic display. The authors found that performance was best when the observation lag corresponded exactly to the actual execution time, and argued that their finding supported the hypothesis that prediction is based on action simulation.

Therefore, Graf et al.’s (2007) study provides some evidence that when predictions could be based on a similar simulation mechanism as what predictions. Interestingly, when predictions are affected by the degree of overlap between observed and performed movements, just as what predictions are (Springer et al., 2011), and priming a different effector (e.g., by having participants move their legs repeatedly before judging the congruency of partially occluded arm movements) suppresses participants’ ability to form when predictions (Springer, Brandstadter, & Prinz, 2013).

Additional, though indirect, evidence comes from Keller et al.’s (2007) finding that pianists find it easier to synchronize with a recording of themselves than with a recording of somebody else. However, note that when Sparenberg et al. (2012) attempted to replicate Graf et al.’s finding with a higher temporal resolution (i.e., varying the lag with smaller increments), they found that in fact observers showed a systematic bias towards continuations that were shifted around 25 ms into the future, suggesting some limitations to the idea that action simulation occurs “in real time”.

Overall, this complex set of findings is broadly compatible with the internal simulation account, as it shows that mechanisms involved in action planning and execution are involved in action prediction. However, as we noted, there are inconsistencies between studies with regard to the level of overlap between action prediction (and, indeed, observation) and action planning and execution. Studies that have manipulated the motor
capabilities of participants, or compared participants with different levels of motor expertise support the claim that simulation during action prediction involves activation of detailed, specific motor plans.

However, other studies did not observe the brain activation patterns that would have been expected if such detailed motor plans were indeed used in action prediction. These studies are perhaps more consistent with a view in which people anticipate others’ actions using higher-level representations, such as goals, distal properties of actions (e.g., Lamm, et al., 2007), or perhaps some form of association between different parts of an action sequence (see Springer, et al., 2012 for a discussion).

Furthermore, some findings are explicitly interpreted with reference to forward-model predictions (Flanagan & Johansson, 2003; Zimmermann, et al., 2013); however, we are not aware of any study that supports the idea that prediction of others’ actions is based specifically on forward models of one’s own motor system, and not on a different mechanism. To the contrary, studies that suggest that the time-course of simulation matches the time course of execution (e.g., Graf, et al., 2007) appear to be at odds with the idea that others’ actions are anticipated using forward models. Forward model computations should be faster than actual execution. Such studies might be interpreted as suggesting that motor activation during action prediction involves execution mechanisms, which are however inhibited so as to prevent overt execution of the observed action.

In addition, note that most of the studies that have looked at action prediction used an explicit prediction task, that is, they asked participants to try and anticipate an observed action. Therefore, it might be argued that these studies are not informative as to whether people routinely predict during action observation in more naturalistic situations. However, there are some exceptions. Both Ramanani and Miall (2004) and Kilner et al. (2004) relied on implicit association between cues and actions in their paradigms; further, studies on anticipatory gaze movements during observation of manipulation tasks used passive viewing
paradigms. Thus, it can be concluded that people can simulate the future actions of others even when they are not explicitly asked to do so.

1.3.3.3 Prediction of others’ actions and coordination in joint tasks

Finally, let us consider the evidence that people simulate and predict their partner’s actions in joint tasks and that this mechanism underlies their ability to successfully coordinate. First, imaging (e.g., Newman-Norlund, et al., 2007) and TMS studies (e.g., Sartori, et al., 2011) reviewed in Section 1.2.3 provide compelling evidence that activation of motor plans during action observation is not limited to motor plans that imitate the observed action, but extends to complementary motor plans that are likely to be useful in joint actions. Second, electrophysiological evidence (e.g., Kourtis, Sebanz, et al., 2013) indicates that motor preparation in anticipation of another’s actions occurs more for an interaction partner than for a by-stander, thus suggesting that use of the motor system to predictively simulate another’s actions might be particularly enhanced in joint tasks.

Third, Loehr et al.’s (2013) finding that the amplitude of the Feedback Related Negativity is indistinguishable for altered feedback affecting one’s own actions or a partner’s actions, suggests that the comparison between expected and actual feedback is based on the same mechanism in both cases. More specifically, the authors argued that participants learned forward models linking finger movements to expected auditory outcomes through extensive practice, and that the FRN reflected a prediction error signal (see p. 1057, for a discussion and rejection of alternative explanations). Crucially, since they learned forward models for both their own and their partner’s piece, prediction error signals were generated in the same way irrespective of whether the mismatch affected self-generated or other-generated actions.

Moreover, some behavioural studies support the hypothesis that prediction is involved in successful coordination. Knoblich and Jordan (2003) argued that their participants learned to anticipate each other’s button presses to avoid interference between
their actions. Pecenka and Keller (2011) showed that musicians who can more reliably anticipate the next stimulus when tapping in synchrony with a computer-generated stimulus are also better at synchronizing their taps with another person, provided that person is also good at prediction (section 1.2.1).

In addition, Ray and Welsh (2011) showed that participants spontaneously anticipated the action needs of their partner in a joint task. When they were asked to hand a jug to a confederate, they overwhelmingly passed the jug with the handle oriented towards the confederate. However, this preference was not modulated by the specific goal of the confederate subsequent action (i.e. whether the confederate ultimately used the jug in a pouring or a placing action), suggesting that the action needs associated with the specific goal (i.e., that pouring requires using the handle more than placing) were either not anticipated or, if they were anticipated, they did not affect the participants’ behaviour. One possibility is that participants opted to be generally “helpful” towards the confederate irrespective of the confederate’s specific needs.

The behavioural results mentioned above do not directly implicate the motor system as the basis of predictions. To the contrary, Vesper et al. (2013) provided evidence for such a link. They asked participants to take a jump with the goal of landing at the same time as another participant; the difference in jumping distance between partners was varied. Despite the lack of continuous feedback from a partner’s jump, participants were able to synchronize their landing times. In particular, the participants who had to travel the shorter distance waited for longer before initiating their jumps. According to the authors, this suggests that the planning of their movements was affected by a simulation of the movements that were being performed by their partner. Interestingly, very similar results were obtained when participants were asked to imagine jumping with another person, but did not perform any actual jumping action (Vesper, Knoblich, & Sebanz, in press).

Moreover, the kinematics, as well as the timing, of the participants’ jumps appeared to be affected by a simulation of their partner’s movements (Vesper et al., 2013), suggesting
that such simulations can interfere not only with planning but also with execution of the jump. However, more research is needed to assess the generalizability of this finding. Interestingly, in another study, participants who had to undo their partner’s actions by returning a knob to the starting position after the partner had turned it, anticipated their partner’s movements, but the kinematics of their actions was not affected by the kinematics of the anticipated (and observed) actions of their partners (Herbort, Koning, van Uem, & Meulenbroek, 2012).

In conclusion, people can use their own motor systems to simulate and predict each other’s actions, and this might underlie coordination in at least some joint activities. This is consistent with the internal simulation approach to acting together. However, a few issues remain open and should be clarified by future research. (i) Is simulation necessary for prediction? (ii) How detailed is simulation-based prediction? (iii) Is simulation-based prediction based on forward models? (iv) Is simulation the only, or the most important mechanism underlying action coordination between people?

1.4 Theories of talking together

In Section 1.3, we discussed three different theoretical approaches to the coordination of actions between people. In this section, we consider theoretical approaches to the coordination of utterances between speakers. As mentioned at the beginning of this chapter (Section 1.1), we focus our attention on the processes or mechanisms that might underlie the coordination of utterances (as in synchronization, alignment, turn-taking, and compound contributions).

The existence of such coordination phenomena in natural conversations has long been noted, and it prompted Clark (1996) to liken conversation to other joint activities, like playing a piano duet (p. 3). Clark regards the process by which individual actors manage to

---

coordinate to be a form of problem solving, and his focus is on “strategies” that they use to attain coordination. What do these strategies consist of? According to Clark, they consist in creating “coordination devices” (p. 64), or exploiting the existence of pre-defined, conventional, coordination devices.

For example, the commit-and-repeat strategy (Clark, 1996; pp. 269-271) is one of several strategies speakers can use to retain the floor when they have not yet completed formulation of the next constituent. It consists of speakers producing the beginning of the new constituent early on, and then repeating it when the rest of the constituent is ready to ensure a fluent delivery. This strategy works by the speaker signaling her commitment to carry on speaking to the listener. The listener acknowledges reception of this signal (in this case by not taking the floor despite a delay in the speaker’s utterance). Early production of part of a constituent functions, in this case, as a coordination device.

More generally, a coordination device is a piece of information that is in common ground between speakers. For a piece of information to be in common ground, according to Clark (1996, p. 66), it must be the case that each speaker possesses that piece of information, and, in addition, is aware that the other speaker possesses it (and also, that each speaker is aware of the other being aware, and so on, in an infinitely recursive way). In the commit-and-repeat strategy example, this requirement is fulfilled by the fact that speakers and listeners jointly attend to the speaker’s utterances.

Some coordination devices are already in place between two speakers for the mere reason that they belong to the same linguistic community. Others, however, can be created by one speaker signaling to the other and by the other acknowledging reception of the signal. This process of signaling is helped by the fact that conversations have a procedural, conventional nature (Levinson, 1992); in other words, they are organized according to a set of principles that facilitate coordination.

For example, the continuity principle, which states that language should be produced fluently whenever possible (H. H. Clark & Wasow, 1998), facilitates coordination in the case
of the commit-and-repeat strategy mentioned above. Similarly, Clark (2002) assumes that
speakers produce certain types of disfluencies to inform listeners that they are violating the
continuity principle (H. H. Clark & Fox Tree, 2002; Fox Tree & Clark, 1997). But he is
silent on the mechanisms that normally allow synchronization of the acts of production and
comprehension, merely pointing out that listener and speaker need to jointly attend to the
speaker’s productions.

To further illustrate, consider turn-taking. As described in Section 1.1, speakers and
listeners take turns by repeatedly switching roles in the conversation. This alternation is
managed “on the fly” by the participants themselves, at least in informal conversations
(Sacks et al., 1974; Clark, 1996). According to Clark’s (1996) approach, turn-taking is the
result of a set of strategies. Sometimes, the current speaker explicitly selects the next speaker
(e.g., in questions). In other cases, listeners monitor the speaker’s utterance and “project”
(anticipate) when this utterance might be completed (see Section 1.1), in order to know when
they might begin their turn. They then signal their intention to begin a turn.

What the account does not explain is what mechanisms listeners might use to
anticipate turn endings. One suggestion is that listeners might rely on the presence of cues,
that is, properties of utterances that reliably signal that the end of a turn is approaching, as
assumed by the signal theory of turn taking (Duncan, 1972). Or they might develop ad-hoc
procedures over the course of the interaction, through a process of explicit or implicit
negotiation (Mills, in press).

In summary, Clark (1996) describes the dynamics of coordination at what we might
call the “intentional” level. Interlocutors coordinate by making inferences about the
intentions underlying their partners’ behavior. Ultimately, coordination is successful if they
develop mutual beliefs about their intentions. In this, they are helped by the existence of
conventions (e.g., turn-allocation norms), and cues (e.g., disfluencies) that map intentions
onto behaviour. But this account says very little about the cognitive (and neural) processes
and mechanisms underlying coordination (though see Brennan, et al., 2010).
Moreover, while we agree with this account that dialogue is a form of joint action, and that it requires coordination between interlocutors, we think that Clark’s (1996) account of how coordination is achieved is limited because it focuses on the coordination of intentions. Interestingly, the declared aim of his account is to treat utterances in dialogue as actions, as opposed to products (Brennan, et al., 2010). However, the emphasis placed on intentions has actually led to substantially ignore the linguistic actions themselves. In other words, the account can explain how language supports the coordination of non-linguistic intentions and plans (e.g., buying something in a shop), but does not explain how the utterances themselves (as unfolding sequences of linguistic constituents) are coordinated.

Other accounts of coordination in dialogue, on the contrary, focus precisely on the coordination of utterances rather than on the coordination of intentions. To illustrate with an example, take the oscillator model of turn-taking introduced in Section 1.1. Wilson and Wilson (2005) proposed a mechanism to explain the observed distribution of inter-turn intervals, which displays few overlaps and relatively short gaps. Their account appeals to the properties of oscillators, systems characterized by a periodic cycle. As seen in Section 1.2.1, mechanical oscillators (e.g., oscillating limbs) tend to spontaneously attune their cycles, so that they become entrained: their cycles come into phase (or anti-phase). In addition, neural populations firing at certain frequencies might act as oscillators, and sensory information regarding the phase of another oscillator (e.g., in another human body) could serve to fine-tune them (see Arnal & Giraud, 2012 for a recent review).

In the oscillator model of turn-taking, the production system of a speaker oscillates with a syllabic phase: the readiness to initiate a new syllable is at a minimum in the middle of a syllable and peaks half a cycle after syllable offset. Wilson and Wilson (2005) argued that interlocutors converge on the same syllable rate, but their production systems are in anti-phase, so that the speaker’s readiness to speak is at minimum when the listener’s is at a maximum, and vice versa. Therefore, in this account, turn-taking behaviour could be
explained by an automatic coordinative mechanism, which does not require the coordination of intentions.

However, the oscillator model cannot fully explain turn-taking. First, regularities in speech appear to take place over very short time-scales, with the cyclic pattern of syllables that Wilson and Wilson (2005) propose as the basis for entrainment occurring at 100-150 ms. If predictions were made on the basis of syllable-level information alone, there would simply be not enough time to prepare the next contribution and leave a 0 ms gap. Anticipation of the end of a turn, instead, must draw on information that spans units larger than the syllable. Indeed, the work by De Ruiter and colleagues has showed that listener’s ability to predict a turn ending depends on their ability to predict with how many and which words the turn will continue (De Ruiter, et al., 2006; Magyari & De Ruiter, 2012). In addition, it has been noted that syllables do not exhibit the regularity of oscillators (Cummins, 2012). Moreover, a corpus study that directly attempted to validate the model found that convergence in speech rate was not a predictor of the variability observed in inter-turn intervals (Finlayson, Lickley, & Corley, 2012). Thus, there must be other mechanisms underlying the coordination of utterances in turn-taking.

More generally, this model is limited in its explanatory scope because it cannot account (at least not with the same mechanism) for the timing of turn-taking and for the pragmatic link between two adjacent turns in a conversation. Consider, for example, how answers complement questions. For an addressee to produce an appropriate answer, it is not enough to talk in anti-phase with the speaker. She must be able to plan in advance not only when to start speaking, but also what to say. This is even more evident in compound contributions (see Section 1.1).

Clearly, the complementary nature of utterances in dialogue can be explained within an intentional account as Clark’s (1996). In fact, it has been argued that the existence of compound contributions can only be explained if one assumes that interlocutors have mutual beliefs about their intentions, and monitor each other’s utterances continuously, comparing
them to expectations derived from these beliefs (Poesio & Rieser, 2010). This might be correct, but we also note that making this assumption is not sufficient. One must also specify how this monitoring is performed; i.e., what sort of mechanism underlies the computation of expectations and their comparisons to one’s partner’s utterances (see Sections 1.4.2 and 1.4.3).

Interestingly, a recent proposal seeks to account for the complementary nature of utterances in dialogue (and for the coordination of utterances in general) from a non-intentional perspective. This proposal uses the concept of synergy or coordinative structure that we have introduced in Section 1.3.1, and is in fact based on the dynamical systems approach to acting together (Fusaroli, Rączaszek-Leonardi, & Tylén, in press; Shockley, et al., 2009). As this account is non-representational, it is not concerned with the coordination of intentions; these are representations contained in the mind/brain of individual interlocutors. Instead, this account takes the system formed by two conversing individuals as the fundamental unit of analysis.

Further, coordination is not treated as a form of problem solving, or as the derivation of inferences about another’s mental states. Rather, interlocutors mutually constrain one another’s behaviour, so that the range of possible behaviours is greatly reduced (cf. dimensional compression) and coordination is facilitated by the (self-organized) emergence of structure. As an example of structure, the authors mention interactional routines and procedures (Levinson, 1992). Such emergent structures are entirely context-dependent, so that different structures can emerge in different conversations. This guarantees flexibility. However, within each context, the existence of a specific function will determine what structure emerges and guarantee its stability (Fusaroli, et al., in press, pp. 4-5). Importantly, because this structure is defined at the level of the system, and not at the level of individual participants, it allows for reciprocal compensations between the interlocutors, and is consistent with the notion of complementarity between utterances.
However, we believe that this account of coordination in dialogue is affected by the same issues we pointed out in Section 1.3.1 in relation to the dynamical systems approach to acting together. That is, it is not clear how variables like intentions, goals, attitudes, and beliefs, that undoubtedly affect the coordination of utterances between interlocutors (Brennan, et al., 2010), and also speech-related gaze patterns (D. C. Richardson, Dale, & Tomlinson, 2009; Shockley, et al., 2009), can be equated with control parameters that govern the behaviour of a dynamical system. In addition, the application of these concepts to linguistic interactions is further complicated by the fact that language is hierarchically organized (Fusaroli, et al., in press), and the account does not yet specify at what level(s) these parameters should operate (phonetic, lexical, et cetera).

In this thesis, we will take a third approach, one that sits at a level of analysis that is intermediate between high-level intentions and low-level observable behaviour. In essence, it is at this level that one can define a cognitive architecture for coordination. This should comprise a set of mechanisms (representations and processes acting on those representation) that underlie the coordination of utterances between interlocutors. More specifically, we propose that the most promising way of identifying these mechanisms stems from a mechanistic account of language processing. This is of course what psycholinguistic theories have traditionally tried to develop. However, most of these theories are concerned with monologue, in which speakers and listeners act in isolation.

Pickering and Garrod (2004) pointed out the need for a theory of dialogue that can explain the seemingly effortless, automatic nature of conversation. In the next section (1.4.1), we present this theory and the evidence that supports it. Then, we outline a few criticisms of this theory that highlight the fact that it is not sufficient to explain the full range of coordination phenomena we encounter in conversations. In the following section (1.4.2), we present a more recent theory (Pickering & Garrod, 2013) of the mechanisms and processes that support coordination in dialogue. Finally, in Section 1.4.3, we present our
proposal of a cognitive architecture for the coordination of utterances, which is based on the theory presented in Section 1.4.2.

1.4.1 Interactive alignment

Pickering and Garrod (2004) proposed that interlocutors come to a mutual understanding via a process of alignment, whereby their representational states tend to converge during the course of a conversation. Alignment occurs at many different levels, including the phonetic, lexical, and syntactic (see Section 1.1). In addition, alignment at one level leads to alignment at other levels, so that ultimately interlocutors align their understanding of the situation (i.e., a situation model). Importantly, Pickering and Garrod argued that the simple mechanism of priming (i.e., facilitation in processing of an item due to having just processed the same or a related item) underlies such alignment, which in turn facilitates successful interaction. In their model, therefore, coordination between interlocutors results from a mechanism of priming that is known to operate within the individual speaker’s production system and the individual listener’s comprehension system.

To account for alignment between speaker and listener, Pickering and Garrod (2004) assumed representational parity between production and comprehension. If the production and comprehension systems make use of the same representations, those representations that have just been built in comprehension can be used again in production and vice versa. Because interlocutors alternate between production and comprehension, their production and comprehension systems become increasingly attuned. Below, we review the evidence for the assumption of parity of representations, starting with sound-related parity (i.e., parity at the phonetic and phonological levels; Section 1.4.1.1), and then focusing on lexico-semantic and syntactic parity (Section 1.4.1.2).
1.4.1.1 Sound-related parity

There is a large body of evidence demonstrating motor activation during the perception of speech, which is analogous to the evidence for motor activation during action observation (reviewed in Section 1.3.2). Similarly to the action perception case, the evidence that will be reviewed below has been usually interpreted as supporting the internal simulation approach (Section 1.3.3), that is the idea that people internally simulate perceived events using their own motor system. However, it can also be interpreted as supporting the idea that production and perception use a common representational format, which is why we opted to present this evidence here. But in addition, this evidence suggests that representations that are shared between perception and production are grounded in production mechanisms. Importantly, the latter assumption is not part of the interactive alignment model (Pickering & Garrod, 2004), but it has been incorporated in a more recent model by the same authors (Pickering & Garrod, 2007; 2013; see Section 1.4.2).

First, Fowler et al. (2003) showed that people are faster at producing a syllable in response to hearing the same syllable than in response to a tone; in fact, shadowing a syllable yielded response latencies that were nearly as fast as those found when the to-be-produced syllable was fixed and known in advance (see also Galantucci, Fowler, & Goldstein, 2009). Moreover, Kerzel and Bekkering (2000) demonstrated an action-perception compatibility effect for speech (due to a task-irrelevant stimulus). They found that participants pronounced a printed syllable while watching a video of a mouth producing the same syllable more quickly than when the mouth produced a different syllable (see also Jarick & Jones, 2008). While the first study involves intentional imitation, the second one provides more compelling evidence for automaticity.

However, they both deal with cases of overt imitation, where there is an overt motor response. Additionally, neuropsychological studies found activation of motor areas during audiovisual speech perception (Skipper, Nusbaum, & Small, 2005; Skipper, van Wassenhove, Nusbaum, & Small, 2007), and also during passive listening to speech (e.g., S.
M. Wilson, Saygin, Sereno, & Iacoboni, 2004). Importantly, motor activation during passive listening is articulator-specific (Pulvermüller & Fadiga, 2010; Pulvermüller et al., 2006).

Moreover, several studies found that listening to speech modulates the excitability of speech-related muscles. For example, Fadiga et al. (2002) showed that MEPs recorded from tongue muscles are larger when listening to words whose articulation would require moving the tongue (e.g., the Italian word *birra*, “beer”, that contains the alveolar trill [r], produced by vibrations of the tip of the tongue against the alveolar ridge) than for words whose articulation requires less tongue involvement (e.g., the Italian *baffo*, “moustache”, that contains the labiodental fricative [f]). Interestingly, the effect occurred also for pseudowords (e.g., *berro* vs. *biffo*), but it was less pronounced than with words (Fadiga, et al., 2002).

The influence of lexical frequency on the magnitude of MEPs was further investigated by Roy et al. (Roy, Craighero, Fabbri-Destro, & Fadiga, 2008), who found that the excitability of tongue muscles was enhanced for rare compared to frequent words; this study also replicated the previous finding that excitability is larger for pseudowords containing tongue-related consonants (e.g., *cillo*) than for pseudowords that require less tongue movement (e.g., *cimmo*), and established that this effect is maximal around 100 ms after the onset of the relevant consonant (i.e., “ll” or “mm”).

Another TMS study showed that listening to speech or watching “silent” mouthing videos both enhance the excitability of the *oris orbicularis* muscle, which controls the lips, compared to listening to non-speech sounds or watching non-speech movements (Watkins, Strafella, & Paus, 2003). Excitability of this muscle was also larger when watching a silent video articulating /ba/ than a silent video articulating /ta/, though a listening-only condition failed to induce enhanced MEPs compared to baseline (Sundara, Namasivayam, & Chen, 2001).

In addition, two recent studies provide converging evidence that listening to speech affects articulation in subtle ways. Yuen et al. (2010) demonstrated that passive listening can modulate the trajectory of concurrently performed articulatory movements so that, for
example, the degree of contact between the tongue and the alveolar ridge during production of syllables starting with /k/ (a phoneme that does not require alveolar contact in its production) was larger when participants were listening to syllables beginning with /t/ (a phoneme that requires alveolar contact). Ito and Ostry (2012) applied stretches of equal magnitude to the skin of participants’ cheeks while the participants listened to /had/ or /head/. Articulation of the vowel in /had/ requires larger jaw displacement and is normally accompanied by an increased feeling of stretch at the cheeks. Accordingly, when participants listened to /had/ they judged the stretch to be greater than when they listened to /head/, even though the force applied was exactly the same in both cases. Taken together, these findings suggest that articulatory and somatosensory representations are involved in speech perception.

Note that the evidence for motor involvement in speech perception that we have reviewed above has often been interpreted as supporting the Motor Theory of Speech Perception (Galantucci, Fowler, & Turvey, 2006). However, this interpretation has been criticized (Lotto, Hickok, & Holt, 2009), and some findings from imitation tasks go against a direct link between speech perception and production, as we will discuss below.

1.4.1.2 Boundary conditions for motor involvement in speech perception

First, two studies failed to find evidence that listening to a phonetic variant (e.g., alveolar /r/) facilitates subsequent production of the same variant compared to listening to an alternative variant (e.g., uvular /r/) that is phonologically equivalent (Mitterer & Ernestus, 2008; Mitterer & Müsseler, 2013). This finding suggests that the link between speech perception and production might be mediated by abstract phonological representations.

Furthermore, there is disagreement between scholars with respect to the functional role of motor involvement. Some argue that motor areas have an active (and possibly causal) role in speech perception and comprehension (Iacoboni, 2008; Pulvermüller & Fadiga,
(2010), while others (S. K. Scott, McGettigan, & Eisner, 2009) pointed out that this role might be served primarily by temporal auditory areas, while the activation of motor areas might instead support coordination between the interlocutors’ speech production systems (for example, during turn-taking).

The former view is supported by studies that modulated activity in motor areas using TMS and showed that this manipulation affected performance on a phonemic task. For example, Meister et al. (2007) reported that repetitive TMS applied to the premotor cortex impaired participants’ ability to identify an auditorily presented consonant against background noise (but not their ability to identify colours). Moreover, Möttönen and Watkins (2009) found that repetitive TMS applied to the lip area selectively impaired participants’ ability to discriminate sounds along a continuum between /ba/ and /da/, where one end of the continuum involves lip movement (but not along a continuum between /ka/ and /ga/, where neither end point involves lip movement). Conversely, D’Ausilio et al. (2009) showed that double-pulse TMS applied to the motor lip area just before speech onset facilitated the identification of labial consonants (/b/ and /p/) in noise, compared to equivalent stimulation applied to the motor tongue area (and vice versa for dental consonants, like /d/ and /t/, that are articulated with the tongue).

Moreover, the perceptual boundary between /had/ and /head/ can be shifted by stretching the listeners’ cheeks downwards or upwards; crucially, the direction of the shift is specific to the direction of the stretch (e.g., participants are more likely to report hearing /had/ for ambiguous stimuli when their cheek is stretched downwards), in a way that suggests that somatosensory representations are causally involved in speech perception (T. Ito, Tiede, & Ostry, 2009). In addition, Adank et al. (2010) showed that overt imitation of an unfamiliar accent improves perception of utterances produced in that accent (under noisy conditions) more than pure exposure and repetition without the explicit instruction to imitate.

Furthermore, silent articulation (mouthing) of a syllable speeds up identification of the same (concurrently presented) syllable and slows down identification of a different
syllable, both when the syllable is auditorily presented (Sams, Möttönen, & Sihvonen, 2005), and when it is presented auditorily and visually (Sato, Troille, Ménard, Cathiard, & Gracco, 2013). Interestingly, inner speech (i.e., imagining articulation of a syllable without moving the articulators) can also bias perception of ambiguous syllables towards the imagined syllable (M. Scott, Yeung, Gick, & Werker, 2013).

However, related findings suggested a more limited role for motor areas in speech perception. First, one study investigated the role of premotor cortex in speech perception in the absence of noise (Sato, Tremblay, & Gracco, 2009) and reported no disruptive effect of TMS when participants were asked to identify the heard phoneme (/p/ or /b/) or judge whether two syllables were similar or dissimilar. To the contrary, TMS had a negative effect on their ability to judge whether the initial phonemes of two syllables were similar or dissimilar, suggesting that the premotor cortex might be involved in speech segmentation, but not in other aspects of speech perception.

Second, Wilson and Iacoboni (2006) compared neural activation while listening to native and non-native phonemes, while also varying the degree to which non-native phonemes were easy or difficult to produce (as assessed in a norming study with a different group of participants). They found motor activation during speech perception, and also that such activation was different for native compared to non-native phonemes. However, the easiness of production of non-native phonemes did not correlate with motor activation; instead, it correlated with activation in temporal auditory areas.

Third, motor areas might be recruited differently depending on the nature of the task (S. K. Scott, et al., 2009). In support of this idea, Möttönen et al. (2013) found that repetitive TMS applied to the lip area suppressed the Mismatch Negativity (MMN), which is normally observed in response to changes in sound sequences; crucially, this occurred only when the changes occurred in sequences of syllables, but not when they occurred in sequences of non-speech tones. However, MMN responses to lip and non-lip related speech sounds were equally suppressed. Interestingly, in this study the listeners had no specific task and were
exposed to the sound sequences while watching an unrelated silent video. The authors therefore speculated that motor activation could be non-specific when the participants are not given an explicit task in relation to the speech input (though cf. Pulvermüller, et al., 2006).

A good candidate for explaining differences in motor involvement revealed by different studies could be the degree of perceptual difficulty associated with the task. Interestingly, a recent meta-analysis of studies that investigate speech perception under difficult conditions concluded that activation of areas that overlap between perception and production is more likely in difficult perceptual conditions (Adank, 2012). More importantly, D’Ausilio et al. (2012) showed that the findings reported in their previous study (D’Ausilio, et al., 2009) hold only when the speech stimuli are presented in noise; for noise-free stimuli, they could not show any effect of TMS stimulation of motor areas on phoneme identification. Finally, Roy et al.’s (2008) finding that the excitability of tongue muscles was enhanced for rare compared to frequent words could suggest that motor involvement is enhanced also when comprehension of speech is difficult in general, and not just when it is perceptually difficult.

1.4.1.3 Lexico-semantic and syntactic parity

In picture-word interference studies, participants automatically comprehend distractor words and this affects their production of target picture names (e.g., Schriefers, Meyer, & Levelt, 1990). Despite the fact that such studies are used to answer questions about language production (e.g., Levelt, Roelofs, & Meyer, 1999), they also provide evidence that lexico-semantic (and phonological) representations are shared between comprehension and production (Pickering & Garrod, 2013). In addition, the rapidity with which some listeners can shadow a speaker’s speech (with latencies which can be as short as 250 ms) also suggests that the link between comprehension and production representations might be direct (e.g., Marslen-Wilson, 1973).
Moreover, structural priming effects occur from comprehension to production, as well as from production to production (Pickering & Ferreira, 2008). Interestingly, there is also evidence that the magnitude of lexical alignment effects depends partially on beliefs interlocutors hold about their partners, with greater alignment occurring when participants believe they are interacting with a computer than when they believe they are interacting with a human (Branigan, Pickering, Pearson, McLean, & Brown, 2011).

Kempen et al. (2012) provided evidence that grammatical encoding (in production) and decoding (in comprehension) might share a common mental “workspace”. They asked Dutch participants to paraphrase sentences from direct (e.g., *De lottowinnaar zei / ‘Ik heb besloten een rode auto te kopen voor mezelf’*, The lottery winner said / “I have decided to buy a red car for myself”) into indirect speech (e.g., *De lottowinnaar zei dat hij had besloten een rode auto te kopen voor zichzelf’*, The lottery winner said that he had decided to buy a red car for himself). Some of the sentences to paraphrase contained an ungrammatical reflexive pronoun (e.g., the third-person reflexive pronoun in the sentence *De lottowinnaar zei / ‘Ik heb besloten een rode auto te kopen voor mezelf’*). Participants were faster producing the reflexive pronouns when the comprehended pronoun matched the to-be-produced pronoun (e.g., comprehended *zichzelf* – produced *zichzelf*), than when they mismatched (e.g., comprehended *mezelf* – produced *zichzelf*). This occurred even though the presence of a match meant that the input sentence in direct speech was ungrammatical, and this should have led to a processing delay in comprehension. The authors concluded that participants’ expectations in comprehension were influenced by the indirect speech sentence they were concurrently encoding (where the same pronoun was perfectly grammatical).

More direct evidence comes from an fMRI study conducted by Menenti et al. (2011). They showed that brain areas that support semantic, lexical, and syntactic processing are largely shared between language production and language comprehension. For syntactic processing, Segaert et al. (2012) showed, in addition, that the same populations of neurons are recruited during language production as during language comprehension.
In another fMRI study, Stephens et al. (2010) compared activation in a speaker with activation in listeners attending to the speech produced by that speaker. The speaker’s and the listeners’ neural activity were not only spatially overlapping, but also temporally coupled. As might be expected, areas of the listeners’ brains were typically activated with some delay relative to the corresponding areas of the speaker’s brain. Intriguingly, some areas showed the opposite pattern: they were activated in the listener’s brain before they were in the speaker’s. These areas might be responsible for anticipatory processing of the sort that seems to be necessary for coordination (see Section 1.4.2). The size of areas showing anticipatory activity was positively correlated with listeners’ comprehension performance.

Kuhlen et al. (2012) reported that the listener’s EEG activity mostly lagged behind the speaker’s (as they were more strongly correlated at a lag of about 12.5 seconds), and proposed that this reflects alignment of larger semantic units, possibly corresponding to situation models. However, this conclusion might have been biased by the nature of the task they used. They exposed listeners to video recordings of two speakers, one superimposed on the other, and asked them to attend to only one of the two speakers. This was done to factor out correlations between the speaker’s and listener’s EEG activity due to purely perceptual factors. However, this somewhat artificial manipulation might have also masked any correlation between language-processing areas at shorter time scales.

1.4.1.4 Summary and discussion

In sum, there is strong evidence to support the assumption of parity between comprehension and production representations in language. This therefore suggests that one mechanism by which speakers coordinate their utterance is via a form of cross-person priming, in which representations that have just been used in comprehension are readily available for subsequent use in production. Note that this mechanisms appears to be largely but not entirely automatic, as indicated by the fact that motor activation during speech
comprehension varies depending on the nature of the task (see Section 1.4.1.1), and also by the fact that beliefs can affect lexical alignment (Branigan, et al., 2011).

However, it has been noted that indiscriminate alignment is not always an indicator of mutual understanding (Fusaroli, et al., in press). Interestingly, one study found that collective performance in a joint perceptual discrimination task, in which participants conversed to resolve disagreement about their perceptions, correlated with the degree to which participants aligned their linguistic choices in a domain that was relevant to the task (i.e., expressions of confidence in their perceptions), but not with the degree to which participants aligned their linguistic choices overall (Fusaroli et al., 2012).

In addition, alignment might not be sufficient to explain all forms of utterance coordination, and in particular turn-taking and cross-person compound contributions (Fusaroli, et al., in press; Poesio & Rieser, 2010). More generally, priming is not sufficient to explain coordination that takes the form of complementary utterances between speakers (e.g., question-answer pairs), rather than repetition of utterances (Fusaroli, et al., in press). A similar criticism could be moved to the shared representations account of acting together (see Section 1.3.2), and indeed the need to account for coordination of complementary action prompted proponents of this account to introduce an additional mechanism: prediction through internal simulation. Crucially, Pickering and Garrod (2007, 2013) proposed that (i) predictions of others’ utterances are formed during language comprehension, and that (ii) these predictions are computed using language production mechanisms. In the next section, we present their proposal and review the evidence available for these assumptions.

### 1.4.2 The integrated theory of language comprehension and language production

Pickering and Garrod (2007, 2013) first proposed that language production mechanisms could be used by comprehenders to make predictions during language comprehension. This claim is based on two assumptions. First, it assumes that a
comprehender can indeed predict what the speaker is about to say. Second, it assumes that a comprehender does so by using mechanisms that she also uses when she is producing language.

### 1.4.2.1 Prediction in language comprehension

The first assumption is supported by several lines of evidence (Huettig, Rommers, & Meyer, 2011; Kutas, DeLong, & Smith, 2011; Pickering & Garrod, 2007, 2013; Van Petten & Luka, 2012). For example, we know that listeners are more likely to look at the picture of a cake after hearing the verb *to eat* (but before hearing the word *cake*) than after hearing the verb *to move* (Altmann & Kamide, 1999). This indicates that they anticipate the most likely referent on the basis of the semantics of the verb (i.e., subcategorization information) and visual information.

Altmann and Kamide (1999) investigated anticipatory eye-movements using the visual-world paradigm (Huettig, et al., 2011; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). In this paradigm, listeners’ eye-movements to a visual display (consisting of real or pictured objects) are tracked in real-time while they attend to auditorily presented utterances. Several other visual-world studies showed that various sources of information, apart from semantics, can be used by listeners to guide anticipatory eye-movements. These include case marking (e.g., Kamide, Altmann, & Haywood, 2003), verb-specific syntactic information (Arai & Keller, 2013), prosody (Nakamura, Arai, & Mazuka, 2012), world-knowledge in combination with sentence meaning (Altmann & Kamide, 2007; Rommers, Meyer, Praamstra, & Huettig, 2013), and beliefs about the speaker’s preferences (Ferguson & Breheny, 2011). In addition, listeners appear to use the presence of disfluencies (Arnold, Hudson Kam, & Tanenhaus, 2007), and beliefs about the speaker’s identity (Arnold, et al., 2007) and previous language uses (Brown-Schmidt, 2009) to constrain the interpretation of temporarily ambiguous linguistic input, which is consistent with the idea that they integrate...
such information very rapidly in their interpretation of the unfolding utterance, though not necessarily via prediction.

As well as being based on various sources of information, prediction in language comprehension can take place at many linguistic levels. First, there is evidence that syntactic features of upcoming words can be predicted (Van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005; see also Wicha, Moreno, & Kutas, 2004). For example, Van Berkum et al. (2005) recorded EEG while participants listened to preambles like “De inbreker had geen enkele moeite de geheime familiekluis te vinden.” (The burglar had no trouble locating the secret family safe.). “Deze bevond zich natuurlijk achter een…” (Of course, it was situated behind a…). The preamble continued either with a highly predictable noun (schilderij, painting) or with a coherent but less predictable noun (boekenkast, bookcase). Importantly, the two nouns differed in gender (neuter vs. common) and were preceded by adjectives congruently marked for gender (groot maar onopvallend schilderij vs. grote maar onopvallende boekenkast, big but unobtrusive painting/bookcase).

Electrophysiological activity time-locked to the onset of the inflection of the first adjective (groot/grote) revealed differential activation in the two conditions. The authors concluded that listeners predicted the most likely noun (and its gender) based on context. When the adjective inflection violated their prediction, the listeners’ brain responded with an enhanced early positive deflection in the ERP signal. In addition, a self-paced reading study confirmed that comprehenders encountered difficulties before reaching the noun while they read sentences similar to the ones presented in the EEG experiment, although this effect emerged only at the second incongruently marked adjective (Van Berkum, et al., 2005).

Similarly, Wicha et al. (2004) had participants read Spanish sentences like “La historia de Excalibur dice que el joven rey Arturo sacó de una gran roca ...” (The story of Excalibur says that the young King Arthur removed from a large stone …), that were completed by a congruent (espada, a sword, feminine) or incongruent (cartera, a wallet, masculine) noun. The nouns were always preceded by a gender-marked determiner that
matched or mismatched the noun’s gender. Nouns mismatching in gender with the preceding
determiner (e.g., un espada/ una cartera) elicited an earlier negativity (N400) and a later
positivity (P600). P600 effects have been observed in response to gender violations in many
other studies in different languages (e.g., in Dutch, Hanuliková, Van Alphen, Van Goch, &
Weber, 2012). Crucially, an increased positivity in the 500-700 ms time-window was also
observed in response to the preceding determiner when it mismatched the gender of the
highly expected noun (e.g., un espada), showing that readers anticipated the noun’s gender.

Second, there is evidence that sound-related features of upcoming words can be
predicted as well. De Long et al. (2005) had English participants read sentence preamble like
‘The day was breezy so the boy went outside to fly ... ’; the preambles could be continued by
a highly predictable noun (e.g., a kite) or a less predictable noun (e.g., an airplane). They
reported that a negativity (N400) was elicited already at the article (an) when it was
incompatible with a highly predictable sentence continuation (kite). Moreover, the amplitude
of the article N400 correlated negatively with the article cloze probability given the
preceding context.

Related evidence comes from a study by Vissers et al. (2006) who had Dutch
participants read high and low cloze sentences that occasionally contained pseudo-
homophones. These are pseudo-words that are phonologically identical to the target word,
but are spelled differently (e.g., peil instead of pijl, arrow). Pseudo homophones elicited an
enhanced P600 effect relative to correct spellings, but only in high-cloze sentences. The
authors concluded that the P600 reflects a monitoring process that resolves conflict between
a predicted word form and incongruent input. Similarly, Severens and Hartsuiker (2009)
found an enhanced P600 in response to phonological errors in high-cloze sentences in a
reading experiment. Interestingly, the amplitude of this component was more pronounced in
response to errors that were pseudo-words (e.g., noom instead of droom, dream, in Dutch)
than to errors with lexical status (e.g., boom, tree).
Third, there is also some evidence that syntactic structures can be predicted and not just single words or their properties (Lau, Stroud, Plesch, & Philips, 2006; Staub & Clifton, 2006; Yoshida, Dickey, & Sturt, 2013). For example, in an eye-tracking reading study, Staub and Clifton (2006) showed that processing of linguistic material following *or* was facilitated in a sentence like “*John borrowed a rake or his wife bought one.*” when the sentence began with the word *either*. They interpreted this facilitation as evidence that readers actively anticipated a coordinative construction when *either* was present.

Furthermore, it is known that words that are less congruent with their sentential or discourse context elicit a neural response known as the N400. The same occurs for words that are congruent but less likely given the context, as measured by offline cloze tasks (Kutas & Hillyard, 1984). In addition, the N400 is smaller for less predictable words that are semantically related to a highly predictable word (e.g., Federmeier & Kutas, 1999), which is compatible with prediction of semantic features. However, scholars debate whether the N400 elicited by less predictable words reflects the computation of predictions, or the ease of integration with the preceding context (Van Petten & Luka, 2012). Interestingly, recent findings seem to support the view that it indexes predictions (Lau, Holcomb, & Kuperberg, 2013). Word predictability is also known to affect reading times in eye-tracking studies (see N. J. Smith & Levy, 2013 for a recent account of such effects).

Finally, as mentioned in Section 1.1, listeners can anticipate the timing of turn endings. De Ruiter et al. (2006) showed that listener’s accuracy in estimating the end of a turn was negatively affected by making speech unintelligible, but not by removing prosodic information. Interestingly, Cummins (2009) showed that speakers’ ability to synchronize with a recording was also impaired by disrupting intelligibility (as well as by altering other features of the speech signal). Further, Magyari and De Ruiter (2008, 2012) showed that listeners’ accuracy in De Ruiter et al.’s (2006) study correlated with the accuracy with which another group of participants could predict how the turn would end (i.e., how many and which words would be uttered next by the speaker).
1.4.2.2 Prediction in language comprehension uses production mechanisms

Given that comprehenders do anticipate upcoming linguistic input, Pickering and Garrod (2007) hypothesized that they might do so using language production mechanisms. Their account was modeled after internal simulation theories of action perception, and in particular after theories that propose that internal simulation is based on forward models (Grush, 2004; Wolpert, et al., 2003). Below, we will first present this account in its most recent formulation (Pickering & Garrod, 2013). Then we will review evidence that prediction in language comprehension is based on production mechanisms. Note that the evidence reviewed in Section 1.4.1.1 shows that production mechanisms are involved in comprehension but not specifically that they support the computation of predictions during language comprehension.

Pickering and Garrod (2013) proposed that people can comprehend others’ utterances using a mechanism called prediction-by-simulation. According to this, the comprehender first covertly imitates the incoming sensory input to recover a production command, a representation of the intended message so far. Having done so, the comprehender can then do three things (see Figure 6, p. 14). First, he can run the recovered production command through his own language production system (which Pickering and Garrod term the production implementer), thus producing an utterance that imitates the comprehended utterance. Importantly, production of the imitative response does not necessarily lead to overt articulation, as the response can be inhibited at any stage during the process.

Second, the comprehender can compute what production command he would be likely to use next if he were speaking (but correcting for differences between himself and the speaker). He can run the new production command through his own implementer, thus producing an utterance that continues the comprehended utterance (but again can inhibit this utterance at any stage during the process).
Third, he can feed a copy of the new production command (i.e., an *efference copy*) into a *forward model*. This third possibility is central to Pickering and Garrod’s (2013) account. In their account, forward models are predictive models of the language system. In particular, they distinguish between a forward production model and a forward comprehension model. The forward production model takes the efference copy as input and computes estimates of the semantics, syntax, and phonology of the to-be-produced utterance. Importantly, such estimates are distinct from production representations retrieved within the production implementer. So, for example, if the production command specifies the intention to name a colourful flying insect, the implementer retrieves the appropriate concept (BUTTERFLY), then a lemma (*butterfly*), and a phonological representation (/bʌtəflaɪ/), which is then used to guide articulation (cf. Figure 1-1 below). The forward production model, instead, might predict aspects of the concept (e.g., that it is concrete), of the lemma (e.g., that it is a noun), and of the phonological representation (e.g., that it composed of three syllables, or that it starts with a consonant). Such predictions can be approximate, but they are ready before the representations computed within the production implementer.

In addition, the estimates computed by the forward production model can be fed into a forward comprehension model. The forward comprehension model takes estimates of the state of the production implementer (i.e., predicted production representations) as input, and computes estimates of the state of the comprehension implementer (i.e., the language comprehension system) as output. The latter correspond to predicted aspects of the semantics, syntax, and phonology of the to-be-comprehended utterance (i.e., what it would be like to comprehend the word “butterfly” once it is produced), and can later be compared (within the *comparator*) to the actual comprehension representations computed within the comprehension implementer once the speaker produces the next utterance and this utterance is perceived by the comprehender.

Importantly, forward production and comprehension models are used both when we produce and when we comprehend language. In production, they predict linguistic aspects of
the speaker’s (upcoming) utterance and her experience of that utterance. In comprehension, they predict linguistic aspects of another speaker’s (upcoming) utterance and the comprehender’s experience of that utterance. However, since forward models are coupled with activation of the production implementer during language production, using a forward model to predict another speaker’s utterance in comprehension can also lead to some activation of the production implementer. Therefore, this theory predicts that language production mechanisms (forward models and, to a certain extent, the production implementer) should be implicated in prediction during language comprehension.

Is there evidence that production processes are involved in prediction?

Behaviourally, speakers are faster naming a picture (e.g., of a car) after reading a high-cloze sentence preamble (e.g., “George taught his son to drive a ...”), than after reading a medium-cloze sentence preamble (e.g., “The commercial was for a new...”), and high cloze preambles eliminate the disadvantage in naming latencies for low frequency picture names (Griffin & Bock, 1998). In addition, the anticipated word is often produced instead of the picture name (e.g., priest) when they are semantically related (e.g., anticipated: nun), and even when the anticipated word is a homophone of another word which is semantically related to the picture name (e.g., anticipated: none), as shown by Ferreira and Griffin (2003).

Note that these findings are not usually interpreted as evidence that prediction is based on production mechanisms. Indeed, they do not constitute direct evidence for this hypothesis. These findings suggest that making a word predictable in comprehension can facilitate its subsequent selection in production, which is consistent with the idea that prediction in comprehension is supported by production mechanisms. But on an alternative explanation, high-cloze sentence preamble might facilitate access of the picture names in comprehension thanks to the spreading of activation in the mental lexicon (without any need to assume a predictive mechanism is operating). Under the assumption of parity, lexical representations are shared between comprehension and production; therefore, subsequently re-accessing the picture names in production could be facilitated via priming.
More compellingly, Federmeier, Kutas, and Shul (2010) reported that a late prefrontal positivity induced by plausible but unexpected nouns (which is thought to index error correction and/or prediction updating; Federmeier, Wlotko, De Ochoa-Dewald, & Kutas, 2007) is greatly reduced in older compared to younger adults. Crucially, the magnitude of this component in the older group correlated with production measures of verbal fluency (see also DeLong, Groppe, Urbach, & Kutas, 2012). Similarly, Mani and Huettig (2012) found that 2-years-olds with larger production (but not comprehension) vocabularies were more likely to predict upcoming referents (by looking at corresponding pictures). These studies suggest that the ability or tendency to predict during language comprehension is correlated with language production abilities both in older adults and in children. However, this evidence is correlational and therefore not conclusive with regard to a causal link between production mechanisms and prediction in comprehension.

More evidence comes from D’Ausilio et al. (2011). They primed Italian participants with pseudowords containing double consonants whose pronunciation either involves (e.g., *birro*) or does not involve (e.g., *bivvo*) tongue movements. Participants then listened to target pseudowords that were the same as the primes on 75% of the trials (to encourage anticipation). TMS was delivered to the tongue motor areas 100 ms after the onset of the target pseudoword, and 300 to 350 ms before the onset of the double consonant. The authors varied whether the co-articulation information carried by the first part of the target (*bi*) was consistent with the up-coming consonant or not.

MEP’s recorded from the tongue showed increased excitability when the participants heard the target *birro*, but only if the first part of the item carried co-articulation information consistent with “rr” and not when the first part carried co-articulation information consistent with a consonant that does not require as much tongue involvement (“vv”). The time-course of this activation is compatible with prediction. Therefore, this finding suggests that language production processes might be implicated in prediction during listening. However, this study does not prove that language production processes are causally implicated in
prediction, because it did not test whether disrupting activation in the tongue area would impair participant’s performance in a perceptual prediction task.

So far, the evidence for production-based prediction in comprehension appears somewhat weak. The next section first reviews findings that implicate forward models in the production and imagination of utterances; then, it shows that prediction in comprehension can sometimes be very rapid, which provides indirect evidence for the implication of forward models during the comprehension of utterances. Finally, it describes two studies that provide more direct evidence for the implication of forward models in language comprehension.

### 1.4.2.3 Forward models are implicated in producing, imagining, and comprehending utterances

In the DIVA (Directions into Velocities of Articulators) and GODIVA (Gradient Order DIVA) models of speech production (e.g., Bohland, Bullock, & Guenther, 2010; Guenther, Ghosh, & Tourville, 2006), a forward model is used to compute the auditory representation corresponding to the current shape of the vocal tract, which in turn is derived from combined proprioceptive feedback and a copy of the motor command sent to the articulators. Another model of speech production that makes use of forward models is Hickok (2012), which also includes an internal feedback loop in which forward model predictions are used to detect and quickly correct errors in executed motor commands, before sensory feedback is available.

In an MEG study, Tian and Poeppel (2010) demonstrated that auditory cortex is activated very quickly (around 170 ms) when participants are asked to imagine themselves articulating a syllable. They therefore proposed that forward models involved in speech production can be decoupled from the movement of the articulators. Their findings open up the possibility that a forward model of the articulation system could be used in covert imitation of perceived speech.
In line with this proposal, another MEG study by the same authors showed that M200 responses in auditory temporal areas are enhanced when the participants have just articulated or imagined themselves articulating the same auditory stimulus (a syllable), compared to a different syllable (Tian & Poeppel, 2013). They argued that this occurred because overt and covert articulation involve the computation of forward models; specifically, articulation activates forward models involved in the computation of somatosensory predictions, which are in turn used to fine-tune auditory predictions, thus leading to enhancement of the M200 in auditory areas (Tian & Poeppel, 2012).

Interestingly, however, they also found that imagining to hear somebody produce a syllable has the opposite effect on the M200, that is, responses are suppressed for repeated compared to novel stimuli. They proposed that this effect reflects a different, memory-based route in perception, as response suppression was also observed after the participants actually heard the stimulus, though it occurred earlier, in the M100 (Tian & Poeppel, 2013). Note that Pickering and Garrod (2013) have also proposed that comprehension can sometimes proceed through a route that does not involve production mechanisms, which they called prediction-by-association. We will not describe this route here (but see Gambi and Pickering, 2013a for a discussion of when one route might be preferred over the other).

In any case, the results of Tian and Poeppel (2013) strongly suggest that forward model predictions can modulate early responses during speech perception. In addition, other studies suggest that during reading comprehenders might compute predictions of the visual forms of words (e.g., Dikker, Rabagliati, Farmer, & Pylkkänen, 2010; Kim & Giley, 2013). This evidence has been reviewed by Pickering and Garrod (2013, p. 15). It suggests that language processing areas can activate visual areas in a top-down manner very rapidly. However, note that these studies do not address the issue of whether language production mechanisms are used in the computation of predictions during language comprehension. Therefore, they can provide only indirect evidence for the involvement of forward models in language comprehension.
Remember that in Section 1.3.3 we mentioned that one piece of evidence in favour of an internal simulation approach is the fact that prediction of one’s own actions appears more accurate than prediction of actions performed by others. Unfortunately, we are not aware of any study that investigated this issue directly in the case of language. However, one study reported that people are better at lip-reading themselves than others, which is at least consistent with this hypothesis (Tye-Murray, Spehar, Myerson, Hale, & Sommers, 2013).

Finally, the most compelling piece of evidence that forward-model predictions are causally implicated in language comprehension comes from a recent TMS study (Lesage, Morgan, Olson, Meyer, & Miall, 2012). In this study, participants looked at visual displays while listening to sentences, as in Altmann and Kamide’s (1999) study on anticipatory eye-movements. Repetitive TMS was administered to the right cerebellum, which is thought to be the neural site of forward model computations (see Section 1.3.3). Crucially, it was found that disrupting activity in this area impaired listeners’ ability to launch anticipatory eye-movements to upcoming sentential referents on the basis of verb semantics.

In sum, different lines of evidence converge to support the idea that production mechanisms are used to generate predictions during language comprehension, and that these mechanisms might be forward production and comprehension models as hypothesized by Pickering and Garrod (2013). However, only one study to date has provided evidence for the hypothesis that forward models are causally related to the ability to compute predictions during language comprehension (Lesage, et al., 2012).

1.4.2.4 Forward models in dialogue

Pickering and Garrod (2013) discussed how the use of prediction-by-simulation might aid coordination in dialogue. First, if interlocutors mutually predict one another using production mechanisms, they can send the production commands derived for their partner’s utterance to their own production implementer; this would allow them to quickly produce appropriate overt responses, which might explain the smoothness of turn-taking and the
existence of cross-person compound contributions. Second, prediction might underlie alignment of representations (Pickering & Garrod, 2004). A version of the latter idea was also proposed by Jaeger and Snider (2013), according to whom listeners and speakers mutually adapt their expectations about the likelihood of occurrence of particular structures in the input. Although their approach links prediction and alignment, they do not commit to any specific cognitive mechanism underlying the computations of expectations.

Importantly, the use of the prediction-by-simulation route is likely to be facilitated in dialogue, for at least two reasons. First, many natural conversations support accurate predictions, because interlocutors can capitalize either on a long interactional history that leads to alignment (Pickering & Garrod, 2004) or on the scriptedness of the activity type (e.g., purchasing an item in a shop, fixing an appointment at the doctor; H. H. Clark, 1996; Levinson, 1992). In particular, prediction-by-simulation tends to be successful in dialogue because interlocutors become sufficiently similar during the interaction, to the extent that they can predict what the other is about to say by predicting what they themselves would say next (without the need for corrections). Second, production processes are likely to be more activated in natural conversations where interlocutors switch between the roles of speaker and listener all the time (Pickering & Garrod, 2013).

Overall, the assumption that comprehenders use prediction-by-simulation both subsumes and extends the interactive alignment theory. In doing so, it also responds to a limitation of the interactive alignment model, that is, its inability to explain the coordination of complementary actions. Therefore, in the next section, we will present a cognitive architecture for the coordination of utterance that is based on Pickering and Garrod’s (2013) integrated theory of comprehension and production. Our cognitive architecture further specifies the theory by making the explicit hypothesis that forward model predictions can be of two kinds: predictions related to the content of production representations and predictions related to the timing with which such representations are built within the production.
implementer. Importantly, we argue that this theory, unlike Clark’s (1996), is mechanistic in that it describes a set of cognitive mechanisms that could underlie utterance coordination.

1.4.3 A cognitive architecture for the coordination of utterances

We propose that interlocutors coordinate their utterances via three mechanisms: (i) they represent others’ utterances in a similar format as their own utterances; (ii) they use these representations as a basis for prediction; and (iii) they integrate self- and other-representations on-line. To illustrate, consider two speakers, A (female) and B (male), producing two utterances roughly at the same time, in response to a shared stimulus, such as a to-be-named picture of a kite. Figure 1-1 illustrates the range of information that A could represent about her own utterance (upper box) and about B’s utterance (lower box).

Before we discuss the nature of these representations, we will briefly illustrate the time course of word production, taking A’s production of “kite” as an example (see the timeline at the top of Figure 1-1). Models of single word production (e.g., Levelt, et al., 1999) involve at least (i) a semantic representation (semₐ) corresponding to the target concept (KITE); (ii) a syntactic representation (synₐ) – sometimes called a lemma – that incorporates syntactic information about the lexical item, such as that it is a noun (kiteₐ/N); (iii) a phonological representation (phonₐ) that specifies a sequence of phonemes and its syllable structure (/kaIt/). Finally, the appropriate articulatory gestures are retrieved and executed (artₐ).
\[ \text{sem}_A = \text{KITE} \quad \text{syn}_A = \text{kite}_m \quad \text{phon}_A = /\text{kalt}/ \quad \text{art}_A \text{ (onset)} \]

\[ \hat{p} (\text{sem}_A) \rightarrow \hat{t} (\text{sem}_A) \]
\[ \hat{p} (\text{syn}_A) \rightarrow \hat{t} (\text{syn}_A) \]
\[ \hat{p} (\text{phon}_A) \rightarrow \hat{t} (\text{phon}_A) \]
\[ \hat{p} (\text{art}_A) \rightarrow \hat{t} (\text{art}_A) \]

A's production system

"kite"
Figure 1-1. Simultaneous production. A produces the word *kite* in response to the picture of a kite. sem_A, syn_A, phon_A are semantic, syntactic, and phonological representations for A’s utterance. t(sem_A), t(syn_A), t(phon_A) indicate the actual time elapsed (from picture onset) when processing is completed at each stage and the corresponding representation has been built (based on Indefrey and Levelt, 2004); t(art_A) marks the onset of A’s utterance. f_(sem_A), f_(syn_A), f_(phon_A), and f_(art_A) are timing estimates computed by A for her own utterance. p_(sem_A), p_(syn_A), p_(phon_A), p_(art_A) are the content predictions for A’s own utterance, on which the timing estimates are based. A believes that B is speaking in response to the same picture. Dotted lines refer to representations of the other. f_(sem_B), f_(syn_B), f_(phon_B), and f_(art_B) are timing estimates computed by A for B’s utterance. p_(sem_B), p_(syn_B), p_(phon_B), p_(art_B) are A’s content predictions, at the various processing stages, for B’s utterance.

Horizontal arrows (from p_(sem_A) to f_(sem_A), from p_(syn_A) to f_(syn_A), etc.) indicate that estimates of the timing at each level are based on content predictions at the same level. Timing estimates at one level could also be directly based on content estimates at other levels, but we ignore this here for simplicity. Vertical arrows from self- and other-predictions to planning represent the integration stage.

Note that each processing level is characterized not only by the content of the associated representation, but also by its timing (t(sem_A), t(syn_A), etc.). Some representations are typically ready before others and the processing stages take different amounts of time. Indefrey and Levelt (2004) derived indicative time windows from a meta-analysis of several word production experiments. Their estimates are also reported at the top of Figure 1-1, though the exact times might depend on the words used or the experimental conditions (cf. Sahin, Pinker, Cash, Schomer, & Halgren, 2009, for estimates based on intracranial electrophysiological recordings).
1.4.3.1 Predicting one’s own utterances

Now, consider the upper box of Figure 1-1. We assume that $A$ can generate predictive estimates of the duration of each processing stage (indicated by $f$ in Figure 1-1). For example, she might generate the estimate $f(syn_A) \approx 250$ ms, meaning that she predicts retrieving the syntactic representation will take approximately 250 ms (from picture onset). These estimates can in turn be exploited by $A$ to guide planning of her own utterance. Interestingly, some studies have shown that individual speakers can coordinate the production of two successive utterances so as to minimize disfluencies (Griffin, 2003; cf. Meyer, Belke, Häcker, & Mortensen, 2007). Similarly, Meyer et al. (2003) demonstrated that the amount of planning speakers perform before articulation onset can depend on the response time deadline they implicitly set for their performance at a naming task. This suggests that timing estimates are computed for one’s own utterances and can be used to guide planning.

Clearly, for a speaker to be able to use the information provided by timing estimates effectively, the estimates must be ready before processing at the corresponding stages is completed. So, for instance, the estimate $f(syn_A) \approx 250$ ms is useful only if it is available before syntactic processing is complete. This means that the estimates are predictions. What are such predictions based on? Importantly, in language production, timing aspects are known to be closely related to the content of the computed representations. For example, word frequency affects $t(phon_A)$, with phonological retrieval being slower for less frequent word forms (e.g., Jescheniak & Levelt, 1994). We therefore assume that $A$ predicts aspects of the content of $sem_A$, $syn_A$, and $phon_A$. In other words, the speaker anticipates aspects of the semantics, syntax, and phonology of the utterance she is about to produce, before the representations corresponding to each level are built in the course of the production process itself. To distinguish these predictions that relate to content from predictions that relate to timing (i.e., the timing estimates), we label them $\hat{p}(sem_A)$, $\hat{p}(syn_A)$, $\hat{p}(phon_A)$, and $\hat{p}(art_A)$. 
Interestingly, recent studies indicate that the amount of planning speakers perform in advance of articulation (i.e., the scope of utterance planning) is not fixed, but varies flexibly depending on a series of factors. In particular, Konopka (2012) showed that speakers use larger planning scopes when reusing a sentence structure that has just been primed than when using a non-repeated structure. In addition, speakers might vary planning scope strategically, as they also used a larger planning scope when both the structure and the first lexical item they had to retrieve were difficult, perhaps because they began retrieving the second lexical item instead. This might indicate that speakers can anticipate production difficulties and one way in which speakers could anticipate difficulties is by predicting aspects of what they are about to say.

Further, speakers often choose structures that appear to minimize planning difficulty, as when they build utterances that agree with the end-weight constraint, that is, the tendency of putting short constituents before long ones (e.g., Bresnan & Ford, 2010). Pickering and Garrod (2013) proposed that this phenomenon could be accounted for if speakers were able to estimate the length of to-be-produced constituents (using forward models), and use such estimates when they apply the end-weight constraint. Interestingly, Wasow (1997) proposed that the constraint itself facilitates production by keeping structural options open; Pickering and Garrod (2013), however, suggested that it might be related to speakers attempting to keep information density constant, in accordance with Jaeger (2010).

Speakers also tend to choose word orders that allow them to avoid producing two phonologically related nouns in close proximity (Jaeger, Furth, & Hilliard, 2012). This finding could similarly be accounted for if speakers were able to predict enough of the phonology of the words to anticipate their similarity and the consequent production difficulties. Note, however, that timing and content predictions for self-generated utterances need not always be as detailed as some of these studies may suggest. The specificity of predictions might depend on task demands (e.g., whether fine-grained control over the production process is needed) and be highly variable.
Having posited that predictions of timing and content can be generated for one’s own utterances, we now propose that representing others’ utterances can also involve the computation of predictions, and that those predictions are in a similar format to the timing and content predictions for self-generated utterances. The lower (dashed) box in Figure 1-1 shows the range of information that A could represent about B’s utterance. Importantly, A may well not represent all of this information under all circumstances; our aim is to provide a comprehensive framework in which such information can be represented. Later, we describe experimental paradigms that can investigate the conditions under which aspects of B’s utterance are represented and how.

First of all, A could estimate the time course of B’s production. Minimally, A could compute \( t^\hat{}(art_B) \), an estimate of B’s speech onset latency. In addition, A might compute timing estimates for the different processing stages, from semantics to phonology (\( t(\text{sem}_B), t(\text{syn}_B), t(\text{phon}_B), \) and \( t(art_B) \) in Figure 1-1), just as she does when anticipating the timing of her own productions. As timing estimates are likely to be based on information regarding the content of the computed representations, and in line with Pickering and Garrod (2013), we suggest that A can also represent the content of B’s utterance. In particular, A builds predictive representations of the semantics, syntax, and phonology of the utterance produced by B (\( p(\text{sem}_B), p(\text{syn}_B), p(\text{phon}_B), \) and \( p(art_B) \) in Figure 1-1).

### 1.4.3.2 Predicting another’s utterances

Following Pickering and Garrod (2013), we also propose that people make content and timing predictions, for both self-generated and other-generated utterance, using forward models of their own production implementer. Our focus will be on forward production models, as we are mainly interested in how forward model predictions can affect ongoing processes in the production implementer. When forward production models are used to simulate the production implementer of another speaker, they are decoupled from the production implementer of the comprehender, so that covertly simulating another’s
utterances does not lead to the actual planning of that utterance or to its articulation. In other words, A does not build $\text{sem}_B$, $\text{syn}_B$, and $\text{phon}_B$ (semantic, syntactic, and phonological representations for the utterance that B is going to produce) just as she does not initiate $\text{art}_B$ (the articulation stage for B’s utterance).

Clearly, however, speakers can overtly imitate a speaker (e.g., in speech shadowing; see Marslen-Wilson, 1973) and they can complete each other’s utterances. On occasion, therefore, covert simulation of B’s utterance, via the computation of a forward model, results in activation of A’s own production implementer. In this case, there will be activation of the semantic ($\text{sem}_B$), syntactic ($\text{syn}_B$), and phonological ($\text{phon}_B$) representations corresponding to B’s to-be-produced utterance, within A’s production system. Depending on the predictability of B’s utterance, and on the speed of the simulation, A might end up shadowing B’s speech, talking in unison with B or even anticipating a completion for B’s utterance.

Note, however, that some activation of A’s production implementer does not necessarily entail that A overtly articulates B’s utterance, possibly because the activated representations are inhibited (see Section 1.4.2). In addition, it is possible that forward model predictions could sometimes be computed without any activation taking place in the production implementer. If this is correct, predicting B’s utterances could sometimes not involve any (detectable) activation flow in A’s production implementer. At present, determining exactly under which conditions A’s production system is activated, and to what extent, is still a matter for empirical investigation.

Another important issue relates to the accuracy of both the timing and content representations of another’s utterances. For example, how similar is $\hat{p}(\text{sem}_B)$ to B’s concept KITE, or how accurate an estimate of B’s speech onset latency is $\hat{t}(\text{art}_B)$? We expect representations of another’s utterances to be generally somewhat inaccurate. First, although context and task instructions might highly constrain the productions of both speakers in experimental settings, normally A would have only limited information regarding what B intends to say. Second, A has limited experience of other speakers’ production implementers.
The forward model she uses to compute predictive estimates is fine-tuned to her own production implementer rather than to B’s production implementer (Wolpert et al., 2003). As a consequence, timing estimates based on a model of A’s production implementer are likely to diverge from the actual time course of B’s production. The degree of error will also depend on how much B differs from A in speed of information processing.

Conversely, we expect accuracy to increase the more A’s and B’s systems are or become similar. Importantly, as proposed by Pickering and Garrod (2013), the two implementers might become increasingly attuned via alignment (Pickering & Garrod, 2004), thanks to priming channels between the production and comprehension systems of the two interlocutors. Alignment could function as coordination smoother, as proposed by Vesper et al. (2010) for the coordination of actions.

Finally, we might ask whether predictions about other-generated utterances can influence the planning of one’s own utterances to the same extent as predictions about self-generated utterances. For example, say that f(art_A) is a prediction of when A will finish articulating her current utterance. A should take this prediction into account as she plans when to start her next utterance. Similarly, if B is the current speaker and A wants to take the next turn, A could compute f(art_B), an estimate of when B will stop speaking. Then the question is, will A pay as much attention to f(art_B) as she would to f(art_A) in the first case? This is likely to depend on the circumstances. For example, f(art_B) might be weighted as less important if its degree of accuracy is low (i.e., previous predictions have proved to be wrong). Alternatively, A might not take f(art_B) into account, simply because she does not share a goal with B; for example, she might be trying hard to be rude and interrupt B as much as possible. Nevertheless, the crucial point is that such predictions could influence ongoing production processes within A’s production implementer, because they are computed using the same mechanisms she uses to anticipate aspects of her own upcoming utterances.
1.4.3.3 The time course of predictions

What is the time course of predictions, both with respect to one another and to the time course of word production? Firstly, predictions should be ready before the corresponding production representations are retrieved in the process of planning an utterance. Secondly, since we assumed that timing estimates are computed on the basis of content predictions, $\hat{p}(\text{sem}_a)$ should be ready before $f(\text{sem}_a)$, $\hat{p}(\text{syn}_a)$ before $f(\text{syn}_a)$, etc. Similarly for other-predictions, $\hat{p}(\text{sem}_b)$ should be ready before $f(\text{sem}_b)$, $\hat{p}(\text{syn}_b)$ before $f(\text{syn}_b)$, etc (see horizontal arrows in Figure 1-1).

However, we do not make any specific claim about the order in which predictions at the different levels (semantics, syntax, and phonology) are computed. It might be tempting to stipulate that the prediction system closely mimics the production implementer in this respect. In fact, however, the prediction system is a (forward) model of the production implementer and such a model need not implement all aspects of the internal dynamics of the modeled system. In particular, the prediction system for language could involve the same representational levels as the production implementer, but the time course with which predictions are computed could differ from the time course of language production (Pickering & Garrod, 2013).

Predictions at the levels of semantics, syntax, and phonology might even be computed separately and (roughly) simultaneously (Pickering & Garrod, 2007). As proposed by Pickering and Garrod (2013), there could be separate mappings from the intention to communicate to semantics, syntax, and phonology. For this reason, in Figure 1-1 we simply list the different predictions. Nevertheless, it is certainly the case that predictions at different levels are related to each other. For example, a prediction that the upcoming word refers to an object (a semantic prediction) and that it is a noun (a syntactic prediction) are related (because nouns tend to refer to objects). It is likely that the prediction system for language exploits such systematic relations between levels.
Once predictions are computed, how are they integrated in the process of planning an utterance (cf. vertical arrows in Figure 1-1)? To illustrate, take the following situation. The speaker needs to initiate articulation ($\text{art}_i$) rapidly, perhaps because of task instructions (in an experiment) or because of an impatient listener trying to get the floor. She might therefore try to minimize $t(\text{art}_i)$, that is the time it takes her to start speaking. If $t(\text{phon}_i)$ exceeds a certain value, it might not be possible for the speaker to complete retrieval of the phonological representation before she starts to speak. How can the speaker make sure that she meets her goal of starting to speak as quickly as possible given $t(\text{phon}_i)$?

Using a forward production model, she could compute $\hat{p}(\text{phon}_i)$, a prediction of the phonology of her intended utterance. On the basis of this, the speaker could estimate, $f(\text{phon}_i)$, roughly how long it will take her to retrieve the phonological representation for that utterance (e.g., that it will take longer if the intended utterance is longer). She could then use this estimate to anticipate whether she is likely to meet her goal or not. If she anticipates a problem (e.g., if $f(\text{phon}_i)$ exceeds her intended $t(\text{art}_i)$, or a certain threshold), she could correct her production command accordingly.

For example, she could invest more resources in planning the utterance to speed up lexical retrieval, or, if processing speed is already at limit, she could articulate the initial part of her utterance, even if the rest has not been retrieved yet (Meyer et al., 2003). In other words, predicted outcomes (i.e., the output of the forward model) can be used to trigger corrections to the ongoing planning process within the production implementer, in case such outcomes do not correspond to the intended goal (see Section 1.3.3 on the use of forward models in planning).

This example illustrates how forward model predictions of one’s own utterances might affect the retrieval of representations within one’s own production implementer. Our account proposes that forward model predictions of another’s utterance could also similarly affect the retrieval of representations within one’s own production implementer, as predicting another’s utterance makes use of the same mechanisms used in anticipating one’s
own utterance. This would in turn explain how interlocutors coordinate their utterances in real time.

Having outlined this theoretical framework, in the next section we describe some experimental paradigms that can help answer the questions raised by this new approach. In fact, we believe that the inadequacy of other accounts of utterance coordination is partly due to the limitations associated with current experimental studies of dialogue. Current paradigms available for the study of dialogue look at how coordination is achieved off-line, over quite long stretches of conversation, using measures such as changes in turn length or choice of referring expressions. Under these circumstances, time constraints might be loose enough to allow for relatively slow and intentional cognitive processes to be the basis of coordination (e.g., H. H. Clark & Wilkes-Gibbs, 1986; Wilkes-Gibbs & Clark, 1992). Studies that focus on alignment have reduced the time scale to consecutive utterances (e.g., Garrod & Anderson, 1987; Branigan, et al., 2000). However, this is still a relatively long time-scale.

More recently, some studies have used a dialogue intervention paradigm (e.g., Howes, et al., 2012), in which the transfer of information between interlocutors is altered in real time. For example, turns can be artificially truncated to investigate under what conditions listeners are more likely to complete their partners’ utterances (see section 1.1). Within this paradigm, it is possible to investigate coordination within an utterance. However, questions about the timing of utterance coordination cannot be addressed, because there is little control over the utterances generated by participants. Furthermore, this paradigm requires the use of written dialogue, where the nature of the coordination is arguably different than in spoken dialogue.

In contrast, no study has looked at that moment-by-moment coordination that might explain how listeners and speakers synchronize and take turns so smoothly. We argue that the obvious way to do this would be to conduct experiments with more than one speaker in which properties of utterances that are known to affect the process of lexical retrieval or
sentence planning are systematically varied. We would then be able to test whether aspects of others’ utterances are indeed anticipated using language production mechanisms and to what extent such predictions are taken into account when planning one’s own utterances. Therefore, these experiments should combine a joint task context with controlled tasks like the ones that have been used to study the psycholinguistics of monologue.

1.5 Methodological rationale: comparing self- and other-representations in joint language tasks

How can we test whether the proposed account is correct? In analogy with the joint action literature reviewed in Section 1.2, we need to compare individual production and joint production. Specifically, we need to test to what extent preparing to produce an utterance and anticipating that another is about to produce an utterance affect concurrent language production processes in a similar manner. Below, we consider two instances of joint production: simultaneous production (section 1.5.1) and consecutive productions (section 1.5.2). In both sections, we first introduce the rationale behind joint production tasks and then present the general predictions of the model outlined in Section 1.4.3.

1.5.1 Simultaneous production

Consider two speakers preparing to produce two different or similar utterances at the same time (see Figure 1-1). If $A$ represents $B$’s utterance using language production mechanisms, then her act of production should be affected by her representation of his utterance, even if there is no need for coordination; the same holds for $B$’s representation of $A$’s utterance. This situation is analogous to the one investigated in studies that asked non-interacting participants to perform actions at the same time as each other (e.g., Sebanz, et al., 2005). By manipulating the relationship between the two speakers’ utterances (e.g., whether
they produce the same or different utterances), we can further investigate the nature of A’s representations of B’s utterances.

Overall, if other-generated utterances are represented using some of the same mechanisms employed during language production, we expect such representations to affect concurrent production of the speaker’s own utterance. Although it might be possible to maintain two representations active in parallel (M. Wilson & Knoblich, 2005; Wolpert, et al., 2003), for one’s own and one’s partner’s utterances, it is likely that using the same mechanism simultaneously for simulating oneself and another will lead to some form of competition, perhaps because of conflict with regard to whose turn it is to speak (cf. Wenke, et al., 2011).

In addition, if other-generated utterances are represented using the full range of mechanisms employed during language production, then we expect representations of B’s utterances to interact with representations of A’s own utterances in the way that representations for different, concurrently activated, self-generated utterances should interact. What would be the effect of such interaction on the production of A’s own utterance? It might lead to facilitation or interference, depending on a variety of factors (e.g., whether B is producing the same word or a different word; in the latter case, whether the two words are related in form or meaning; cf. Schriefers et al., 1990).

Besides, since some representations are harder to retrieve than others, variables that affect retrieval difficulty of self-generated utterances should also exert an effect in relation to other-generated utterances. Consider, for instance, the following situation. A and B name different pictures. The frequency of picture names is varied, so that on some trials B produces low-frequency words, whereas on others he produces high-frequency words. Given that it is harder to access the phonological representation of a low-frequency word than a high-frequency word (cf. Miozzo & Caramazza, 2003), representing B’s utterance should interfere more with A’s naming in the low-frequency condition than the high-frequency condition.
To sum up, paradigms that involve two speakers’ simultaneously or near-simultaneously producing utterances serve two purposes: they test whether self- and other-generated utterances are represented in a similar way (i.e., using language production mechanisms), and they can elucidate the nature of other-representations, in particular whether they involve the activation of one’s own production implementer to the same extent as representations for self-generated utterances. Study 1 of this thesis tested precisely these predictions of the account (see Chapter 2).

1.5.2 Consecutive productions

One concern with the study of simultaneous production is that it is comparatively rare in real conversations (see Section 1.1). Of course, speakers do occasionally contribute at the same time, for example when two listeners both claim the ground (e.g., in response to a question; M. Wilson & Wilson, 2005) or in intended choral co-production (e.g., mutual greetings; Schegloff, 2000). But it may be that speakers do not need a system that is specialized for representing their own utterance and a simultaneous utterance by their partner.

In contrast, consecutive production occurs all the time in conversation (see Section 1.1). Thus, speakers have much more need of representing their own utterance and their partner’s upcoming utterance. Consecutive production paradigms should then somewhat mimic the naturalistic situation exemplified by cross-person compound contributions. For example, A and B could be shown two pictures (e.g., of a wig and of a carrot), one on the right and one on the left of a computer screen. A first names the left picture (wig); then B names the right picture (carrot; see Figure 1-2, Panel A).
Figure 1-2. Consecutive utterances: pictures of a wig and a carrot appear simultaneously. (A) JOINT: A names the left picture, then B names the right picture. (B) SOLO: A names the left picture, then A names the right picture. (C) NO: A names the left picture. art$_B$ stands for the articulation stage of B’s utterance, where present. Where two utterances are produced, we indicate the temporal relation between them by way of number subscripts (1 for the first utterance, 2 for the second utterance). $\hat{\text{p}}$(sem$_B$) is the semantic content prediction that A generates in relation to B’s utterance. All other details as in Figure 1-1.

Figure 1-2 (Panel A) presents a schematic description. Given the complexity of the situation, in order to ensure that the figure is readable, we illustrate what happens from the perspective of A, the speaker that names the first picture. An additional reason to focus on A is that, because she is the person who speaks first, she cannot comprehend any aspects of B’s utterance before or while she is speaking. This means that any effect of B’s utterance on A’s utterance must be due to A forming a predictive representation of B’s utterance. The timeline at the top shows the time course of word production for A’s utterance (and the onset of B’s utterance). Just as for the simultaneous production paradigm, we assume that A generates timing estimates for her own utterance and that these estimates are based on content predictions (left box). In addition, we hypothesize that A represents B’s upcoming utterance in the same format and computes timing estimates and content predictions for that utterance, as well (right box).

To test these hypotheses, we again compare joint tasks with solo tasks. In the solo task (see Figure 1-2, Panel B), A produces both pictures’ names. Clearly, A goes through all the processing levels for both words and builds production representations at each level. The timeline at the top of Panel B differs from the one in Figure 1-1: most notably, t(art$_A$) corresponds to 1200 ms, instead of the 600 ms posited by Indefrey and Levelt (2004). This reflects the finding that participants tend to delay the onset of the first word, presumably
because they perform advance planning of the second word (Griffin, 2003). We also assume that \( A \) computes timing estimates and content predictions for the second word, as well as for the first word.

If other-generated utterances are represented using the same mechanisms employed during language production, then content and timing predictions computed for \( B \)’s utterance in the JOINT condition should be similar to those computed for \( A \)’s own second utterance in the SOLO condition. We would therefore expect the JOINT and the SOLO condition to show similar patterns of results. We know that speech onset latencies depend on various properties of the planned material, such as its length (Meyer et al., 2003) or syntactic complexity (e.g., Ferreira, 1991); the likelihood of being disfluent is also larger at the beginning of longer than at the beginning of shorter constituents (H. H. Clark & Wasow, 1998). In addition, there is some evidence that the duration of spoken articulation of words and constituents also depends on properties of the upcoming linguistic material (Gahl & Garnsey, 2004; Kuperman & Bresnan, 2012; Tily et al., 2009).

Therefore, speech onset latencies, likelihood of disfluencies, and duration of the first word might be affected by properties of the second word in the SOLO condition. This would reflect an influence of predictions of the second word’s features on the production of the first word. In the JOINT condition, we predict \( A \)’s speech onset will be similarly affected, despite the fact that the second word is actually produced by \( B \). This would show that predictions of the second word’s features are computed using the same mechanisms regardless of whether the second word is generated by the same or a different speaker.

Additionally, the JOINT condition could be usefully contrasted to the NO condition, depicted in Panel C of Figure 1-2. The NO condition is equivalent to an instance of isolated production of a single word by \( A \). Importantly, \( A \)’s task is the same in the NO and the JOINT conditions (i.e., producing Utterance 1), the only difference being that \( B \) does not produce Utterance 2 in the NO condition. The NO condition can therefore act as a control in which a
null effect is expected. Chapter 3 and 4 in this thesis present five experiments that are based on variations of this consecutive production paradigm.

1.6 Summary

To summarize, we propose a cognitive architecture for the coordination of utterances in which coordination is primarily achieved via a mechanism of anticipation, of one’s own and another’s utterances, that makes use of language production processes. If our account is correct, we expect speakers to form representations of other speakers’ utterances that are in a similar format to representations of their own utterances. Therefore, we expect other-representations to affect concurrent production. We investigated this hypothesis in three studies, which used either the simultaneous production paradigm (Study 1, Chapter 2), or the consecutive production paradigm (Study 2, Chapter 3, and Study 3, Chapter 4).
2. Study 1 Experiments 1-4: simultaneous speaking and imagining

2.1 Introduction

Is the way in which people speak affected by their beliefs about other people’s speech? To address this question, we asked whether performance in a joint picture naming task differs depending on whether the speaker is told that his partner is concurrently performing the same or a related task, or no task. This would provide evidence that speakers do not separate themselves from their partners, but instead represent their partner’s speech as well as their own, even when the partner is absent and their speech is not perceived but only imagined. Importantly, it would also suggest that speakers’ representations of their partners’ speech are in the same format as speakers’ representations of their own speech (as the former can affect the latter).

We do not know whether speakers form representations of an absent partner’s utterance or whether such representations are in the same format as their own utterances. One reason to believe the latter claim is that people appear to represent their own and other people’s language in the same way. In other words, there is evidence that language production and comprehension share representations (the parity hypothesis; see Section 1.4.1).

For example, in the picture-word interference paradigm participants name pictures while ignoring written (or auditory) distractor words. Responses are fastest when the

---

2 Experiment 1 in this study was designed and carried out by the author in collaboration with Joris Van de Cavey, while he was visiting the Department of Psychology at the University of Edinburgh as part of an MSc Program in Experimental Psychology at the University of Ghent, Belgium. This chapter is based on a manuscript under review in Journal of Experimental Psychology: Learning, Memory, and Cognition (Gambi, C., Van de Cavey, J., & Pickering, M.J. (2013). Naming is inhibited by the belief that one’s partner is also naming: A joint interference effect in picture naming. Unpublished manuscript, The University of Edinburgh.)
distractor word is the picture’s name. They are slower when the distractor is an unrelated word and slowest when it is a semantically related word (e.g., Glaser & Düngelhoff, 1984; Schriefers, et al., 1990). This shows that comprehension of the distractor word can affect production of the target word (at least under the right task conditions). But even if parity is correct, we do not know whether comprehending another person’s utterance and imagining such an utterance have comparable effects on production.

Interestingly, as discussed in Section 1.4.2.3, Tian and Poeppel (2013) showed that imagining oneself produce a syllable and imagining to hear someone else produce a syllable have different effects on perception of the same syllable. Here, we ask what effect imagining somebody else produce a word has on production of the same or a different word. To answer this question, we combined a picture naming task with a joint task paradigm, as suggested in Section 1.5.1. By varying instructions in the joint task, we manipulated the participants’ beliefs about the task their partner was performing while they were naming pictures.

It is known that participants represent their partner’s task as well as their own in joint task paradigms (see Section 1.2.2). Importantly for the current study, they do so if they are not interacting with their partner and might in fact be seated in different rooms (see, for example, Experiment 3 in Atmaca, et al., 2011). In other words, the belief of acting with another person is sufficient to induce representations of that person’s task in the participant. If speakers similarly represent an absent partner’s utterances, and if they do so in a way that is similar to how they represent their own utterances, then they should be affected by these representations while producing their own utterances.

In this study, we tested two speakers simultaneously, but they were seated in separate rooms and could not hear each other. However, they could infer what their partner was about to say because they could see their instructions as well as their own (i.e., they held beliefs about their partner’s task). Therefore, we tested whether representations of others’ utterances are formed in a non-interactive joint task setting, in which two individuals produce language alongside each other, but are not using language to communicate.
Participants were presented with pairs of pictures (e.g. of an apple and of a blouse) and asked to produce the pictures’ names while their partner either named the pictures or remained silent. We varied the relationship between the participants’ responses: they either produced the same response as their partner (e.g., they both said “apple blouse”), or a different response (e.g., one participant said “apple blouse” while the other said “blouse apple”). In addition, the two pictures were either semantically unrelated (e.g., apple blouse) or related (e.g., apple banana). In other conditions, one partner did not respond (or responded in a way that did not involve naming). These manipulations allowed us to test whether representations of others’ utterances are formed and, in addition, they allowed us to investigate whether representations of others’ utterances are content-specific – whether people represent what their partner is saying, or whether they simply represent that their partner is naming or, more generically, that their partner is acting in some way.

2.1.1 Representing another’s task

A clear demonstration that one’s partner’s task is automatically represented comes from experiments that reported joint interference effects (see Knoblich, et al., 2011 and Section 1.2.2 for reviews). For example, the spatial compatibility effect disappears (or is reduced) when an individual go/no-go version of the Simon task is used (Sebanz, et al., 2003, 2005). In the individual go-no version, one participant responds to only one colour with one hand, rather than to two colours with two hands). In the joint version of the go/no-go Simon task, instead, two participants take turns performing complementary halves of classic Simon task; i.e., each responding to one colour of the two colours with one hand. The joint go/no-go version, unlike the individual go/no-go version, is characterized by the presence of a joint interference effect.

What is the source of joint interference effects? As discussed in Section 1.2.2, there is substantial disagreement in the literature over this issue. According to the co-representation account, the partner’s task (set of SR mappings) is automatically represented
in a joint-task setting. The irrelevant features of the stimulus activate one’s partner’s response, which then interferes with the participant’s own response (Sebanz, et al., 2005). Specifically, this account predicts that interference occurs when the partner’s response and the participant’s own response are incongruent or incompatible.

According to the agent-conflict account (Wenke, et al., 2011), however, joint interference effects are due to agent conflict rather than to response conflict. In other words, it is the fact that one’s partner might (potentially) be responding on the current trial that matters, not the nature of their response. The account predicts that double-response trials (i.e., trials that evoke a response from both co-actors) should show longer latencies than single-response trials (i.e., trials that evoke a response only from one co-actor). This should occur irrespective of whether a congruent response or an incongruent response is evoked.

The literature on joint tasks has focused on tasks involving manual responses. However, a few studies have shown joint interference effects with verbal responses. Philipp and Prinz (2010) asked participants to respond to coloured shapes by uttering their own or their partner’s name, while their own or their partner’s face appeared as distractors. Responses were faster when participants saw their own face than when they saw their partner’s face, regardless of which name they used. This finding has been interpreted as supporting the agent conflict account (see Section 1.2.2).

In addition, Pickering and McLean (2013) had participants perform a joint version of the Stroop task, in which each participant responded to only one colour. Stroop interference was larger in this joint version than in the go/no-go version, where only one participant responded to only one colour, and the other did not respond (at least when the participants in the joint version of the task provided feedback to each other’s responses). This showed that participants represented their partner’s potential response and that this caused additional interference with their own response on incongruent trials.
2.1.2 Representing another’s utterances

The present study investigated whether joint task effects can occur when participants are asked to produce the names of pictured entities. Unlike previous studies, the response set was large and stimulus-response mappings were not arbitrary. Participants were cued to name the entities while their partner was cued not to respond or to name the entities either in the same (“congruent”) way or in a different (“incongruent”) way. This allowed us to test whether joint interference effects occur when people produce language alongside one another.

Note that our participants could not interact: They named pictures alongside each other, but could not hear each other. Therefore, participants might not represent their partner’s utterances at all. The studies that showed that non-interacting participants display joint interference effects (e.g., Atmaca, et al., 2011) all used manual responses. We do not know whether the same would hold for verbal responses, particularly because language is more tightly associated with communicative situations than manual actions. If other-representations are not formed, we expect no difference between conditions in which participants believe their partner is responding and conditions in which they believe their partner is not responding, because beliefs about one’s partner’s task should not matter at all. We term this the no-representation account. But if, on the contrary, other-representations are formed in our joint naming task, then the participants’ own responses should be affected by their beliefs about what their partner is doing.

Importantly, picture-naming responses are subject to varying degrees of congruency. For example, if one participant names a picture of an apple, her partner could concurrently produce the same word (i.e., apple), an unrelated word (e.g., blouse), or a semantically related word (e.g., banana). Importantly, different degrees of congruency do affect naming in solo tasks, as demonstrated by many picture-word interference studies (Damian & Bowers, 2003; Damian & Martin, 1999; Levelt, et al., 1999; Schriefers, et al., 1990; Starreveld & La Heij, 1999). Picture-word interference effects are explained by noting that
people automatically comprehend distractor words, which activate lexical representations that interact with the lexical representations accessed in the process of retrieving the picture’s name (Levelt, et al., 1999).

Interestingly, a recent study showed that the tendency to activate lexical representations corresponding to distractor words is not modulated by the belief that another speaker is naming these words, probably because such activation is too automatic (Sellaro, Treccani, & Cubelli, 2013, July; Experiment 1). Accordingly, when comprehension of distractor words was rendered more difficult (by presenting them in alternate upper and lower case), the semantic interference effect was eliminated if participants believed that another speaker was taking care of naming the distractor words (but not if they believed their partner was naming the ink colour of the pictures rather than the words). This suggests that beliefs about another speaker’s task might affect the allocation of attention in a picture-word interference task (Sellaro, et al., 2013, July; Experiments 2 and 3).

However, this study was not designed to investigate whether speakers form representations of other speakers’ utterances that are in a similar format to representations of their own utterances. The authors compared a condition in which participants named pictures individually to conditions in which they named pictures while they believed another person was concurrently speaking (either naming the word distractors or the ink colour of the pictures). Therefore, there was no condition in which participants believed a co-actor was present but silent. In addition, when the participants believed another person was speaking at the same time as them, they always believed that this person was producing utterances that differed from their own utterances.

In conclusion, Sellaro et al.’s (2013) findings show that speakers represent a co-actors’ task (i.e., whether another speaker is naming distractor words or not), and are affected by such representations. However, these findings cannot directly answer the question of whether speakers represent the content of other speakers’ utterances in a format that is similar to how they represent the content of their own utterances. Moreover, this study
indicates (Experiment 1) that in the standard picture-word interference paradigm, activation of distractor words occurs regardless of beliefs about a co-actor’s task, presumably because written words trigger lexical access in an automatic manner.

To circumvent the latter issue, in our study participants saw pairs of pictures rather than picture-word pairs. When distractor words are replaced by distractor pictures semantic interference effects often disappear (Damian & Bowers, 2003; Navarrete & Costa, 2005), and in at least one case were replaced by semantic facilitation effects (La Heij, Heikoop, Akerboom, & Bloem, 2003). Lack of interference effects suggests either that lexical retrieval does not normally occur for distractor pictures names (Bloem & La Heij, 2003) or that it is often too weak to out-weigh facilitatory effects at the conceptual level (Damian & Bowers, 2003; Navarrete & Costa, 2005).

But if participants are led to believe that their partner is naming the distractor picture at the same time as they are naming the target picture, and they represent their partner’s utterance in a way that is similar to how they represent their own utterance, they may carry out some of the same processes they would carry out if they were naming the distractor picture themselves. The time they take to respond may then be affected by this representation of their partner’s response. In particular, if the co-representation account holds for verbal responses, it predicts that participants represent the content of their partner’s response and activate the corresponding lexical representation.

This means that naming times should be affected by the belief that one’s partner is naming, and, in addition, that they should be differently affected depending on the content of the activated representations (i.e., whether they are congruent or incongruent). Specifically, the co-representation account predicts that activating a different set of production representations for one’s partner response (incongruent trials) should lead to interference compared to activating the same set of representations as for one’s own response (congruent trials). When the activated other-representations are different they can compete with self-
representations, whereas if other-representations and self-representations happen to coincide, the former would just strengthen activation of the latter.

In addition, if the co-representation account is correct and other-representations involve retrieval of the lexical item corresponding to one’s partner’s response, the semantic relationship between other- and self-representations might also matter. Specifically, participants should take longer to name a picture when they believe their partner is naming a different and semantically related picture than when they believe they are naming a different and unrelated picture.

We also included a condition in which the participant was cued to name a picture, while their partner was cued not to respond. If the agent-conflict account is correct, participants should take longer when their partner is cued to name a picture than when their partner is cued not to respond, regardless of whether they are cued to name that picture in a congruent or incongruent way. This is because participants would represent that their partner is responding, but not what she is saying.

We explored these issues in four experiments in which we varied the relationship between the participant’s and their partner’s response. Pairs of participants concurrently named pictures displayed on a computer screen. To rule out the possibility that any effect would be due to auditory feedback from the partner’s responses, participants performed the task in adjacent soundproof rooms. We asked whether picture naming would be slowed down by the belief that one’s partner is also naming pictures and whether it would be slowed down to varying degrees depending on the relationship between self- and other-representations. If one’s partner task is automatically represented, even in the absence of interaction, the mere belief that one’s partner is producing a naming response should interfere with the execution of one’s own naming response (Atmaca, et al., 2011; Knoblich, et al., 2011). To manipulate participants’ beliefs about their partner’s task, the partner’s instructions were displayed on the screen together with the participants’ instructions, and participants attended to the same stimuli as their partner.
2.2 Experiment 1

In Experiment 1, two differently coloured pictures were presented simultaneously to both participants. On each trial, before the pictures appeared, an instruction screen showed the names of the two participants accompanied by the words *red, blue, or no*. *Red* and *blue* corresponded to “go” trials: The participant was instructed to name the picture presented in the given colour first, and then name the other picture. *No* corresponded to “no-go” trials: The participant was instructed to give no response.

We manipulated the order in which the other participant (the partner) concurrently named the pictures (Partner’s task), as follows. On trials on which the two participants were assigned the same colour (*blue-blue* or *red-red*), they named the pictures in the same order, therefore producing the same verbal response (SAME condition). On trials on which the two participants were assigned different colours (*blue-red* or *red-blue*), they named the pictures in reverse order, therefore producing different verbal responses (DIFFERENT condition). Finally, when either of the participants was assigned a “no-go” trial (*red-no*, *blue-no*, *no-red*, *no-blue*), one participant named the pictures while their partner produced no response (NO condition). See Figure 2-1 for a summary of the Partner’s task manipulation employed in Experiment 1.
Figure 2-1. Partner’s task manipulation in Experiment 1 (sample trial from the unrelated condition; *apple* is displayed in blue, *blouse* is displayed in red).
In addition, participants saw either two semantically related (e.g., *apple – banana*) or two semantically unrelated pictures (e.g., *apple – blouse*). The semantic relatedness manipulation served two purposes. The first was to provide a manipulation check. Picture-word interference studies indicate that when two lexical items from the same semantic category are activated concurrently, the net result is usually interference (e.g., Schriefers, et al., 1990; see Section 2.1.2). In addition, Smith and Wheeldon (2004) showed that the time to initiate utterances containing two semantically related words is longer than the time to initiate comparable utterances containing unrelated words. Furthermore, Aristei, Zwitserlood, and Abdel Rahman (2012) showed a similar effect when German speakers were asked to utter the names of two pictures in close proximity (forming a novel noun-noun compound).

We therefore asked participants to produce both pictures’ names. The pictures were presented for a very short time (see Figure 2-1) to encourage participants to carry out a substantial amount of processing of both pictures in parallel (Wagner, Jescheniak, & Schriefers, 2010). The speed of picture presentation should make it difficult for participants to use colour information strategically (i.e., delaying processing the second picture until they have begun uttering the first picture’s name). Overall, we expected longer latencies when naming two related than two unrelated pictures (a main effect of semantic relatedness).

Most importantly, we expected the Partner’s task to affect naming latencies. Specifically, according to the co-representation account, naming latencies should be longer in the DIFFERENT condition than in the SAME condition. This is because it assumes that other-representations are content-specific; that is, they specify the lexical items that the partner is retrieving. Note that, because the speakers always named both pictures, their utterances always contained the same lexical items. However, when the order differed, the picture that the speaker had to name second was the picture that their partner had to name first. Therefore, in the DIFFERENT condition the representation of the partner’s response might enhance activation of the second picture’s name. This would in turn result in greater interference with selection of the first picture’s name, and lead to longer naming latencies.
Instead, when the order is the same, the first picture’s name was the word that one’s partner also had to retrieve first. Therefore, its activation level might be raised and competition with the second picture’s name might be reduced.

In addition, according to the co-representation account, the degree of relatedness might interact with Partner’s task in the following way: The semantic interference effect should be enhanced in the DIFFERENT condition (compared to the SAME condition) because in the DIFFERENT condition the second picture’s name receives additional activation from a representation of one’s partner response. The predictions of the co-representation account are presented in Figure 2-2 (panel A).

Alternatively, according to the agent-conflict account, speakers do not represent the content of their partner’s response, but they represent whether their partner is responding on the current trial or not. If this is the case, the relationship between self- and other-representations would not affect processing, and hence naming latencies would be equivalent in the SAME and DIFFERENT conditions. For the same reason, there should be no interaction between the relatedness manipulation and Partner’s task. However, naming latencies should be longer in the SAME and DIFFERENT conditions than in the NO condition. The predictions of the agent-conflict account are summarized in Figure 2-2 (panel B).

Finally, according to the no-representation account, another person’s utterances should not be represented at all under the conditions tested in our experiment (i.e., in the absence of interaction). The account therefore predicts that the Partner’s task manipulation will have no effect (i.e., there will be no difference between the SAME, DIFFERENT, and NO conditions). This scenario is presented in Figure 2-2 (panel C).
A  Co-representation

Mary(Mary)  Apple  Blouse
Mary(John)  Apple  Blouse

Predicted pattern of naming latencies

DIFFERENT  SAME  NO

B  Agent conflict

Mary(Mary)  Apple  Blouse
Mary(John)  Word

Predicted pattern of naming latencies

DIFFERENT  SAME  NO

C  No-representation

Mary(Mary)  Apple  Blouse
Mary(John)  Word

Predicted pattern of naming latencies

DIFFERENT  SAME  NO
Figure 2-2. Hypothesized effects of Partner’s task according to three different accounts: (A) co-representation account; (B) agent conflict account; (C) no-representation account; *apple* is in blue, *blouse* is in red. The left-hand side of the figure presents static depictions of the relevant nodes in Mary’s mental lexicon, with nodes making up Mary’s self-representations at the top and nodes making up Mary’s representation of John’s response (that is, *Mary(John)*) at the bottom. The right-hand side of the figure presents a snapshot of the activation level of the nodes just before the onset of the word “apple” when Mary is preparing to utter “apple blouse” (unrelated condition) and John is either preparing to utter “blouse apple” (DIFFERENT), “apple blouse” (SAME), or nothing (NO). The degree of activation is indicated by the thickness of the circles. Single-headed arrows indicate facilitation, while double-headed arrows indicate competition. In the related case, *banana* would replace *blouse* and the strength of competition between *banana* and *apple* would be greater than in the depicted unrelated case.
2.2.1 Method

Participants

Twelve pairs of previously unacquainted participants from the University of Edinburgh student community were paid £6 to participate. All participants reported being native speakers of English with no speaking or reading difficulties.

Materials

Fifty pictures were paired in two different ways to yield 50 picture-picture pairs (25 semantically related, 25 semantically unrelated; see Appendix A, Table A-1). For example, apple was paired once with banana (related) and once with blouse (unrelated). In turn, banana was paired once with apple (related) and once with frog (unrelated). Since one picture was embedded inside the other, half of the pictures were relatively small (about 250x200 pixels), whereas the others were relatively large (about 600x500 pixels). Of these 50 pairs, 11 were taken from Damian and Bowers (2003), 4 pairs from Navarrete and Costa (2005), and 10 were modelled after materials used by the same studies, but had to be modified to avoid phonologically related names (as the two studies where not conducted in English). Damian and Bowers (2003) and Navarrete and Costa (2005) controlled for visual similarity (i.e., related pairs were not visually more similar than the unrelated pairs). When adapting their materials, care was taken to pair pictures that were visually dissimilar to each other. Eight additional pictures were selected from Snodgrass and VanderWart (1980) to be used on practice trials.

Design

We manipulated three factors, all within participants: Partner’s task (henceforth, Partner; SAME vs. DIFFERENT vs. NO), Relatedness (unrelated vs. related), and Size (i.e., size of the first named picture: big vs. small). Partner and Relatedness were also manipulated
within items, whereas Size was manipulated between-items. An item was defined in terms of the first named picture (so apple-blouse and blouse-apple counted as different items). Partner refers to the task assigned to the participant’s partner: she named the pictures in the same order (SAME; e.g., participant: apple-blouse, partner: apple-blouse), in reverse order (DIFFERENT; e.g., participant: apple-blouse, partner: blouse-apple), or did not name any pictures (NO; e.g., participant: apple-blouse, partner: “”). Partner varied on a trial-by-trial basis.

Each picture was repeated 16 times across the experiment. In the SAME and DIFFERENT conditions, each picture was presented four times (twice in a related pair, twice in the unrelated pair). In the NO condition, each picture was presented 8 times (four times in a related and four times in an unrelated pair), in order to get the same amount of data as in the other two conditions. Each participant named each picture in first position 6 times, once per cell of the design.

There were 400 trials in total. These were presented in 4 blocks of 100 trials. Each block comprised an equal number of trials in each condition for both participants. Because the number of pictures could not be divided by four, and because of the requirement that participants named big and small pictures equally often in first position, it was not possible to ensure that each picture was named an equal number of times in each block. However, the order of presentation was pseudo-randomized, separately for each pair and for each block, with the constraint that the same picture never appeared on two consecutive trials. The order of blocks was also counterbalanced across pairs.

On every trial, one picture was red and the other was blue. Participants were cued to start either with the red or the blue picture. To prevent response strategies, we counterbalanced the following factors within each block: colour-participant pairing (whether a given participant named the red or the blue picture in first position), colour-size (whether the red picture was small and the blue picture was large or vice versa), and order of
instructions (whether the colour cue for a given participant was displayed in the top half or in the bottom half of the screen).

Procedure

Participants were tested in adjacent soundproof rooms. Each was seated at a distance of about 90 cm in front of a 48-cm 60 Hz LCD monitor; both monitors were connected to the same machine in the control room (so stimulus presentation was simultaneous). Stimulus presentation was controlled using E-Prime (Version 2.0). There was a window between the two rooms, but participants could only perceive each other peripherally when facing the monitors. The two rooms were linked via a communication system that allowed the experimenter to control whether the two participants could hear each other or not. Participants wore headphones through which they could hear their own voice and spoke into high-quality directional microphones (AKG Acoustics, Vienna, Austria, www.akg.com).

Upon entering the lab, the participants were introduced to one another and were taken to a randomly assigned room. They first learned the names of the pictures individually. The experimenter showed all the (practice and experimental) pictures on the computer screen once, one at a time, with the corresponding names, and asked participants to repeat them to aid memorization. Immediately afterwards, the pictures were shown again (without names), and participants were asked to produce the correct name. Correct feedback was provided in case either participant made a mistake or could not retrieve the name (both participants could hear the experimenter during this phase, but they could not hear each other). Participants were then informed that they would “work together” and were called out of the booths. Instructions were delivered to both participants at the same time in the control room. The instructions stressed that they should try to name both pictures as quickly as possible, while still preserving accuracy and clear pronunciation.

The participants then went back to their respective booths and performed 20 practice trials. These were similar to experimental trials but used only practice pictures, which were
matched to form semantically unrelated pairs. They were presented in a different random order for each pair. For each participant, on four of the practice trials the partner named the pictures in the same order; the partner named the pictures in reversed order on four other trials, and in six trials the partner named no picture. In the remaining six trials, the participant remained silent while the partner named the pictures. Finally, participants began the experimental phase.

On each trial, first a fixation cross was displayed for 1000 ms, then a display (2000 ms) that showed the participants names, each followed by an instruction word. After a 500-ms blank, two pictures (one red, one blue) were displayed simultaneously (for 400 ms). Each trial was concluded by a 1500 ms inter-stimulus interval. The 4 blocks were separated by breaks of variable length. The participants were left free to rest for as long as they required. The experimenter checked with both participants that they were happy to continue before resuming the experiment. An experimental session lasted about 1 hour.

**Recording and data analysis**

A 75-ms beep was used to mark stimulus presentation and was recorded together with the participants’ responses (on three separate channels, sampling rate: 48000 Hz) via an M-Audio FireWire 1814 device (inMusic, Cumberland, RI, www.m-audio.com) in Adobe Audition (Version 4.0). Beep onsets were automatically tagged using Audacity (Version 1.2.5). Recordings were then pre-processed to reduce background noise. Speech onsets were first automatically tagged using the Silence finder algorithm in Audacity and later checked manually for lip smacks and other non-speech noises. Naming latencies were defined as the time from beep onset to response onset.

We analysed the data using Generalized Linear mixed-effects models (Baayen, Davidson, & Bates, 2008; Bates, Maechler, & Dai, 2008) in R (Version 2.13.1). We started by fitting the complete model (including the main effects of Partner, Relatedness, and Size, the three two-way interactions, and the three-way interaction); we then removed predictors
that were not significant from the model, using a backward stepwise procedure, and stopped whenever removing a predictor caused a significant loss of fit (assessed using a log-likelihood ratio test). All predictors were contrast-coded. For Partner, we defined two planned contrasts: naming vs. no compared the DIFFERENT and SAME conditions against the NO condition; same vs. different compared the SAME against the DIFFERENT condition. For the accuracy data, we used a logistic link function (Jaeger, 2008).

Regarding random effects, we started with the full random effect structure, including random slopes (for all factors and their interaction) and random intercepts for both subjects and items. Since random slopes are only appropriate for within factors, we included by-subjects random slopes for Partner, Relatedness, and Size, and by-items random slopes for Partner and Relatedness. If the model with full random structure did not converge, we simplified it by removing higher order terms; we started with the three-way interaction by-subjects and proceeded with removing the two-way interactions, one at a time, first by subjects, then by-items. Once we found a model that converged, we used backward selection to select the slopes that contributed to model fit (with the alpha-level set to .1 to account for the conservativity of such tests).³

2.2.2 Results

We only report descriptive statistics and the results of likelihood ratio tests in this section. See Appendix B for full reports of the best fitting models.

Accuracy

Responses were coded as correct and entered into the onset time analysis (see below) only if both pictures were named correctly. Incorrect responses fell into 4 different conditions.

³ We also performed analyses that included random slopes for the factor of interest (Partner), for both items and participants, regardless of whether those slopes were found to contribute significantly to model fit according to the above-mentioned criteria. These analyses yielded exactly the same pattern of results as the ones reported below.
categories: naming errors (the wrong name was used); disfluency (the correct name was used, but the response contained hesitations or repetitions); order error (the second picture was named before the first picture); no response (the participant remained silent when he or she had to respond). Error counts and percentages are reported in Table 2-1.

Table 2-1.

<table>
<thead>
<tr>
<th></th>
<th>DIFFERENT</th>
<th>SAME</th>
<th>NO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unrelated</td>
<td>95 (7.9%)</td>
<td>81 (6.8%)</td>
<td>76 (6.3%)</td>
</tr>
<tr>
<td>Related</td>
<td>97 (8.1%)</td>
<td>64 (5.3%)</td>
<td>59 (4.9%)</td>
</tr>
</tbody>
</table>

Removing by-items random slopes for the factor Relatedness significantly harmed fit ($\chi^2 (4) =19.70, p<.001$). Speakers made (marginally) more incorrect responses when naming two unrelated than two related pictures ($\chi^2 (1) =3.54, p=.06$). Crucially, they produced more incorrect responses when their partner named than when he remained silent, and also fewer incorrect responses in the SAME than in the DIFFERENT condition. This was confirmed by a significant main effect of Partner ($\chi^2 (2) =13.10, p<.01$) and by $z$-values $> |2|$ for both planned contrasts (see Table B-1).

**Naming Latencies**

Naming latencies longer than 3000 ms or shorter than 300 ms were considered outliers and removed from all analyses. However, there were no such cases in Experiment 1. Then by-participant means and standard deviations were computed. Values that were more
than 3 standard deviations from the by-participant mean (1.5%) were replaced with the cut-off value.$^4$ Mean latencies are reported in Table 2-2.

Table 2-2.

*Mean voice onset times in ms (and standard deviations) by Partner and Relatedness in Experiment 1.*

<table>
<thead>
<tr>
<th></th>
<th>DIFFERENT</th>
<th>SAME</th>
<th>NO</th>
<th>Tot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unrelated</td>
<td>869 (220)</td>
<td>869 (225)</td>
<td>855 (217)</td>
<td>864</td>
</tr>
<tr>
<td>Related</td>
<td>881 (221)</td>
<td>886 (230)</td>
<td>872 (229)</td>
<td>880</td>
</tr>
<tr>
<td>Tot</td>
<td>875</td>
<td>877</td>
<td>864</td>
<td></td>
</tr>
</tbody>
</table>

No slopes contributed to fit, so only random intercepts were retained. The main effects of Partner ($\chi^2 (2) =7.80, p<.05$) and Relatedness ($\chi^2 (1) =11.32, p<.001$) contributed to fit. The three-way interaction between Partner, Relatedness, and Size was only marginally significant ($\chi^2 (2) =5.15, p=.08$) and is not included in Table B-2. All two-way interactions, including the interaction of Partner and Relatedness were not significant. Therefore, the prediction of the co-representation account that the relatedness effect would be larger in the DIFFERENT than in the SAME condition was not confirmed. Participants took longer to start speaking when their partner was also preparing to speak (naming: 876 ms) than when their partner remained silent (NO: 864 ms). However, average naming latencies were very similar in the DIFFERENT (875 ms) and SAME (877 ms) conditions. Therefore, the main prediction of the co-representation account (i.e., that latencies would be longer in the DIFFERENT than in the SAME condition) was also not confirmed. In addition, this finding is inconsistent with the no-representation account, but fully consistent with the agent-conflict account. Finally, in all conditions, participants took longer when the two pictures were

$^4$ Additional analyses performed on the complete data set (with only very extreme outliers removed) yielded a similar pattern of results.
semantically related than when they were unrelated (semantic interference effect: 12 ms in DIFFERENT, 17 ms in SAME, 17 ms in NO).

2.2.3 Discussion

Experiment 1 showed that participants took longer to initiate a naming response when they believed their partner was also preparing a naming response. This finding rules out the no-representation account (Figure 2-2, Panel C).

The results of Experiment 1 do not fully support the co-representation account (Figure 2-2, Panel A) for two reasons. First, while participants made more errors when their partner was preparing an incongruent (DIFFERENT) than a congruent (SAME) response, this pattern was not confirmed by the latency data. Participants were no slower when they believed their partner was preparing an incongruent response than when they believed they were preparing a congruent response. Second, while there was a semantic interference effect (longer latencies for related than unrelated responses), which replicated previous findings (Aristei, et al., 2012), this effect was no greater in the DIFFERENT (12 ms) than in the SAME condition (17 ms). This further suggests that the content of the partner’s response was not represented.

Instead, the results are consistent with the agent-conflict account (Figure 2-2, Panel B), which claims that people represent whether a response occurs or not, but they do not represent the content of the response itself. Consistently, latencies were longer when both participants named pictures (i.e., naming conditions) than when only one did (i.e., NO condition), regardless of the relationship between the participant and their partner’s response. However, we must consider alternative explanations of this finding.

The slowest conditions (SAME and DIFFERENT) are the ones in which two “go” instructions are displayed on the screen. It might be that, despite being addressed by their first name (a highly salient stimulus; e.g., Wood & Cowan, 1995), participants were
distracted by the presence of their partner’s instruction more when it was a “go” instruction than when it was a “no-go” instruction, for a number of reasons.

First, “go” instructions are words of the same type (colour names), whereas “no-go” instructions used a clearly different word (“no”). Therefore, “go” instructions are more similar to each other than they are to “no-go” instructions. Similarity might cause interference between memory representations for one’s own and the partner’s instructions. Note that participants rarely performed their partner’s task by mistake, which seems to suggest that they rarely misremembered the instructions. There is however some indication that this occurred more often in the DIFFERENT condition (on 2.3% of trials speakers named the pictures in their partner’s order), compared to the NO condition (on 1.2% of trials speakers gave no response). But more importantly, this explanation cannot account for why latencies were no longer in the DIFFERENT than in the SAME condition (where instructions were identical, so interference between memory representations is unlikely to have occurred). We return to this issue in the Discussion to Experiment 2, where we replaced the “no-go” instructions with (a different kind of) “go” instructions. Here we note that interference between memory representations for the instructions cannot account for the finding that naming latencies were slower in both naming conditions (DIFFERENT and SAME) than in the NO condition.

Therefore, we conclude that participants represented whether it was their partner’s turn to respond on any given trial and that they experienced interference whenever both they and their partner were preparing a response concurrently. This is consistent with the agent conflict account. But what sort of mechanism could be responsible for this interference

---

5 Of course, it was not possible to make a mistake of this type in the SAME condition, as both participants were given the same instruction.
6 In addition, in another study (Van de Cavey, Gambi, MacKenzie, Nelissen, & Pickering, 2012, September) we conducted a related experiment where participants described simple scenes using active or passive sentences. In that study, we replaced the word “no” with “grey” (keeping instructions the same; i.e., to remain silent) and found that latencies were still longer when participants believed their partner was speaking than when they believed their partner was silent.
effect? Participants might represent that their partner was “doing something” at the same time that they prepared their response. If this version of the agent conflict account is correct, we expect that a belief that one’s partner is performing any task would slow down the process of naming to the same extent as a belief that they are naming pictures.

Alternatively, the interference effect could arise because the same mechanisms (i.e., language production mechanisms) are used to represent one’s partner naming response and to prepare one’s own naming response. If this alternative version of the agent conflict account is correct, we predict less interference when one’s partner is performing a different (non-naming) task than when one’s partner is preparing a naming response. Experiment 2 was designed to decide between these alternatives.

2.3 Experiment 2

In Experiment 2 we replaced the NO condition with a semantic categorization (hence, CAT) condition. The SAME and DIFFERENT conditions were the same as in Experiment 1. In the CAT condition, participants named the pictures while their partner judged whether the two pictures were from the same semantic category (as discussed below). For example, goat and pig were considered as belonging to the same semantic category (animal), while goat and cup were not. Responses to the categorization task were given by speaking “yes” or “no” into the microphone.

Therefore, all trials required an overt verbal response from both participants. Consequently, if imagining one’s partner performing any task was driving the effect we observed in Experiment 1, we should find no difference between the SAME, DIFFERENT, and CAT conditions. We chose the CAT task because it would be maximally similar to the naming task and therefore more likely to share the same pool of attentional resources. First, both tasks involve visual processing of the pictures and retrieving the concepts associated with the depicted entities. In addition, both tasks involve the articulation of an overt verbal
response. Crucially, however, the two tasks differ in the extent to which they engage language production mechanisms (i.e., lexical retrieval).

In the naming task, speakers clearly need to access the lexical items corresponding to the concepts depicted in the pictures. In the categorization task, however, speakers need not access the lexical items, as the task can be performed on the basis of conceptual semantic information alone. That indeed speakers do not access the names of pictures when they do not perform a naming task is confirmed by an experiment reported in Jescheniak and Levelt (1994). In Experiment 2 in their study participants had to decide whether a picture matched a previously presented word or not. Despite the fact that participants were exposed to the picture names on matching trials, there was no evidence that they (re-)accessed the picture names at the moment of performing the task, as shown by the lack of a frequency effect.

More generally, it is common practice in psycholinguistics to contrast picture categorization with picture naming in order to disentangle the contribution of perceptual and conceptual processing from lexical retrieval processes proper (for a similar logic, see for example Almeida, Knobel, Finkbeiner, & Caramazza, 2007 and references therein). Therefore, we assume that the naming task engages language production mechanisms (i.e., lexical retrieval) to a greater extent than the categorization task. If the interference effect in Experiment 1 was specific to representing that one’s partner is preparing to name (i.e., is engaging in lexical retrieval) it should be replicated in Experiment 2.

Finally, we retained both the SAME and the DIFFERENT conditions from Experiment 1 to provide another test of their comparative effects. The accuracy data in Experiment 1 seemed to suggest that the DIFFERENT condition might induce more interference than the SAME condition. Therefore, we wanted to check whether this effect would be replicated in Experiment 2. The semantic relatedness manipulation was also retained in this experiment. Apart from ensuring that positive and negative responses were balanced in the semantic categorization task, it also provided a further test of the co-
representation account’s prediction that semantic interference effects should be larger in the DIFFERENT than in the SAME condition.

2.3.1 Method

Participants

Sixteen further pairs of participants from the same community as the pairs in Experiment 1 were paid to participate.

Materials and Design

These were the same as in Experiment 1, except that the NO condition was replaced with the CAT condition. In order for participants to practice the semantic categorization task, we replaced two of the original practice pictures with two new pictures from Snodgrass and VanderWart (1980), so that it would be possible to form semantically related pairs (this also involved re-pairing the original pictures). Various semantic categories were represented in the materials from Experiment 1 (see Appendix A).

Procedure

For the semantic categorization task, participants were instructed to respond to the word question (which replaced the word no) by answering the following question: “Are the two pictures from the same category?” They responded by saying “yes” into the microphone if they thought the answer was positive, or “no” if they thought the answer was negative. The experimenter provided two examples to clarify what it meant for two pictures to be “from the same category” (one example mentioned dog and snake as requiring a positive answer, dog and lemon as requiring a negative answer; the second example mentioned pen and ruler, as requiring a positive answer, pen and door as requiring a negative answer). The experimenter
also mentioned the relevant superordinate category (i.e., animal; stationery) while illustrating the examples. Otherwise, the instructions and procedure were the same as in Experiment 1.

**Recording and data analyses**

These were exactly the same as in Experiment 1 with regard to naming responses. Responses to the semantic categorization task were also analysed; latencies for both positive and negative answers were extracted automatically using the Silence Finder feature in Audacity (without manual correction).

**2.3.2 Results**

**2.3.2.1 Semantic categorization task**

Four types of responses were coded as incorrect: disfluencies (hesitations, repetitions), wrong responses, missing responses, and task errors (i.e., when participants performed the naming task instead of the categorization task). Overall, participants were highly accurate: they responded correctly on 94.7% of the unrelated trials and on 93.6% of the related trials. There was no significant difference in the number of errors between related and unrelated trials. Task errors amounted to 2.3% of the trials (N=77). We also looked at the latency to respond on correct trials. Based on the overall distribution of responses, we removed the 4 responses that were shorter than 250 ms or longer than 2500 ms. Again, there was no difference between related (“yes”) trials (M = 936 ms, SD = 289 ms) and unrelated (“no”) trials (M = 944 ms, SD = 291 ms).
2.3.2.2 Picture naming task

Accuracy

Incorrect responses were coded as in Experiment 1, except that another type of error was possible; occasionally (on 2 trials in DIFFERENT, 4 in CAT and 5 in SAME) participants performed the categorization task instead of the naming task. Counts and percentages are reported in Table 2-3. Removing by-items random slopes for the factor Relatedness significantly harmed fit ($\chi^2(4) = 18.38, p<.005$). Participants produced marginally more incorrect responses on related than on unrelated trials ($\chi^2(1) = 2.98, p=.08$). No other factor gave a significant contribution to model fit (Table B-3).

Table 2-3.

Error counts (percentages) by Partner and Relatedness in Experiment 2.

<table>
<thead>
<tr>
<th></th>
<th>DIFFERENT</th>
<th>SAME</th>
<th>CAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unrelated</td>
<td>89 (5.6%)</td>
<td>100 (6.3%)</td>
<td>96 (6.0%)</td>
</tr>
<tr>
<td>Related</td>
<td>115 (7.2%)</td>
<td>114 (7.1%)</td>
<td>93 (5.8%)</td>
</tr>
</tbody>
</table>

Naming Latencies

Naming latencies longer than 3000 or shorter than 300 ms were considered outliers and removed from all analyses. There were only 2 such cases in Experiment 2. Then by-participant means and standard deviations were computed. Values than were more than 3 standard deviations from the by-participant mean (1.7%) were replaced with the cut-off value.\(^7\)

\(^7\) Additional analyses performed on the complete data set (with the two outliers removed) yielded a similar pattern of results.
Table 2-4.

*Mean voice onset times in ms (and standard deviations) by Partner and Relatedness in Experiment 2.*

<table>
<thead>
<tr>
<th></th>
<th>DIFFERENT</th>
<th>SAME</th>
<th>CAT</th>
<th>Tot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unrelated</td>
<td>881 (257)</td>
<td>879 (257)</td>
<td>874 (255)</td>
<td>878</td>
</tr>
<tr>
<td>Related</td>
<td>898 (257)</td>
<td>907 (277)</td>
<td>885 (259)</td>
<td>897</td>
</tr>
<tr>
<td>Tot</td>
<td>889</td>
<td>893</td>
<td>880</td>
<td></td>
</tr>
</tbody>
</table>

By-participants random slopes for Size ($\chi^2(5) = 46.16, p < .001$) and by-items random slopes for Relatedness ($\chi^2(4) = 8.21, p < .10$) were included. Relatedness contributed significantly to model fit ($\chi^2(1) = 11.04, p < .001$). Crucially, Partner also contributed significantly to model fit ($\chi^2(2) = 6.54, p < .05$; see Table B-4). Participants took longer to start naming when their partner was also preparing to name (891 ms) than when their partner was preparing to categorize the pictures (880 ms). However, average naming latencies were no longer in the DIFFERENT (889 ms) than in the SAME (893 ms) condition. Finally, in all conditions, participants took longer when the two pictures were semantically related than when they were unrelated (semantic interference effect: 17 ms in DIFFERENT, 28 ms in SAME, 11 ms in CAT; see Table 2-4).

### 2.3.3 Discussion

The results of Experiment 2 are again not consistent with the no-representation account, as Partner’s task affected naming latencies. In addition, they are not consistent with the co-representation account. As in Experiment 1, naming latencies did not differ in the DIFFERENT and SAME condition. In addition, and unlike in Experiment 1, the likelihood of producing an incorrect response also did not differ in the two conditions (and did not differ significantly from the CAT condition, either). Therefore, the co-representation
account’s prediction that participants experience more interference when they believe their partner is naming the pictures in reverse order than when they believe they are naming in the same order was not supported by the results of Experiment 2. Furthermore, as in Experiment 1, we found that speakers were slower at naming pairs of semantically related than unrelated pictures. However, and once again as in Experiment 1, the semantic interference effect was not larger in the DIFFERENT than in the SAME condition. This further suggests that participants did not represent the content of their partner’s response, contrary to the predictions of the co-representation account.

Most importantly, naming latencies were longer when speakers believed that their partner was also naming a picture than when they believed that their partner was performing a semantic categorization task. Given that the two tasks share all processing stages except lexical retrieval, we conclude that the process of naming pictures is inhibited by the belief that another speaker is concurrently retrieving the pictures’ names. This is not consistent with a version of the agent-conflict account in which speakers only represent whether it their partner’s turn to respond on the current trial. Rather, the results of Experiment 2 suggest that speakers specifically represent whether another speaker is planning to name, and not just whether their partner is about to respond (in any way). We return to this issue in the General Discussion.

In addition, we note that naming latencies were longer in the Partner-naming than in the CAT conditions, even though the CAT condition was also associated with “go” instructions for the participants’ partner. This weakens the concern raised in the Discussion to Experiment 1 that interference on Partner-naming trials in that experiment was due to the presence of similar instructions (both “go” instructions) and not to a representation of the partner’s response.

However, it is possible that the finding that participants represent whether their partner is about to name a picture or not is specific to the situation in which participants have to encode both pictures’ names, while simultaneously formulating an utterance plan that
specifies order of mention. These requirements might make the task rather demanding and perhaps more sensitive to interference from a representation of the other person’s response. To investigate whether similar effects would occur when speakers were naming single words, we conducted Experiment 3.

Another aim of Experiment 3 was to test whether representations of others’ responses are formed when those responses bear no relationship to one’s own responses. In Experiment 1 and 2 participants might have formed representations of their partner’s responses because those responses always overlapped in content with their own responses, and were therefore perceived as somehow associated with their own responses. In Experiment 3, we tested a condition in which the partner’s response was completely unrelated to the participant’s concurrent response (except for the fact that the corresponding visual stimulus was co-located with the stimulus the participants responded to).

Finally, in Experiment 3 we provided yet another test of the co-representation account. Proponents of this account might note that in Experiments 1 and 2 the content of the partner’s response (i.e., the identity of the retrieved lexical items) was in fact identical in the SAME and DIFFERENT conditions, as the only difference was in the order of mention. It is conceivable that people do indeed represent the content of their partners’ responses, but not order. Therefore in Experiment 3 we changed the instructions so that the partner named either the same picture or a different picture than the participant.

2.4 Experiment 3

Experiment 3 was similar to Experiment 1, except that participants named one picture rather than both pictures. Participants named the picture in the assigned colour, and ignored the other picture. Therefore, we included a condition in which participants named a picture while their partner remained silent (NO), a condition in which participants named the same picture (SAME), and a condition in which participants named different pictures
(DIFFERENT). Of course, participants therefore believed their partner was naming one picture or no picture.

We reasoned that the task would be much less demanding than the task in Experiment 1. Speakers had to retrieve and produce only one word. Therefore, we expected them to respond at shorter latencies. Of course, this task also requires speakers to successfully ignore the non-target picture. There is evidence that distractor picture names are accessed during target picture naming (e.g., Meyer & Damian, 2007; Morsella & Miozzo, 2002; Navarrete & Costa, 2005). However, distractor picture names should be less activated than in the previous experiments at least. This is because in Experiment 1 and 2 participants retrieved and articulated both pictures’ names, while in the current experiment they were explicitly instructed to ignore the distractor picture. If the reduced demands of this task make it somewhat more impenetrable to interference, then the effect of Partner’s task might be reduced in Experiment 3 compared to Experiment 1.

In addition, participants in the DIFFERENT condition of Experiment 3 produced a response that did not overlap in lexical content with their partner’s response. Therefore, this made it easier for participants to ignore their partner’s response in this condition. It is possible that this will further reduce the size of the interference effect (in the DIFFERENT condition only). Nevertheless, the literature on shared representations of manual responses shows that one’s partner’s responses are represented even when they are irrelevant to one’s own task (see Knoblich, et al., 2011). Therefore, it is also possible that participants will form representations of their partner’s response even when they are not related to their own responses, as in the DIFFERENT condition of Experiment 3.

Finally, Experiment 3 provides an additional test of the co-representation account. Arguably, the lack of evidence in favour of the co-representation account in Experiments 1 and 2 might be a consequence of the task used. Because both picture names had to be uttered, activation levels for both might have been very high. In turn, this could have made it impossible to detect any additional activation due to representing the content of one’s
partner’s response. As a result, we would not have observed an interaction between Partner’s Task and Relatedness, even if representations of the partner’s response were in fact content-specific.

This account would have still trouble explaining why participants did not experience greater interference in the DIFFERENT than in the SAME condition. However, proponents of the co-representation account could claim that this occurred because representations of the partner’s responses specify the content of those responses (i.e., the lexical items) but not the order in which the responses are produced (rendering the SAME and DIFFERENT conditions in fact equivalent in terms of what other-representations are formed).

In contrast, in Experiment 3, the SAME and DIFFERENT conditions involved distinct target pictures and, as discussed above, we expected distractor picture names to be less activated than in previous experiments. Furthermore, we did not expect any semantic interference effect in the NO condition (Damian & Bowers, 2003; Navarrete & Costa, 2005). Therefore, any additional activation received by the distractor picture name in the DIFFERENT condition (due to representing one’s partner response) should be more easily detected. Specifically, if the activation of the distractor picture’s lemma is strong enough to out-weigh conceptual facilitation from the associated concept, then we might observe semantic interference in the DIFFERENT condition. No such effect should occur in the SAME condition, which would be equivalent to the NO condition in terms of activation of the distractor’s picture name (and perhaps might involve even stronger activation of the target picture’s name, compared to the NO condition). In contrast, the agent-conflict account makes the prediction (the same as for Experiment 1 and 2) that naming will be similarly inhibited in the SAME as in the DIFFERENT condition, and that semantic relatedness will not interact with Partner task.
2.4.1 Method

Participants

Thirteen further pairs of participants from the same community as participants in Experiment 1 and 2 were paid to participate. One pair was removed from the analyses because one of the participants produced exceptionally long naming latencies.

Materials, Design, Procedure, Recording, and Data Analysis

These were the same as in Experiment 1, except that participants were instructed to interpret the colour cue as indicating that they had to name the picture presented in that colour (and ignore the other picture).

2.4.2 Results

Accuracy

Recordings for 21 (0.2%) trials could not be analysed due to experimental error or technical problems. For the remaining trials, we coded whether the response was correct or incorrect. Incorrect trials were trials on which the wrong name for the picture was used, the correct name was used but produced disfluently, the participants performed the wrong task (either did not name a picture when they had to, or named the wrong picture; there were 10 such cases in DIFFERENT, 11 in SAME, and 10 in NO).
Table 2-5.

*Error counts (percentages, out of the total number of scorable trials) by Partner and Relatedness in Experiment 3.*

<table>
<thead>
<tr>
<th>Relatedness</th>
<th>DIFFERENT</th>
<th>SAME</th>
<th>NO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unrelated</td>
<td>45 (3.8%)</td>
<td>38 (3.2%)</td>
<td>31 (2.6%)</td>
</tr>
<tr>
<td>Related</td>
<td>31 (2.6%)</td>
<td>46 (3.8%)</td>
<td>37 (3.1%)</td>
</tr>
</tbody>
</table>

As can be seen from Table 2-5, the likelihood of producing an incorrect response was not affected much either by Partner or by Relatedness. Participants tended to be less accurate when the two pictures were semantically related than when they were unrelated, but only in the SAME and NO conditions. The trend appeared to be reversed in the DIFFERENT condition. No random slopes contributed to model fit (all p’s > .1). The interaction of Relatedness and Partner marginally contributed to fit ($\chi^2$ (2) = 4.73, p = .09; *naming vs. NO*: $B = -.30$, SE = .31, $z =$ -.97; *SAME vs. DIFFERENT*: $B = .45$, SE = .24, $z =$ 1.86). No other factor or interaction was significant.

**Naming Latencies**

Naming latencies longer than 3000 or shorter than 300 ms were considered outliers and removed from all analyses. There were 6 such cases in Experiment 3. Values that were more than 3 standard deviations from the by-participant mean (1.6%) were replaced with the cut-off value.$^8$ Mean latencies are reported in Table 2-6. By-participant random slopes for Size ($\chi^2$ (5) = 10.17, p = .07) and by-item slopes for Relatedness ($\chi^2$ (4) = 22.10, p < .001) were included. Among fixed factors, only the interaction of Relatedness and Size significantly contributed to fit ($\chi^2$ (1) = 4.74, p < .05). The factor Partner was not significant ($\chi^2$ (2) = 4.33, p = .11). See Table B-5.

$^8$ Additional analyses performed on the complete data set (with the six outliers removed) yielded a similar pattern of results.
Table 2-6.

*Mean voice onset times in ms (and standard deviations) by Partner, Relatedness, and Size in Experiment 3.*

<table>
<thead>
<tr>
<th></th>
<th>DIFFERENT</th>
<th>SAME</th>
<th>NO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Unrelated</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Big</td>
<td>694 (174)</td>
<td>698 (171)</td>
<td>691 (168)</td>
</tr>
<tr>
<td>Small</td>
<td>709 (148)</td>
<td>710 (176)</td>
<td>697 (152)</td>
</tr>
<tr>
<td>Tot unrelated</td>
<td>701</td>
<td>704</td>
<td>694</td>
</tr>
<tr>
<td><strong>Related</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Big</td>
<td>688 (175)</td>
<td>692 (172)</td>
<td>681 (166)</td>
</tr>
<tr>
<td>Small</td>
<td>716 (164)</td>
<td>722 (168)</td>
<td>719 (169)</td>
</tr>
<tr>
<td>Tot related</td>
<td>702</td>
<td>707</td>
<td>700</td>
</tr>
</tbody>
</table>

In order to resolve the interaction, we used treatment coding to fix one of the two levels of the factor Size to zero and refitted the model shown in Table B-5 to derive estimates for Relatedness separately for small and big pictures. While there was a non-significant tendency towards semantic facilitation for big pictures ($B = -8$, $SE = 7$, $t = -1.14$), small pictures showed a tendency towards semantic interference, which approached significance ($B = 14$, $SE = 7$, $t = 1.96$).

### 2.4.3 Discussion

Unlike in Experiments 1 and 2, the effect of Partner was not significant in Experiment 3, although we found a numerical tendency in the same direction as in the previous two experiments. Given that the current experiment was comparable to Experiment
1 for number of items and participants, it is possible that the magnitude of the effect is reduced under simple naming conditions. This could be because the task is less demanding (and thus less susceptible to interference). We explore this possibility further in Experiment 4.

Importantly, no interaction involving Partner approached significance in the latency analysis, thus suggesting that representing one’s partner response did not modulate the effect of the Relatedness manipulation. There was a marginal interaction between Partner and Relatedness in the error analysis, but it was in the opposite direction to the one predicted by the co-representation account (i.e., the semantic interference effect was larger in the SAME than in the DIFFERENT condition). Therefore, the results of Experiment 3 do not support the co-representation account.

Another aspect of the findings deserves attention. We found a marginal semantic interference effect when speakers where naming small pictures. Interestingly, we note that Damian and Bowers (2003), who reported no effect of semantic relatedness in a similar task, only asked their participants to name the pictures that were large in size. Unfortunately, it is not possible to determine the source of this interference effect. Given that our materials were not fully controlled for visual similarity, we cannot rule out that the relatedness effect was in fact caused by visual similarity (that could be greater between related than unrelated picture pairs). Perhaps the relatedness effect was limited to small target pictures because big distractor pictures were harder not to attend to than small distractor pictures.

However, the presence of a relatedness effect, even if only marginal, might also indicate that speakers were accessing the distractor’s picture name, at least when the target picture was small in size, contrary to expectations. Therefore, proponents of the co-representation account might take this as evidence that Experiment 3 suffers from some of the same limitations as Experiments 1 and 2 (where participants had to name both picture names). We addressed this concern in Experiment 4.
2.5 Experiment 4

Experiment 4 was identical to Experiment 3, except that all stimuli were visually degraded. Degradation is known to cause quite large increases in picture naming latencies (e.g., Mädebach, Jescheniak, Oppermann, & Schriefers, 2011). If processing the target picture is made more demanding, speakers will be less likely to encode the distractor’s picture name. Therefore, baseline activation of the distractor picture’s name (i.e., in the NO condition) should be reduced compared to Experiment 3. Hence the distractor picture name should interfere less with retrieval of the target picture name.

Mädebach et al. (2011) demonstrated that this is indeed the case. They found that phonological facilitation effects from distractor pictures’ names are eliminated when the target picture, the distractor picture, or both are visually degraded. They argued that processing constraints limit the amount of resources that can be devoted to encoding the name of distractor pictures.

Experiment 4, therefore, provides an especially favourable test of the co-representation account. Since degradation should reduce activation of the distractor picture’s name, it should be easier to detect an increase in activation as a result of representing the content of the partner’s response. Hence, participants should be slower in the DIFFERENT than in the SAME condition, and there should be a semantic interference effect in the DIFFERENT condition, but not in the NO or the SAME condition. The agent-conflict account, instead, would still predict slower latencies in the DIFFERENT and SAME conditions than in the NO condition, as in Experiments 1, 2, and 3.
2.5.1 Method

Participants

Twelve further pairs of participants from the same community as participants in Experiment 1, 2, and 3 were paid to participate.

Materials

These were the same as in Experiments 1 to 3, except that a mask of parallel white lines (see Mädebach, et al., 2011) was superimposed on the pictures in order to conceal part of the lines. The proportion of masked lines varied from picture to picture; we tried to keep the proportion of masked lines constant for each picture across the related and unrelated condition, but this was not always possible because the mask superimposed on small pictures partly overlapped with the contours of the big picture. However, as explained above, we were not expecting participants to retrieve the distractor’s name in this experiment and, therefore our interest was not focused on the semantic relatedness manipulation.

Design, Procedure, Recording and Data Analysis

These were the same as in Experiment 3.

2.5.2 Results

Accuracy

Recordings for 2 trials could not be analysed because of experimental error. The remaining trials were coded as correct or incorrect as in Experiment 3. The participants performed the wrong task (either did not name a picture when they had to, or named the wrong picture) on 31 trials in the DIFFERENT condition, 24 in the SAME condition, and 33 in the NO condition. Counts (and percentages) of incorrect trials are given in Table 2-7,
broken down by Partner, Relatedness, and Size of the named picture. By-participant random slopes for Size contributed to model fit ($\chi^2 (5) = 35.97, p<.001$). The interaction of Partner and Size contributed to model fit significantly ($\chi^2 (2) = 10.57, p<.01$). When naming big pictures, speakers made 58 errors in DIFFERENT, 39 in SAME, and 34 in NO. When naming small pictures, they made 48 errors in DIFFERENT, 58 in SMALL and 68 in NO. The best fitting model is reported in Table B-6.

Table 2-7.

*Error counts (percentages, out of the total number of scorable trials) by Partner, Relatedness, and Size in Experiment 4.*

<table>
<thead>
<tr>
<th></th>
<th>DIFFERENT</th>
<th>SAME</th>
<th>NO</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Unrelated</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Big</td>
<td>33 (5.5%)</td>
<td>21 (3.5%)</td>
<td>15 (2.5%)</td>
</tr>
<tr>
<td>Small</td>
<td>24 (4.0%)</td>
<td>28 (4.7%)</td>
<td>32 (5.3%)</td>
</tr>
<tr>
<td>Tot unrelated</td>
<td>57</td>
<td>49</td>
<td>47</td>
</tr>
<tr>
<td><strong>Related</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Big</td>
<td>25 (4.2%)</td>
<td>18 (3.0%)</td>
<td>19 (3.2%)</td>
</tr>
<tr>
<td>Small</td>
<td>24 (4.0%)</td>
<td>30 (5.0%)</td>
<td>34 (5.7%)</td>
</tr>
<tr>
<td>Tot related</td>
<td>49</td>
<td>48</td>
<td>53</td>
</tr>
<tr>
<td><strong>Tot</strong></td>
<td>106</td>
<td>97</td>
<td>100</td>
</tr>
</tbody>
</table>

To resolve the interaction, we fixed one level of the factor Size to 0 using treatment coding. In addition, we tested for differences between DIFFERENT and NO and differences between SAME and NO separately (i.e., we used treatment coding for the factor Partner, and took NO as the reference level). We did this, instead of using planned contrast coding as in
previous analyses, because the interaction was not predicted. When participants named big pictures, they made significantly more errors in the DIFFERENT than in the NO condition (log-odds $B = .59$, SE=.23, $z=2.60$), but they made similar amounts of errors in the SAME compared to the NO condition (log-odds $B = .15$, SE=.25, $z=.61$). When speakers where naming small pictures, instead, they made marginally fewer errors in the DIFFERENT than in the NO condition (log-odds $B = -.37$, SE=.21, $z=-1.82$); again, they made comparable amounts of errors in the SAME as in the NO condition (log-odds $B = -.15$, SE=.20, $z=-.79$).

So, it appears that in both cases participants’ behaviour differed when they believed that their partner was naming a different picture compared to when they believed their partner was not naming.

**Naming Latencies**

Naming latencies longer than 3000 or shorter than 300 ms were considered outliers and removed from all analyses. There were 4 such cases in Experiment 4. Values that were more than 3 standard deviations from the by-participant mean (1.5%) were replaced with the cut-off value. As shown in Table B-7, by-participant random slopes for Size contributed to model fit ($\chi^2 (5) =12.91$, $p<.05$). Only the main effect of Partner was significant ($\chi^2 (2) =11.20$, $p<.01$). As shown in Table 2-8, latencies were longer when participants believed their partner was responding (naming: 817 ms) than when they did not (NO: 801 ms). However, latencies were no longer in the DIFFERENT (818 ms) than in the SAME condition (817 ms).

---

9 Additional analyses performed on the complete data set (with only the four outliers removed) yielded a similar pattern of results.
Mean voice onset times in ms (and standard deviations) by Partner and Relatedness in Experiment 4.

<table>
<thead>
<tr>
<th></th>
<th>DIFFERENT</th>
<th>SAME</th>
<th>NO</th>
<th>Tot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unrelated</td>
<td>817 (208)</td>
<td>813 (215)</td>
<td>798 (198)</td>
<td>809</td>
</tr>
<tr>
<td>Related</td>
<td>818 (222)</td>
<td>820 (214)</td>
<td>805 (206)</td>
<td>815</td>
</tr>
<tr>
<td>Tot</td>
<td>818</td>
<td>817</td>
<td>801</td>
<td></td>
</tr>
</tbody>
</table>

Of course, this finding contrasts with the results of Experiment 3, where the main effect of Partner was not significant. However, in Experiment 3 we found a non-significant trend in the same direction. To directly compare the two experiments, we ran a combined analysis of Experiment 3 and 4. We found a significant main effect of Partner ($\chi^2 (2) =12.54$, $p<.005$) and no interaction between Partner and Experiment ($\chi^2 (2) =2.15$, $p=.34$). There was (as expected) a main effect of Experiment ($\chi^2 (1) =118.63$, $p<.001$), with latencies being longer when pictures were degraded than when they were not (see Table B-8).

2.5.3 Discussion

In Experiment 4, we replicated the finding that naming is inhibited by the belief that one’s partner is naming using a task in which participants name single pictures that have been visually degraded. This contrasts with the failure to replicate the effect in Experiment 3, when participants were naming single pictures that were visually intact. However, in Experiment 3 we found a non-significant trend in the same direction, and we did not find a significant Experiment-by-Partner interaction in a combined analysis of the two experiments.

This could indicate that a similar phenomenon is at play in both experiments. However, it is also possible that naming responses to non-degraded pictures are affected by beliefs about the partner’s task to a lesser degree. This, in turn, could be because the task of
naming single, non-degraded pictures is less demanding in terms of resources (as proved by shorter response times), and thus less penetrable to interference from a representation of the partner’s response. Alternatively, visually intact pictures might be more strongly linked to the response (i.e., the name of the pictured entity) than visually degraded pictures, for which the link might be less automatic, and hence more susceptible to interference.

In Experiment 4, the interaction between Partner and Size in the accuracy analyses was not predicted. Overall, it seems that speakers were mostly affected when naming big pictures in the DIFFERENT condition. In particular, they produced more incorrect responses, so that the proportion of incorrect trials when naming big pictures in the DIFFERENT condition (4.8%) approached the proportion of incorrect trials when naming small pictures (DIFFERENT: 4.0%, SAME: 4.8%, NO: 5.5%; averaged across conditions: 4.8%), whereas the proportion of errors on big pictures was much lower in the other two conditions (SAME: 3.3%, NO: 2.8%). This could be interpreted as suggestive evidence in favour of the co-representation account. We found that errors where more frequent in the DIFFERENT than in the SAME condition in Experiment 1 as well. However, note that the effect was not replicated in Experiment 2. In addition, the interaction between Condition and Size in Experiment 4 is difficult to interpret. In particular, since neither Size nor Partner interacted with Relatedness the increased error rate in the DIFFERENT condition could be due to low-level (visual) interference, rather than interference at the conceptual and lexical level.

In any case, the latency data did not provide any evidence in favour of the co-representation account. Experiment 4 successfully addressed the limitations of Experiment 3. There was no semantic interference effect, nor an interaction between Size and Relatedness in the latency analysis. This is consistent with demonstrations that phonological facilitation from distractor picture names is eliminated when processing resources are limited.
(Mädebach, et al., 2011). It suggests that distractor picture names were not routinely accessed in this experiment. This should provide a particularly favourable test of the predictions of the co-representation account.

Crucially, however, there was no interaction between Relatedness and Partner in the latency analysis. In addition, latencies were no longer in the DIFFERENT than in the SAME condition. Therefore, we found no evidence to support the co-representation account. In contrast, the results conformed to the predictions of the agent-conflict account.

### 2.6 General Discussion

In this study we found evidence that people represent whether their partner is naming, and that such representations interfere with concurrent language production processes. To our knowledge, this is the first demonstration of a joint interference effect in picture naming. Across four experiments we found no evidence to support the co-representation account of joint task effects. In other words, there was no evidence that participants were representing the content of the response their partner was preparing. Our results are more consistent with a version of the agent-conflict account, in which participants represent whether their partner is naming (and not just whether it is their partner’s turn to respond).

In Experiment 1, we established that speakers who are preparing two-word utterances respond more slowly when they believe that another speaker is preparing an

---

10 As already mentioned (see Discussion of Experiment 3), another possibility is that the relatedness effect was due to visual similarity between target and distractor pictures. If this explanation is correct, the effect was eliminated in Experiment 4 because degrading the pictures reduced visual similarity between semantically related pictures. We note that the combined analysis of Experiments 3 and 4 revealed also a Relatedness*Size interaction ($\chi^2(1) = 4.27, p<.05$) that mirrored the marginal trend in Experiment 3: a trend towards semantic facilitation when participants where naming big pictures ($B=-5$, SE=6, $t=-.76$), and a reliable semantic interference effect when participants where naming small pictures ($B=13$, SE=6, $t=2.17$; see Table B-8). We believe it might be due to residual visual interference between targets and distractors (on some subset of the items) in Experiment 4, which reinforced the effect found in Experiment 3.
utterance composed of the same two words (irrespective of whether they believe they will utter the words in the same or in reverse order). In Experiments 3 and 4, we showed that inhibition occurs even when speakers are preparing to utter single words while their partners are preparing to utter either the same or a different word. The effect was reliable when the pictures were visually degraded (Experiment 4); there was a non-significant trend when the pictures were visually intact (Experiment 3).

Interestingly, the result of Experiment 4 shows that participants represented their partner’s responses (to shared visual stimuli) when they did not overlap in content with their own responses. This is consistent with the results of the many studies that reported joint interference effects in manual tasks (see Knoblich et al., 2011). In all these studies, participants represented their partner’s (potential) response (which was evoked by irrelevant features of the stimulus) even though it was completely unrelated to their own response (as there was no overlap between the S-R mappings assigned to the two participants in a pair).

Additionally, in Experiment 2 participants were faster when they believed their partners were categorizing the pictures (with a verbal yes/no response) than when they believed they were naming the pictures. This finding is interesting for two reasons. First, it suggests that interference is specifically due to the belief that one’s partner is preparing a naming response (as opposed to any response). Second, it suggests that interference is (at least partially) due to the belief that another speaker is concurrently engaged in the process of translating a concept into language (i.e., the process of lexical access; Levelt, et al., 1999), and that it is not entirely due to the belief that another speaker is producing a verbal response.

In sum, our experiments showed that speakers represent whether another speaker is concurrently engaged in language production processes and that doing so interferes with the production of their own utterance. Therefore, we propose that speakers use their own language production mechanisms to represent whether another speaker is about to produce an utterance. Below, we first discuss whether our findings are consistent with any of the
three accounts put forward in the Introduction, and then we discuss how they could be accommodated within Pickering and Garrod’s (2013) model of language comprehension and production, and our own account presented in Section 1.4.3

First, our findings clearly rule out the no-representation account, as they show that speakers represent their partner’s verbal responses even when they are not interacting with them. Second, our findings do not support the co-representation account, as speakers experienced interference from representing their partner’s responses to a similar degree when they named the same pictures in the same order as their partner and when they did not (i.e., interference occurred irrespective of congruency).

Third, the findings from Experiments 1, 3, and 4 are generally consistent with the agent-conflict account (Wenke, et al., 2011), as they show that speakers represent whether it is their partner’s turn to respond. Interestingly, we know that people tend to avoid speaking at the same time (H. H. Clark, 1996; Sacks, et al., 1974; see also Schegloff, 2000). Turn-taking is a particularly important component of successful conversations, and knowing when it is one’s turn to speak is crucial to avoid overlapping with other speakers (see Section 1.1). This observation is compatible with our findings that speakers represent when another person is about to speak, even in the absence of interaction.

However, the agent-conflict account in its more general formulation (Wenke et al., 2011) cannot explain the results of Experiment 2. In that experiment we found indication that interference was larger when speakers believed their partners were naming pictures than when they believed they were producing a yes/no response. This finding suggests that speakers did more than just keeping track of their partners’ turns (respond vs. not respond). It suggests that speakers also represented the “type” of task their partners were about to perform (name vs. not name). But what kind of mechanism allowed them to do so? We propose that speakers covertly simulated their partner’s task.

The finding that speakers represent whether another speaker is about to produce an utterance based on the instructions assigned to the other speaker is consistent with the idea
that beliefs about other people’s behaviour can drive (or, indeed, are based on) internal simulations of their behaviour (e.g., Goldman, 2006). In addition, it is consistent with the idea that mental imagery is a form of internal simulation (Grush, 2004; Moulton & Kosslyn, 2009; Tian & Poeppel, 2010, 2013). Specifically, the finding that representing whether another speaker in concurrently engaged in language production processes interferes with production of one’s own utterances is consistent with accounts in which production mechanisms (i.e., representations recruited during utterance planning and/or execution) are activated when representing (that is, imagining, or comprehending) somebody else’s utterances.

Pickering and Garrod (2013), as described in Section 1.4.2, proposed that people can comprehend others’ utterances using a mechanism called prediction-by-simulation. According to this, the comprehender covertly imitates the incoming sensory input to recover a production command, a representation of the intended message so far. Having done so, the comprehender can run the recovered production command through his own language production system, thus producing an utterance that imitates the comprehended utterance. Importantly, production of the imitative response does not necessarily lead to overt articulation, as the response can be inhibited at any stage during the process.

In addition, the comprehender can compute what production command he would be likely to use next if he were speaking, and run the new production command through his own language production system. Third, he can feed the new production command into a forward model. In sum, another speaker’s utterance can implicate the comprehender’s production mechanisms via imitation and prediction. If the comprehender is concurrently preparing to speak, there may be interference between comprehension and production.

Pickering and Garrod’s (2013) account is concerned with the representation of other people’s utterances during comprehension. However, the mechanisms used for covert imitation and prediction-by-simulation can also be used in situations in which the other person’s utterance is not heard, as in our experiments. In this case, the participant does not

143
recover his partner’s production command from the initial part of his partner’s utterance (as described above), but rather recovers it from the instructions (and hence the belief that his partner is preparing to speak at the same time as him). In this way, Pickering and Garrod’s account could be extended to include the possibility that representations of other speakers’ utterances can be formed even in the absence of any perceptual information about those utterances.

In our experiments, participants recovered their partners’ production command (i.e., they represented their partner’s intention to name), but they did not begin production of an utterance that would imitate their partner’s utterance. If they had done so, we would have observed more interference when the utterances were incongruent than when the utterances were congruent (see Section 1.5.1). So, why did participants experience interference and took longer to name pictures when they believed their partners were naming pictures themselves.

One possibility is that the recovered production command specified only the partner’s intention to name, but not the content of the partner’s upcoming utterance. When the participants ran the recovered production command through their own production systems, at the same time as they were preparing to name (according to their own instructions), interference was caused by a conflict between the speaker’s own intention to name and a representation of the partner’s intention to name.

Another possibility, however, is that participants ran the recovered command through their forward production model, thus anticipating the production of an utterance. This forward-model prediction could have triggered the allocation of resources away from the concurrent process of producing the participant’s own utterance, thus leading to slower naming latencies when the participants believe their partner was naming at the same time as them.

Future studies should investigate the conditions under which naming responses are affected by beliefs about another speaker’s speech. For example, is it necessary that the participants believe another speaker is naming the pictures (and not, say, a speech
synthesizer)? Tsai et al. (2008) showed that joint Simon effects are not present when participants believe their partner is not an intentional agent (i.e., a computer). In a similar way, participants might not represent whether their partners are naming if these partners were not intentional agents. This would be compatible with Pickering and Garrod’s (2013) suggestion that the prediction-by-simulation route might be used only if the speaker is sufficiently similar to the comprehender.

In conclusion, we have shown that naming responses are inhibited by the belief that another speaker is concurrently preparing to produce a naming response. This suggests that language production mechanisms can be used to represent whether another speaker is about to engage in language production, even in non-interactive contexts.

2.7 Acknowledgments

We thank Joris Van de Cavey, who contributed to an equal degree as the author to the design, the testing of participants, and the scoring of participants’ responses in Experiment 1. He also contributed to the initial stages of analyses and interpretation of the findings, and wrote a script to facilitate the scoring of participants’ responses.

We also thank Sven Radcke for assistance in scoring the participants’ responses in Experiment 3 and Experiment 4. Finally, we thank Eddie Dubourg and Ziggy Campbell for assistance with recording equipment.
3. Study 2. Experiment 5: effect of anticipating another’s utterance on stopping speech

3.1 Introduction

In Section 1.4.2 we reviewed evidence that comprehenders predict upcoming linguistic input, on the basis of multiple sources of information, and at multiple linguistic levels (Huettig, et al., 2011; Kutas, et al., 2011; Pickering & Garrod, 2007, 2013; Van Petten & Luka, 2012). We then introduced the hypothesis (Pickering & Garrod, 2007, 2013) that comprehenders compute predictions using language production mechanisms, and concluded that at present the evidence in support of this hypothesis is mostly indirect (Sections 1.4.2.2 and 1.4.2.3).

Only one study showed that production mechanisms might be causally involved in prediction (Lesage, et al., 2012). In that study, repetitive TMS applied to the right cerebellum impaired listeners’ ability to make semantic predictions during comprehension in a visual world paradigm. The conclusions of Lesage et al.’s study rest on the assumption that the right cerebellum is the site of forward model computations during language production. This assumption is incorporated in influential models of language production (Guenther, et al., 2006; Hickok, 2012). However, direct evidence that language production mechanisms are implicated in the computations of predictions would require a paradigm in which the computation of predictions about another’s utterance and production of an utterance take place concurrently. In this chapter, we report a study that uses one such paradigm. In

---

11 The study presented in this chapter was designed and carried out by the author in collaboration with Uschi Cop, while she was visiting the Department of Psychology at the University of Edinburgh as part of an MSc Program in Experimental Psychology at the University of Ghent, Belgium. This chapter is based on a manuscript prepared for publication (Gambi, C., Cop, U., & Pickering, M.J. (2013). How do speakers coordinate their utterances? Evidence for prediction of another’s utterances in a joint language task. Unpublished manuscript, The University of Edinburgh.)
particular, it is based on the rationale for consecutive production tasks we introduced in Section 1.5.2.

In a nutshell, this study tested the hypothesis that comprehenders compute predictions using language production mechanisms by investigating whether predictions can affect language production on-line. If the same mechanism is used concurrently to produce an utterance and to predict another speaker’s utterance, then we would expect the latter process to affect the former. This means that production of a speaker’s utterance should be affected by the fact that this speaker simultaneously predicts another speaker’s (upcoming) utterance.

We devised a joint language production task that requires participants to take turns speaking. Two speakers, $A$ and $B$, sit next to one another in front of the same computer screen. Speaker $A$ produces an utterance (a picture’s name), and then speaker $B$ produces a second utterance (another picture’s name, which is unrelated to the first picture’s name). We call this the joint word-replacement task. We ask whether and how the way speaker $A$ produces the first utterance is affected by the fact that speaker $B$ will later produce another utterance. We are interested in speaker $A$’s (rather than $B$’s) utterance because it occurs first. Therefore, any effects of $B$’s utterance on $A$’s utterance would be due to $A$’s prediction of $B$’s utterance (and not to $A$’s comprehension of $B$’s utterance).

### 3.1.1 Three accounts of the relationship between prediction and production mechanisms

If speaker $A$ indeed predicts $B$’s utterance, the mechanism(s) she uses to compute this prediction could stand in one of three relationships with respect to the mechanism(s) she uses to produce her own utterance. One possibility is that prediction mechanisms are entirely independent from production mechanisms. For example, predictions could be computed using a comprehension-based mechanism. Note that speaker $A$ does not comprehend any part of $B$’s utterance before she has finished her own utterance (as $B$ starts speaking only after $A$
has stopped). But it is possible that A predicts B’s utterance by inferring what B is most likely to do given the instructions A received about the task, perhaps using mechanisms involved in elaborative (i.e., predictive) inferences (e.g., McKoon & Ratcliff, 1986).

Note that many scholars assume (either implicitly or explicitly) that in language comprehension people use different mechanisms than the ones they use in language production (Pickering & Garrod, 2013). Therefore, if this account is correct, predicting another speaker’s utterance should have no direct and immediate effect on production processes. Say that speaker A infers that B is about to produce the second picture name. Even if A constructs this inference very quickly (within the space of planning and uttering a single word), there is no reason to expect that doing so should affect A’s production of her own utterance. Therefore, we term this the separate mechanisms account.

A second possibility is that the mechanism that speaker A uses to predict B’s utterance is precisely the mechanism she uses to produce her own utterance. In other words, predicting B’s utterance would entail the same processes on the part of speaker A that she would use to produce B’s utterance herself. We term this the shared mechanisms account, as the computation of such predictions is implemented in the language production system. If this account is correct, predicting another speaker’s utterance should have the same effect as preparing to produce that utterance.

A third possibility is that predictions of others’ utterances are computed using some of the mechanisms used when producing utterances, but not all. We term this the overlapping mechanisms account of prediction. If this account is correct, predicting another speaker’s utterance should have some effect on production processes, but this effect might be different from the effect of preparing to produce the same utterance.

In order to distinguish among these three possibilities, we tested two more versions of our word-replacement task in addition to the joint version described above. In the no-replacement version of the task, speaker A names the first picture but speaker B (who is present) does not name the second picture. Therefore, speaker B remains silent and there is
simply no utterance to predict. In the solo word-replacement task, instead, speaker $A$ names the first picture and then names the second picture as well. (Speaker $B$ is still present and performs the same solo task on different trials.) Therefore, speaker $A$ needs to retrieve the second picture name and will go through all the stages of language production (from concept selection to articulation), as assumed by all theories of single word production (e.g., Dell, 1986; Levelt, et al., 1999).

If prediction mechanisms are completely independent of production mechanisms, the joint word-replacement task should be equivalent to the no-replacement task: Even if $A$ predicts $B$’s utterance in the joint version, $A$’s predictions should not affect the way $A$ produces her utterance. On the contrary, if the shared mechanisms account is correct, then the joint version of the word-replacement task should be equivalent to the solo word-replacement task, as $A$ would predict $B$’s utterance in the joint task using the same mechanisms that she uses to prepare her second utterance in the solo task.

Finally, if the overlapping mechanisms account is correct, the joint word-replacement task will not be equivalent to the no-replacement task, because in the joint task $A$ predicts $B$’s utterance using some mechanisms that can affect production. However, the joint version need not be equivalent to the solo version, because in the joint task $A$ predicts $B$’s utterance using some mechanisms that are used during language production, but not the full range of language production mechanisms.

Importantly, the solo word-replacement and the no-replacement tasks have already been used by Hartsuiker, Catchpole, De Jong, and Pickering (2008), so that we can use their findings to formulate specific hypotheses about the joint word-replacement task. Hartsuiker et al. were interested in how speakers coordinate the planning and articulation of two utterances in speech repairs. We briefly review this literature below to consider which factors might affect performance in the solo version of the task, and then ask whether similar factors would affect our joint task.
3.1.2 Coordinating stopping and resuming speech

In instances of self-repair, the speaker coordinates the planning and articulation of two utterances (phrases, words, or sounds): the initial utterance and the replacement. Thus in “Left – er – right in front of me” (Levelt, 1989, p. 484), the initial word (left) is completed, and then the replacement (right) follows after an editing expression (er). But sometimes the initial word is not completed, as in “To the left side of the purple disk is a v – a horizontal line” (most likely, the intended word was vertical; Levelt, 1989, p. 474). These examples illustrate that there is variability as to where speakers stop (between words or within words) when they detect an inappropriate word and correct themselves.

This observation has sparked considerable theoretical (Hartsuiker & Kolk, 2001; Levelt, 1983, 1989; Nooteboom, 1980) and empirical investigation, both in the form of corpus studies (Blackmer & Mitton, 1991; Seyfeddinepur et al., 2008) and experiments (Hartsuiker, et al., 2008; Hartsuiker, Pickering, & De Jong, 2005; Tydgat, Stevens, Hartsuiker, & Pickering, 2011; cf. Van Wijk & Kempen, 1987). In these experiments, self-repairs are induced by asking participants to describe an initial picture, which is then replaced by a target picture. This sometimes causes participants to reformulate their utterances.

Note that the situation in such experiments differs from situations in which the speaker detects an internally generated error. Specifically, the need to reformulate is caused by a change in the environment (see Tydgat et al., 2011, p. 360 for discussion). However, this feature of the task is useful for our purposes, as it makes the solo version of the task more comparable to the joint version of the task.

According to the account proposed by Hartsuiker and Kolk (2001), and modified by Hartsuiker et al. (2008) and Tydgat et al. (2011), the speaker simultaneously initiates two processes when executing a self-repair: the process of stopping articulation and the process of planning the replacement. These processes proceed in parallel and share a limited pool of resources. Therefore, the process of planning the replacement competes with the process of
stopping the initial word. In other words, the speaker uses production mechanisms to begin preparing the target word while also trying to stop the planning and articulation of the initial word, therefore incurring interference. This hypothesis is supported by two findings.

First, Hartsuiker et al. (2008) found that speakers complete initial words more often when they subsequently have to produce a replacement (53.9%) than when they simply have to stop speaking (21.5%; see also Tydgat et al., 2011). Thus, speakers find it harder to stop the initial word when they also need to start preparing a replacement than when they do not. This could reflect competition between a “go” and a “stop” signal, in line with the account proposed by Hartsuiker et al.\(^ {12}\) Alternatively, speakers might strategically evaluate whether it is better to interrupt the initial word as quickly as possible or to continue with articulation (Tydgat et al., 2011; cf. Seyfeddinipur et al., 2008, when self-repairing internally generated errors); by continuing to articulate previously planned material, speakers could in fact re-allocate resources from stopping to replacement planning, and ensure that the replacement is ready in a reasonable time span. Second, Hartsuiker et al. (2008; Experiment 1) found that the spoken duration of the initial word is longer when planning the replacement is more difficult, as when the target picture is degraded.

To sum up, replacing a word with another involves the coordination of two processes that compete for resources: 1. stopping the initial word; 2. planning the replacement. In the solo version of the word-replacement task, one speaker carries out both processes. In the no-replacement task, instead, the speaker stops the initial word but does not plan the replacement (so only one process is involved). Crucially, in the joint version, the two processes are distributed between two speakers: Speaker A plans the initial word and stops, and speaker B plans the replacement. The question we ask in this study is what sort of mechanism does speaker A use when she predicts that B is about to speak. The next section

\(^ {12}\) Hartsuiker et al. (2008) did not comment on this aspect of their results (and did not test for it statistically).
describes the experimental conditions and presents the expected findings according to the three accounts of prediction discussed above.

### 3.1.3 An experimental comparison of the accounts

In three conditions, a pair of participants viewed a picture that appeared on a shared screen, and we cued one or the other participant to name that picture. On a small proportion (9%) of trials, the initial picture changed into a target picture (as in Hartsuiker et al., 2008). When the change occurred, the participant was instructed to stop naming the initial picture as quickly as possible.

In all conditions, the cued participant varied across trials. Instructions about the target picture depended on the condition to which the participant was assigned. In the SELF condition (solo task), the cued participant also named the target picture. This condition therefore followed Hartsuiker et al.’s (2008) Experiment 1, except that it involved two (co-present) participants. In the OTHER condition (joint task), the other (non-cued) participant named the target picture. In the NO condition (no-replacement task), neither participant named the target picture. This last condition therefore followed Hartsuiker et al.’s Experiment 2, except that it again involved two participants. Following the results of Hartsuiker et al., we hypothesized that participants in the SELF condition would complete the initial word more often than participants in the NO condition. This finding would confirm that participants in the SELF condition were planning the target picture name before stopping the initial name and that these processes competed for resources.

Note that the presence of another person can affect individual performance in complex ways, sometimes yielding facilitation, sometimes interference (e.g., Klauer, Herfordt, & Voss, 2008). Indeed, the presence of another person serves as a retrieval cue for words that have been uttered by that person and facilitates picture naming (Horton, 2007). So it was important to investigate whether Hartsuiker et al.’s (2008) results would be replicated in the presence of another person. To further ensure comparability between our results and
theirs, in all conditions the target picture was either intact or degraded (with 50% of its contours removed). Based on Hartsuiker et al.’s Experiment 1, we expected participants in the SELF condition to stop naming the initial picture later when the target picture was degraded versus intact (i.e., a degradation effect); based on Hartsuiker et al.’s Experiment 2, we expected no degradation effect in the NO condition.

Consider now the novel OTHER condition. Let us assume that, at some point during the process of stopping the initial word, speaker A predicts that speaker B is about to name the target picture. If prediction mechanisms are completely independent of production mechanisms (as the separate mechanisms account assumes), speaker A’s prediction will not affect her ability to stop producing the initial word. Therefore, A should find stopping the initial word no harder in the OTHER condition than in the NO condition. More specifically, she should be no more likely to complete the initial word in the OTHER than in the NO condition.

But if, on the contrary, A predicts that B is about to name the target picture using the same mechanisms she would use when she prepares to name the target picture herself (as the shared mechanisms account assumes), then A’s prediction should affect her ability to stop producing the initial word. More precisely, A’s prediction of B’s utterance should affect her ability to stop producing the initial word in the same way as the process of planning the target word would affect her ability to stop (i.e., as in the SELF condition). Therefore, A should be more likely to complete the initial word in the OTHER than in the NO condition and, moreover, A should be as likely to complete the initial word in the OTHER as in the SELF condition.

Finally, if A predicts that B is about to name the target picture using some of the mechanisms that are used in production, but not all of them (as the overlapping mechanisms account assumes), then A’s prediction should still affect her ability to stop producing the initial word. Crucially, however, this effect need not be the same as in the SELF condition. One possibility is that A might find it less hard to stop in the OTHER than in the SELF
condition (while still finding it harder than in the NO condition, where production mechanisms are not used at all). Table 3-1 summarizes the differences between the accounts.

Table 3-1.

\textit{Anticipated likelihood of completing the initial word in the SELF, OTHER, and NO conditions according to the separate mechanisms, shared mechanisms, and overlapping mechanisms accounts.}

<table>
<thead>
<tr>
<th>Condition</th>
<th>SELF</th>
<th>OTHER</th>
<th>NO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Separate mechanisms</td>
<td>High</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Shared mechanisms</td>
<td>High</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Overlapping mechanisms</td>
<td>Highest</td>
<td>Higher</td>
<td>Low</td>
</tr>
</tbody>
</table>

### 3.2 Method

**Participants**

Ninety-six students from the University of Edinburgh participated in the experiment. They were either paid £6 or received course credit in return for participation. All were native English speakers and reported no language impairment. Participants were matched to form 48 pairs, which were then randomly assigned to each of the three conditions. Thus, we tested 16 pairs of participants in each condition. Most participants did not know their partners beforehand.

**Materials**

The materials were simple black and white line drawings. There were 32 target pictures, each of which appeared in an intact and a degraded format. These were the target
pictures used by Hartsuiker et al. (2008), derived from a set originally developed by Meyer et al. (1998). To create the degraded versions, Meyer et al. deleted “50% of the black pixels, in regions where they could be reconstructed by straight or smoothly curved lines” (p. 27). There were 32 initial pictures, also taken from Hartsuiker et al., and 128 filler pictures from Snodgrass and Vanderwart (1980). Each of the 64 experimental items constituted a unique combination of an initial picture and a target picture. The pictures were combined in such a way that every target picture occurred after 2 initial pictures and every initial picture preceded 2 target pictures (e.g., glasses-mouse, glasses-wall, orange-mouse, orange-wall). In each item, the names for the initial and target pictures had different initial phonemes and unrelated meanings (see Appendix A, Table A-2 for a complete list of the experimental items). Initial pictures and filler pictures were presented inside a colored frame (green or red) in order to cue one participant to name that picture (see Procedure). Target pictures were presented without a frame (the instructions made clear who was to name a given target picture; see Procedure). Initial and target pictures were used both on change (experimental) and no-change (filler) trials, whereas filler pictures were used only on no-change trials.

**Design**

Degradation (intact vs. degraded) was varied within participants and items. Condition (SELF, OTHER, NO) was varied between participants but within items. We first created four lists containing the 64 experimental items (change trials). Every initial picture and target picture occurred twice in each list of change trials. Each target picture appeared once degraded and once intact. If the degraded version of a target picture was preceded by an initial picture with a green (red) frame on one change trial, the intact version of the same target picture was preceded by an initial picture with a red (green) frame on another change trial; also, each initial picture occurred once in each color. This meant that each participant in a pair named each initial picture and each target picture only once on change trials (though, of course, they saw each initial and each target picture twice). In addition, we divided change
trials into two blocks and each initial picture and each target picture appeared once in each block. For each experimental item, every combination of color-assignment and target degradation (red initial – degraded target, red initial – intact target, green initial – degraded target, green initial – intact target) occurred once across lists. For each of the 4 lists, we derived 4 random orders, with the constraint that each block appeared first in half these orders.

We also constructed 2 lists of no-change trials. Each contained 640 items: the 32 target pictures twice (once degraded, once intact) in isolation; the 32 initial pictures twice in isolation; and the 128 fillers, four times each. The two lists were constructed so that the target pictures and initial pictures that were presented in one color in the first list were the other color in the second list. They were also split into two blocks, with repetitions of the same picture being equally distributed between blocks. To create running lists, one no-change trial list was combined with one change trial list. The pairing was done in such a way that target pictures in the change trial list had the opposite color and the opposite degradation relative to their instances (within the same block) in the no-change trial list. Change trials were quasi-randomly interspersed with no-change trials, with the constraint that each change trial was separated by at least three no-change trials. The same 16 running lists were then presented in each of the three between-participants conditions.

Procedure

The experiment was controlled using E-Prime (Version 2.0). First the participants were introduced to each other and told that they were going to do a task together. They were then familiarized with the materials in individual booths. They were shown the 192 pictures (32 initial pictures, 32 target pictures, 128 fillers) with the corresponding names, and were instructed to read the names out loud to aid memory. Next, the two participants were seated in front of the same computer screen. Half of the pairs were instructed that the pictures in the
green frame were to be named by the person on the left, and the pictures in the red frame were to be named by the person on the right. The other half of the pairs were instructed that the pictures in the red frame were to be named by the person on the left, and the pictures in the green frame were to be named by the person on the right. They were told to use the names that they had learned during the naming phase.

The instructions about change trials depended on condition. For pairs in the SELF condition, cued participants were instructed to stop naming the initial picture and name the interrupting picture as fast as possible. Therefore, in the SELF condition the participant who responded to the initial picture also responded to the target picture on the same trial. Cued participants in the OTHER condition were also instructed to stop naming the initial picture as fast as possible, but this time the other participant had to name the interrupting picture. Therefore, in the OTHER condition one participant responded to the initial picture and the other responded to the target picture on the same trial. The color of the initial picture frame indicated who was to perform which task on any given change trial. Both participants performed each task equally often in each block, while they took turns according to a randomized sequence. Finally, in the NO condition, cued participants were again instructed to stop naming the initial picture as fast as possible, but they were told to ignore the interrupting picture. Therefore, in the NO condition none of the participants responded to the target picture on change trials (see Figure 3-1).
Figure 3-1.

Schematic representation of the three experimental conditions with an example degraded change-trial; A is the cued participant.
Before starting the experiment, the participants completed 8 practice trials. These were 5 no-change trials and 3 change trials on which filler pictures were used instead of experimental pictures. After the practice, the instructions were quickly summarized again and the participants were warned that some of the pictures would consist of dashed lines.

All trials started with a fixation cross which remained on the screen for 2500ms. On no-change trials an initial picture (with a colored frame) then appeared for 500ms. On change trials the initial picture appeared for 300ms and was then replaced by a target picture (without a colored frame) that appeared for 500ms. The inter-trial interval was 3300ms after a change trial and 3000ms after a no-change trial. Participants spoke into head-mounted microphones and their responses were digitally recorded on two separate channels. For each change trial, two audio files were generated (and automatically stored), one time-locked to initial-picture onset, the other time-locked to target-picture onset. An experimental session lasted approximately 45 minutes.

**Scoring**

Only change trials are relevant for our hypotheses, so only the audio files recorded during these trials were analyzed. Data from 7 pairs (3 in the SELF condition, 2 in the OTHER condition, 2 in the NO condition) had poor audio quality, and so background noise was reduced by batch processing their change-trial files, using Adobe Audition. Responses that were still inaudible or could not be categorized were excluded from further analyses; if there were more than 10 such trials for a single pair, the whole set of data for that pair was discarded. This resulted in the loss of 1.8% of the data in the SELF condition, 1.4% in the OTHER condition, and 3.5% in the NO condition, in which one pair was discarded.

The remaining responses were annotated off-line (half by the author, half by her collaborator, see Footnote 1). We first noted errors and disfluencies (e.g., *um*, repetitions) in producing the initial or target name (in SELF and OTHER); see the Results section for percentages. For two target pictures (*mouth, steps*), participants responded with an
alternative name (*lips*, *stairs*) on at least 39% of the trials. As these were clearly acceptable responses, we included these trials in the analyses. All other naming errors were coded as such and the corresponding trials were discarded. Correct and fluent initial responses were divided into three response types: completed initial (*e.g.*, *apple chair*), interrupted initial (*e.g.*, *ap-chair*), and skipped initial (*e.g.*, *chair*).

Second, three time-points were manually annotated on the audio files using the phonetic analysis software Praat (Boersma & Weenink, 2010): the onset of the initial name, the offset of the initial name, and the onset of the target name (in the SELF and OTHER conditions). We used these time-points to determine the following time measures: Initial Onset (onset of initial name relative to onset of the initial picture); Initial Duration (onset of initial name to offset of initial name); Target Onset (onset of target name relative to onset of the target picture, in the SELF and OTHER conditions); Interval (offset of initial name to onset of target name, in the SELF and OTHER conditions). The primary measure of interest was Initial Duration. However, we also analyzed the other time measures, in part to determine whether our results were consistent with Hartsuiker et al. (2008). Summary tables and a brief description of these results can be found in Appendix B. Additional data exclusion and trimming criteria for the time measures are reported in the Results section (for Initial Duration) and in Appendix B (for the other measures).

**Data Analysis**

The data were analyzed using Generalized Linear mixed-effects models (Baayen, et al., 2008; Bates, et al., 2008) in R (Version 2.7.2). For the response type data, we used a logistic link function (Jaeger, 2008) and conducted a binomial analysis (comparing the likelihood of a completed response against the likelihood of observing any of the two other kinds of responses; *i.e.*, an interrupted or a skipped response). For Initial Duration, we only included completed initial responses in the analysis. This was motivated by the fact that there were no specific predictions for the factor Response Type. In addition, completed responses
were more evenly distributed than interrupted responses (see Table 3-2) and we hoped, in this way, to minimize issues related to the imbalance in the proportion of response types across conditions (see below). Consequently, the predictors of interest for the analysis of Initial Duration were only Condition and Degradation. See Appendix B for details of the analyses of the other time measures.

In all analyses, we started by fitting the complete model; we then removed predictors that were not significant from the model, using a backward stepwise procedure, and stopped whenever removing a predictor caused a significant loss of fit (assessed using a log-likelihood ratio test). We report coefficients, standard errors, and Wald’s t-tests from the complete model together with results of the likelihood ratio test for each predictor (Barr, 2008; Quené & van den Bergh, 2008). Regarding random effects, we started with the full random effect structure, including random slopes (for all factors and their interaction) and random intercepts for both subjects and items (defined as a combination of initial and target picture). Given that random slopes are only appropriate for within-subjects and within-items factors, we included by-subjects random slopes for Degradation and by-items random slopes for Degradation, Condition, and their interaction. If the model with full random effects specification did not converge, we simplified it by removing the higher-order term (interaction of Condition and Degradation). We then tested whether specific random effects significantly contributed to model fit using likelihood ratio tests. We report estimates of the variances and covariances of all random effects that passed the test (with an alpha-value of .1 instead of .05 to account for the conservativity of these tests).

We used sum coding for our predictors, both in the response type analyses and in the analyses of Initial Duration. For the analyses of the other time measures, we used contrast (Helmert) coding, so that the coefficients associated with the factor Condition could be more easily interpreted (see Appendix B for further details). Because Response Type was not under experimental control and was in fact affected by Condition (see below), the number of observations per cell varied widely, leading to a highly imbalanced design for the time
analyses. This means that in order to have weighted estimates for the fixed effects (and to make sure the contrasts are as close to orthogonal as possible), it is necessary to weight the contrasts by the observed cell counts. We therefore used weighted contrast coding (Cohen, Cohen, West, & Aiken, 2003; Serlin & Levin, 1985; West, Aiken, & Krull, 1996) for all the predictors entered in the analyses of the time measures (see Appendix B for an example).

3.3 Results

3.3.1 Response type data

As stated in the Scoring section, for the analyses of response type we excluded the trials where the initial picture was not named correctly or the initial name contained hesitations or repetitions (5.1% in the SELF condition, 5.8% in the OTHER condition, 4.7% in the NO condition). This left us with 963 data points in the SELF condition, 962 in the OTHER condition, and 914 in the NO condition. The percentages of Completed, Interrupted, and Skipped initial responses in each condition are reported in Table 3-2, separately for degraded versus intact trials.
Table 3-2.

Percentages of Completed, Interrupted, and Skipped Initial Responses by Degradation Level and Condition.

<table>
<thead>
<tr>
<th>Degradation</th>
<th>Completed</th>
<th>Interrupted</th>
<th>Skipped</th>
</tr>
</thead>
<tbody>
<tr>
<td>SELF</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Degraded</td>
<td>84.8%</td>
<td>6.3%</td>
<td>9.0%</td>
</tr>
<tr>
<td>Intact</td>
<td>82.0%</td>
<td>6.8%</td>
<td>11.2%</td>
</tr>
<tr>
<td>Total</td>
<td>83.4%</td>
<td>6.5%</td>
<td>10.0%</td>
</tr>
<tr>
<td>OTHER</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Degraded</td>
<td>64.9%</td>
<td>20.8%</td>
<td>14.3%</td>
</tr>
<tr>
<td>Intact</td>
<td>61.5%</td>
<td>23.3%</td>
<td>15.2%</td>
</tr>
<tr>
<td>Total</td>
<td>60.0%</td>
<td>23.5%</td>
<td>14.8%</td>
</tr>
<tr>
<td>NO</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Degraded</td>
<td>46.1%</td>
<td>28.2%</td>
<td>25.7%</td>
</tr>
<tr>
<td>Intact</td>
<td>46.9%</td>
<td>23.5%</td>
<td>29.6%</td>
</tr>
<tr>
<td>Total</td>
<td>46.5%</td>
<td>25.8%</td>
<td>27.7%</td>
</tr>
</tbody>
</table>

*Note.* The percentages are based on a total of 963 trials in the SELF condition, 962 trials in the OTHER condition, and 914 trials in the NO condition.

Participants completed the initial name more often in the OTHER condition (60.0%) than in the NO condition (46.5%). They also completed the initial name more often in the SELF condition (83.4%) than in the OTHER condition. The best-fitting model included only Condition as a predictor, whereas Degradation had no effect on the proportion of completed responses, nor did the interaction. No random slopes were justified, so only random intercepts were retained (see Table 3-3). In order to test the hypotheses of the three accounts laid out in the Introduction, we then set the OTHER condition as the reference level, and we defined two contrasts, one comparing the mean of the OTHER condition to the mean of the NO condition (Condition1), the other comparing the mean of the OTHER condition to the mean of the SELF condition (Condition2).
The first contrast therefore tests whether speakers tend to complete the initial word more often in the OTHER than in the NO condition, which would be compatible with both the shared mechanisms and the overlapping mechanisms account (but not with the separate mechanisms account). The second contrast tests whether speakers tend to complete more in the SELF than in the OTHER condition, which would be compatible with the overlapping mechanisms account but not with the shared mechanisms account. Importantly, when we fit a model that included only the factor Condition, with the two contrasts defined above, both contrasts were associated with estimates significantly different from zero (Condition1: $B= -0.99$, SE = .46, $z = -2.16$, $p<.05$; Condition2: $B= 1.57$, SE = .47, $z = 3.36$, $p<.001$). Overall, these results are compatible with the overlapping mechanisms account, but not with the shared mechanisms account or the separate mechanisms account.

Table 3-3.

*Response Type analyses: complete model.*

<table>
<thead>
<tr>
<th>Fixed effect</th>
<th>Coef.</th>
<th>SE</th>
<th>z</th>
<th>$\chi^2$</th>
<th>p</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept (unweighted grand mean)</td>
<td>1.01</td>
<td>0.20</td>
<td>5.07</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Degradation (Intact – Degraded)</td>
<td>-0.15</td>
<td>0.10</td>
<td>-1.49</td>
<td>1.73</td>
<td>.19</td>
<td>1</td>
</tr>
<tr>
<td>Condition1 (NO-grand mean)</td>
<td>-1.19</td>
<td>0.27</td>
<td>-4.43</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Condition2 (SELF – grand mean)</td>
<td>1.38</td>
<td>0.27</td>
<td>5.06</td>
<td>26.56</td>
<td>&lt;.001</td>
<td>2</td>
</tr>
<tr>
<td>Degradation*Condition1</td>
<td>0.21</td>
<td>0.14</td>
<td>1.52</td>
<td>2.2</td>
<td>.32</td>
<td>2</td>
</tr>
<tr>
<td>Degradation*Condition2</td>
<td>-0.17</td>
<td>0.17</td>
<td>-1.11</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject: Intercept</td>
<td>3.05</td>
<td>NA</td>
</tr>
<tr>
<td>Item: Intercept</td>
<td>0.24</td>
<td>NA</td>
</tr>
</tbody>
</table>
3.3.2 Initial duration

For the analyses of Initial Duration, we removed all trials that were more than 2.5 SD from the grand mean or more than 3 SD from the by-subject mean (2.5% in SELF, 2.0% in OTHER, 1.7% in NO). As stated above, we limited our analyses to completed initial responses. Apart from this, we conducted the same analyses as for the response type data. In addition, we conducted separate analyses for the three conditions in order to compare our results directly to Hartsuiker et al.’s (2008) findings.

Table 3-4.

*Initial Duration by Initial Response Type and Degradation in the three conditions.*

<table>
<thead>
<tr>
<th>Degradation</th>
<th>Completed</th>
<th>Interrupted</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SELF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Degraded</td>
<td>413 (80)[397]</td>
<td>279 (88)[26]</td>
<td>405 (87)[423]</td>
</tr>
<tr>
<td>Intact</td>
<td>401 (82)[393]</td>
<td>272 (105)[28]</td>
<td>392 (90)[421]</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>407 (81)[790]</td>
<td>275 (96)[54]</td>
</tr>
<tr>
<td></td>
<td>OTHER</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Degraded</td>
<td>400 (91)[301]</td>
<td>248 (94)[99]</td>
<td>363 (112)[400]</td>
</tr>
<tr>
<td>Intact</td>
<td>398 (93)[291]</td>
<td>245 (91)[113]</td>
<td>355 (115)[404]</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>399 (92)[592]</td>
<td>246 (92)[212]</td>
</tr>
<tr>
<td></td>
<td>NO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Degraded</td>
<td>394 (102)[201]</td>
<td>248 (114)[121]</td>
<td>339 (128)[322]</td>
</tr>
<tr>
<td>Intact</td>
<td>385 (92)[216]</td>
<td>250 (97)[105]</td>
<td>341 (113)[321]</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>389 (97)[417]</td>
<td>249 (106)[226]</td>
</tr>
</tbody>
</table>

*Note. Mean values in ms (standard deviation within round brackets)[cell count in square brackets].*
Table 3-5.  

*Initial Duration analyses (completed responses only): complete model.*

<table>
<thead>
<tr>
<th>Fixed effect</th>
<th>Coef.</th>
<th>SE</th>
<th>t</th>
<th>$\chi^2$</th>
<th>p</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept (weighted grand mean)</td>
<td>402.51</td>
<td>7.78</td>
<td>51.76</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Degradation (Intact – Degraded)</td>
<td>-10.09</td>
<td>6.05</td>
<td>-1.67</td>
<td>2.96</td>
<td>.09</td>
<td>1</td>
</tr>
<tr>
<td>Condition1 (NO-grand mean)</td>
<td>-20.90</td>
<td>15.82</td>
<td>-1.32</td>
<td>1.60</td>
<td>.45</td>
<td>2</td>
</tr>
<tr>
<td>Condition2 (SELF-grand mean)</td>
<td>13.83</td>
<td>15.38</td>
<td>0.90</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Degradation*Condition1</td>
<td>12.11</td>
<td>18.82</td>
<td>0.64</td>
<td>5.20</td>
<td>.07</td>
<td>2</td>
</tr>
<tr>
<td>Degradation*Condition2</td>
<td>-35.37</td>
<td>15.86</td>
<td>-2.23</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random effect</th>
<th>Variance</th>
<th>Corr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject: Intercept</td>
<td>2539.26</td>
<td></td>
</tr>
<tr>
<td>Item: Intercept</td>
<td>1837.29</td>
<td></td>
</tr>
<tr>
<td>Item: Condition1</td>
<td>55.231</td>
<td>1.00</td>
</tr>
<tr>
<td>Item: Condition2</td>
<td>464.51</td>
<td>-.60</td>
</tr>
</tbody>
</table>
Participants took 12 ms longer to stop before they named degraded than before they named intact targets in the SELF condition (see Table 3-4, completed responses). The inclusion of by-item random slopes for the factor Condition significantly improved fit (see Table 3-5). The main effect of Degradation marginally improved fit (p=.09), as did the interaction of Degradation and Condition (p=.07); the main effect of Condition was not significant When we fitted separate models to the three conditions (Table 3-6), we found a
degradation effect in the SELF condition (p<.01) but not in the OTHER or NO conditions (both ts < 1).

3.4 Discussion

We investigated whether participants in a joint language production task predict that their partner will speak using language production mechanisms and whether such prediction affects production of their own utterance. To do so, we compared a solo word-replacement task (the SELF condition), a joint word-replacement task (the OTHER condition), and a no-replacement task (the NO condition).

We found that participants completed their first utterance more often in the OTHER condition (60.0%) than the NO condition (46.5%), but less often in the OTHER condition than the SELF condition (84.3%). Therefore, we replicated Hartsuiker et al.’s (2008) findings in a two-person setting, as participants completed the initial word more often when they later named the target word than when they did not name the target word. The tendency to complete the initial word was greater in our SELF condition than in their Experiment 1 (53.9%), and similarly larger in our NO condition than in their Experiment 2 (21.5%), perhaps because the lower percentage of change trials in our study (9%) than in theirs (12.5%) made the task of stopping overall harder for our participants. In addition, in the SELF condition, we replicated the effect of Degradation in their Experiment 1 on the duration of the initial word. Similarly, in the NO condition, we found no effect of Degradation on the duration of the initial word, as in their Experiment 2.\(^\text{13}\)

Participants tended to complete the initial word more often in the OTHER than in the NO condition. This suggests that they predicted that their partner was about to speak and that this prediction interfered with the process of stopping speech. This finding supports the

\(^{13}\) Note that Hartsuiker et al. (2008) analysed both completed and interrupted responses, whereas we analysed only completed responses. They did so because the proportions were much less unbalanced in their experiments than in our experiment.
claim that predictions of others’ utterances can be computed using production mechanisms, and is therefore not consistent with the separate mechanisms account. In addition, participants tended to complete the initial word less often in the OTHER condition than in the SELF condition. This suggests that they did not activate production mechanisms to the same extent in the OTHER as in the SELF condition, and hence the study is not consistent with the shared mechanisms account. In sum, this set of findings is compatible only with the overlapping mechanisms account.

### 3.4.1 Predicting that you are about to speak

The overlapping mechanisms account states that prediction uses some, but not all of the mechanisms used for production. It therefore raises the question: What is the precise nature of prediction mechanisms? We know that they are used in language production, but what kind of mechanisms are they? Below, we sketch two related proposals about the relationship between language production mechanisms and the mechanisms used in prediction.

The first possibility is that others’ utterances are predicted via activation of linguistic representations within the language production system. If \( A \) predicts that \( B \) is going to say *chair*, she does this by going through the stages of language production (e.g., accessing semantics, syntax, phonology) that she goes through when she prepares to utter *chair* herself. Importantly, the finding that our participants completed their utterance less when they predicted that their partner was about to produce an utterance than when they were about to produce a second utterance indicates that the production system was only *partly* activated when \( A \) predicted \( B \)’s utterance. Accordingly, at some point during the process of language production, \( A \) might inhibit her production system (so that she does not actually speak).

In Section 1.4.1.1 we reviewed evidence that language production mechanisms are activated during language comprehension (e.g., Yuen, et al., 2010). Moreover, in Section
we noted how D’Ausilio et al.’s (2011) TMS study suggests that activation of the production system (up to articulation) might be involved in prediction. However, it is an empirical question whether activation always occurs at all levels, with inhibition only suppressing overt articulation (as the findings reviewed above suggest), or whether inhibition can occur at any stage during the process of language production (see Section 1.4.3). This is reminiscent of a discussion concerning the nature of inner speech, where some accounts posit fully-specified sub-phonemic features (Corley, Brocklehurst, & Moat, 2011), whereas others maintain a more abstractionist view in which inner speech is specified only up to the phonological level (Oppenheim & Dell, 2010). But it is also possible that inhibition can occur at an even earlier stage (e.g., before or during lexical selection).

The second possibility is that other speakers’ utterances are predicted using forward models, as proposed by Pickering and Garrod (2013) and in Section 1.4.3 of this thesis. If this is the case, predictions of others’ utterances could be computed without the need to activate the production implementer, as forward models can be decoupled from the implementer, as it occurs in imagining (see Chapter 2). However, forward models are routinely used in the control of language production processes that take place within the implementer. Therefore, predictions of others’ utterances computed using forward models could affect ongoing language production processes even if they do not lead to direct activation of the implementer.

To illustrate with an example, imagine a sport commentator reporting live on a soccer match. She might be providing some statistics about previous matches, when one of the players suddenly performs an amazing pass that could lead to a goal. The commentator might then issue a stop signal to the articulators (depending on various factors; e.g., how much she values fluency; cf. Seyfeddinipur et al., 2008). She also forms the intention of producing the player’s name, and starts retrieving that name from memory (using the production implementer). This process requires resources, and will therefore interfere with
other language production processes that are active in parallel, including the process of stopping speech.

But in addition, and before retrieving the player’s name, the commentator predicts that she will soon produce an utterance starting with the player’s name (using a forward model). The commentator has had to stop and reformulate before. She might have learned that it is difficult to stop speech while attempting to formulate a new utterance. Once she has predicted that she will produce the player’s name, she can make use of this prediction to remove resources from the process of stopping speech and allocate those resources to the process of retrieving the player’s name, thus performing the latter task more efficiently (see Tydgat et al., 2011). Therefore, forward-model predictions could affect how quickly the commentator stops speaking. Crucially, they could do so very rapidly, as they could affect whether resources are reallocated predictively, that is, in anticipation of a need for additional resources.

In the same way, in our SELF condition, the cued participant sends a stop signal to the articulators. The participant also intends to name the target picture, and therefore sends a command to the production implementer. At the same time, a copy of the command is sent to a forward production model that computes a prediction that a word will be produced. At this point, the participant has not completed the process of stopping and is therefore still naming the initial picture. The prediction that the target word will be produced triggers the (predictive) reallocation of resources from the process of stopping to the process of retrieving the target word, thus delaying the stopping process. In addition, the cued participant retrieves the target picture’s name using the production implementer. This process takes up resources and further delays the process of stopping speech, thus increasing the tendency to complete the initial word.

When a speaker is planning an utterance, the predictions generated by the forward model are always accompanied by the activation of representations within the production implementer, which in turn normally leads to articulation. Crucially, following Pickering and
Garrod’s (2013) account, in Section 1.4.3 we have proposed that forward-model predictions can be computed for another speaker’s upcoming utterance as well. To return to our sport commentator example, imagine a situation in which she is assisted by a partner commentator. While providing the statistics, she realizes that her partner has noticed the action. She predicts that he is about to speak. However, the commentator will not retrieve the semantics, syntax, or phonology of her partner’s utterance and will therefore not have to take resources away from the production of her current utterance. But because she predicts her partner’s utterance using the same mechanism that she would use to predict her own upcoming utterance, she might nevertheless predictively take some resources away from stopping (because this has proved effective in the past).

In the same way, in our OTHER condition, the cued participant knows that her partner intends to name the target picture, and therefore forms a representation of his production command. A copy of the command is sent to the forward production model and, as in the SELF condition, the prediction that a word will be produced triggers the reallocation of resources, away from the process of stopping. Unlike in the SELF condition, however, the participant does not use the representation of her partner’s production command to drive retrieval of linguistic representations within the implementer. Therefore, the (partner’s) naming of the target picture does not interfere with the process of stopping speech. This explains why the tendency to complete the initial word was weaker in the OTHER than in the SELF condition.

In summary, while forward production models are production mechanisms, and can affect ongoing language production, they can do so without activating the production implementer. According to this version of the overlapping mechanisms account, therefore, the production implementer is not causally involved in prediction of other people’s utterances and the activation in D’Ausilio et al. (2011) is incidental, rather than the source of the prediction.
In conclusion, our results indicate (i) that some production mechanisms are implicated in prediction; (ii) that production mechanisms are not as strongly activated when speakers predict others’ utterances as when they prepare their own utterances. The results of the present study are therefore compatible with both versions of the overlapping mechanisms account, that is with a version where predictions are computed by first activating and then inhibiting the production implementer, and a version where predictions are computed using forward models.

3.4.2 Prediction and the coordination of utterances in dialogue

What is the relevance of these results for our understanding of the coordination of utterances in natural conversations? Clearly, our task is very different from natural conversation. First, speaker A and speaker B produce two completely unrelated utterances. Second, the moment at which the speaker-switch occurs, and the direction of the switch, are fixed and determined by the experimenter. Third, the experimental conditions were the most favorable for prediction: A could see what picture B was going to name, and she knew, because of the instructions, that B was about to name it.

In addition, it is possible that participants in the OTHER condition developed a tendency to attend to the target picture even on trials on which they did not have to name it. The cued participant does the same thing in the NO and OTHER conditions, except that the cued participants names some target pictures in the OTHER condition, and these pictures of course have the same characteristics as the pictures that their partner names on the other trials. This process might enhance the activation of production processes, and in turn cause the cued participant to complete the names of the initial pictures more often in the OTHER condition than the NO condition. If so, it may be the case that people are only affected by their partner’s tendency to speak under conditions in which they sometimes have to speak themselves. A future study could test whether participants who always name either the initial
or the target pictures (i.e., they never switch roles) would predict their partner’s utterances to the same or to a smaller extent as they do in the OTHER condition.

Despite some limitations, we argue that our experiment provides evidence about mechanisms that could be used in conversation to predict other people’s utterances (at least when conditions are favorable). Moreover, many natural conversations support accurate predictions, because interlocutors can capitalize either on a long interactional history that leads to alignment (Pickering & Garrod, 2004) or on the scriptedness of the activity type (e.g., purchasing an item in a shop, fixing an appointment at the doctor; H. H. Clark, 1996; Levinson, 1992). And production processes are certainly likely to be activated in natural conversations where interlocutors switch between the roles of speaker and listener all the time.

In addition, while we have only provided evidence that speakers can predict that their partner is going to speak, it is possible that similar mechanisms underlie the ability to predict what one’s partner is going to say and when. Clearly, interlocutors would greatly benefit from the ability to predict (i) that their partner is going to speak; (ii) what their partner is going to say; (iii) when their partner is going to speak. Prediction (i) would allow them to decide whether to continue or stop speaking themselves. Prediction (ii) would help them prepare an adequate response to the current speaker’s contribution, or in completing the speaker’s utterance. Prediction (iii) would be useful for smooth turn-taking. In the next chapter we present a series of experiments designed to investigate whether speakers can predict the “what” and “when” of others’ utterances.

3.5 Acknowledgements

We thank Uschi Cop, who contributed to an equal degree as the author to the design of the experiment, the testing of participants, and the scoring of participants’ responses; she also contributed to the first stages of analysis and interpretation.
We thank Rob Hartsuiker for helpful discussions and for making the target pictures available to us, Eddie Dubourg (Edinburgh University) for invaluable technical assistance, and Michael Stevens (Ghent University) for patiently discussing various statistical issues with us.
4. Study 3. Experiments 6-8: effect of anticipating another’s utterance on time to start speaking and time spent speaking

4.1 Introduction

In this chapter, we present four experiments that investigated whether speakers can predict aspects of the “what” and “when” (i.e., of the content and timing) of other speakers’ utterances using production mechanisms. To answer this question, we used a design similar to the one employed in Study 2, that is, a consecutive production paradigm in which speaker A produces an utterance and then speaker B produces a second utterance. In this case, in addition, B’s utterance continued A’s utterance, so that the two utterances taken together would constitute a full sentence.

Crucially, we varied B’s continuation so that it would be more or less difficult (i.e., require more or less processing resources) to plan. The precise details of how difficulty was manipulated varied between experiments (see below). However, the rationale was the same in all experiments. As described in Section 1.5.2, it is known that speakers plan their utterances ahead of articulation. The extent (scope) of this advance planning is variable (Konopka, 2012). But more importantly, it is possible to measure it by looking at the extent to which speech onset latencies, articulation durations, and likelihood of disfluencies are affected by properties of (parts of) an upcoming utterance. These properties can in turn be controlled experimentally by asking participants to describe pictured scenes whose features are systematically manipulated.

In the joint task, we kept the properties of A’s utterance constant, and varied only the properties of the continuation that B was about to produce. If A anticipates aspects of the content and timing of B’s upcoming utterance using language production mechanisms, the
production of $A$’s utterance could be affected by $A$’s representation of $B$’s utterance, in a way that is similar to how the production of $A$’s utterance is affected by $A$’s representation of $A$’s own utterance when $A$ plans ahead.

Note that, in the literature on the scope of utterance planning, researchers have carefully distinguished between different stages in the language production hierarchy. According to a model of utterance planning proposed by Bock and Levelt (1994), conceptual planning generates a message (i.e., a representation of the meaning the speaker wishes to express). Then, the message goes through a process of grammatical encoding, which begins the translation of the message into a linguistic form. Grammatical encoding is itself composed of two stages. The functional processing stage involves the selection of lemmas and the assignment of syntactic functions (e.g., subject, object) to lemmas. The positional processing stage involves the construction of a sentence frame, which specifies the structural dependencies between constituents and also determines the linear order in which such constituents will appear in the utterance. Finally, phonological processing involves the retrieval of the phonological forms corresponding to the selected lemmas and creates representations that are used to control articulation.

Smith and Wheeldon (1999) investigated higher level conceptual planning and grammatical encoding. They showed that longer utterances (composed of two clauses) are initiated later than shorter utterances (composed of only one clause). In addition, they showed that utterances of comparable length are initiated later when the first phrase is complex than when it is simple (but not when the second phrase is complex than when it is simple). Overall, they concluded that grammatical encoding (including lemma selection) is completed only for the items in the first phrase of an utterance, but some higher level processing is carried out for later part of an utterance before speech onset (see also F. Ferreira, 1991).

Other studies investigated lemma selection and the retrieval of phonological forms. They did so by manipulating the accessibility of lexical items that occupy different positions
within an utterance. Some of these studies showed that the lemmas for all the nouns in simple sentences (also beyond the first phrase) are selected and accessed before speech onset, but the phonological form of only the first noun is retrieved (e.g., Meyer, 1996; M. Smith & Wheeldon, 2004). However, other findings suggest that lexical selection is restricted to only the first phrase of an utterance (e.g., Allum & Wheeldon, 2009; Griffin, 2001). For example, Allum and Wheeldon (2009) showed that when English speakers were asked to produce sentences starting with an noun phrase (NP) followed by a propositional phrase (PP; e.g., the dog above the table is green), onset latencies were sped up by preview of a picture corresponding to the noun in the first phrase (i.e., dog) but not by preview of a picture corresponding to the noun in the second phrase (i.e., table). Further, Schriefers et al. (1998) showed that distractors semantically related to the verb affected onset latencies only when the verb was in initial position in German transitive sentences (a recent review of these issues is provided by Konopka, 2012).

Here, we are not concerned with the question of how far ahead speakers can plan their utterances. However, we will take advantage of the fact that speakers do. Whenever it was possible, we ran pre-tests to verify that speakers would indeed be affected by properties of their upcoming utterances. In addition, note that we did not systematically manipulate the complexity of the continuations independently of their length. Rather, more difficult continuations were always longer than easier continuations. Nor did we directly manipulate the accessibility of particular lexical items within the continuations. Accordingly, our manipulations might have affected the difficulty of the earlier stages of planning, those involved in the conceptualization of the events depicted in the pictures and/or in grammatical encoding (lemma selection, or the creation of a sentence frame), as well as later stages of planning, including retrieval of the phonological forms corresponding to the selected lemmas. It was not our aim to distinguish between these different possibilities. Given the novelty of our design, we opted for manipulations that would guarantee as large a difference as possible in terms of planning difficulty.
4.2 Experiment 6

In this experiment, pairs of participants took turns describing scenes that consisted of two parts. The first part showed a character (e.g., *soldier*) performing a transitive action (e.g., *chase*). Let us call this part the *preamble*. The preamble always had a fixed structure and length, and it was designed to elicit carefully controlled descriptions. The participants learned the characters' name beforehand and could read the verb (corresponding to the depicted action) off the screen. The second part of each scene showed three abstract shapes, one of which was the target shape the participants had to describe. Let us call this part the *continuation*. Continuations were designed to elicit spontaneous descriptions, and the participants could choose freely both the content and the structure of their descriptions.

However, we manipulated the difficulty of the continuations in the following way. On each trial, the target shape that participants were asked to describe was accompanied by two context shapes. The participants only described the target shape, but they were told that their descriptions had to contain enough information to allow a potential (imagined) addressee to correctly select the target shape amongst the others. On easy trials, the context shapes were similar to each other, but dissimilar to the target shape. On hard trials, one of the context shapes was similar to the target shape (while the other context shape was dissimilar; see Figure 4-1).
Figure 4-1. The easy (top) and the hard (bottom) version of an item used in Experiment 6, with sample descriptions produced by a pair of participants. The parts of the picture corresponding to the preamble and the continuation, and the target shape are marked in this example for explanatory purposes, but were not in the pictures shown to participants.
We hypothesized that participants would incorporate more details in their descriptions of the target shape on hard trials, in order to better distinguish the target shape from the similar context shape, and help their (imagined) listeners. Note that it is not clear to what extent speakers spontaneously design their utterances in a way that is helpful for listeners (e.g., Brennan & Hanna, 2009; Shintel & Keysar, 2009). However, we pre-tested our materials to make sure that participants would produce longer descriptions on hard than on easy trials. Both the results of the pre-test and of the main experiment confirmed that participants did take the imagined listeners’ needs into account to some extent, and produced longer descriptions when the target shape shared more features with one of the context shapes.

Pairs of participants were assigned to one of two conditions. In the SELF condition, one of the participants described both the preamble and the continuation on any given trial, but participants took turns speaking (according to a random sequence) between trials. Importantly, on the basis of the pre-test, we expected participants in the SELF condition to spend more time articulating the preamble before producing hard than before producing easy continuations. This would indicate that participants were planning the continuation while still producing the preamble, and that the former interfered with the latter (presumably because both are production processes that make use of similar mechanisms).

Note that in the pre-test we found no evidence that the difficulty of the continuation affected speech onsets. This could reflect the fact that participants planned incrementally. However, for our purposes it was important to establish an effect of continuation difficulty on production of the preamble, but we had no specific hypotheses as to which measure of production difficulty (likelihood of disfluencies, preamble onset, or preamble duration) would show the effect.

In the OTHER condition, participants took turns speaking within each trial. One participant described the preamble, followed by the other participant who described the
continuation. Participants were seated in front of the same screen and could therefore see the same stimuli as their partners (i.e., preambles and continuations were visible to both participants at all times). Participants performed both roles, describing preambles and continuations equally often, and switching roles between trials (according to a random sequence). Again, we expected the participants who described the continuations to produce longer continuations on hard than on easy trials.

We were interested in the participants who produced the preambles. If these participants anticipate that their partner is about to describe the continuation, and, in addition, if they represent the approximate length or complexity of their partner’s utterance (based on the shared visual displays) using language production mechanisms, then they might be affected by this representation in a similar way to participants in the SELF condition. In other words, if participants in the OTHER condition represent some aspect of the content and timing of the utterance their partner is about to produce, this representation might interfere with the production of the preamble, causing them to spend more time articulating the preamble before hard than before easy continuations. This would indicate that speakers can form predictive representations of others’ utterances using language production mechanisms.

More precisely, we can distinguish between two versions of this hypothesis (see Chapter 3). If speakers form predictive representations using the same mechanisms they use when producing language (the shared mechanisms hypothesis), then we expect the effect of continuation difficulty to be the same in the SELF as in the OTHER condition (i.e., a main effect of continuation difficulty, with no interaction). However, if speakers form predictive representations using mechanisms that are partly dependent on production mechanisms, but they do not recruit the full range of production mechanisms, we would expect an interaction, as the effect of continuation difficulty might well be smaller in the OTHER than in the SELF condition. Crucially, though, we would still expect an effect of continuation difficulty in the OTHER condition.
4.2.1 Method

Participants

Twenty-four pairs of previously unacquainted participants took part in the main experiment. Twelve pairs were assigned to the SELF condition (6 female pairs, 2 male pairs, 4 mixed-gender pairs), and twelve pairs were assigned to the OTHER condition (7 female pairs, 1 male pair, 4 mixed-gender pairs). One additional pair of participants was tested in the SELF condition, but it was discarded because one of the participants reported that her partner had not been naming the target picture but the distractor pictures instead. Participants were all native speakers of British English. After the experiment, they performed two additional tasks (not related to the current study and not reported here). They were paid at the rate of £6.20/hour (with a minimum compensation of £6.00 if they took less than an hour to complete all tasks).

Materials

To construct the materials, we paired each of 20 preambles with four abstract shapes. The shapes in each set were organized in two groups of pairwise similar shapes, for a total of 40 pairs of similar shapes (see Figure 4-1 for an example and Appendix A, Table A-3, for a complete list of items). Preambles consisted of one of ten characters (ballerina, painter, pirate, nun, policeman, waitress, sailor, cook, soldier, and cowboy) combined with one of ten action verbs (push, touch, shoot, welcome, see, follow, chase, hit, punch, choose). Each character was combined with two different action verbs, and each action verb was combined with two different characters. However, we treated each unique combination of a character and an action verb as a separate item. Pictured characters were selected from a gallery of black and white line drawings available at the University of Edinburgh. The action verbs were printed in Arial 36-point bold capital letters underneath the characters. We selected 40 abstract shapes from various sources: 6 were taken from Krauss and Glucksberg (1969); 4
from the set of tangram figures used by Wilkes-Gibbs and Clark (1992); 2 from the set of tangram figures used by Schober and Clark (1989), and 12 from a publicly available database of images (http://wiki.cnbc.cmu.edu/TarrLab), courtesy of Michael J. Tarr, Center for the Neural Basis of Cognition and Department of Psychology, Carnegie Mellon University (Tarr, 2012) The remaining 16 shapes were created for this study, drawing inspiration from various non-Latin-based scripts. Each of the selected shapes was then modified to produce a visually similar shape.

Pre-test. Eight pairs of previously unacquainted participants from the University of Edinburgh student community took part in the pre-test (4 female pairs, 1 male pair, 3 mixed-gender pairs). They were all paid £6.00 for their participation. All but one participant reported to be a native speaker of English. This participant was highly proficient in English. All the analyses reported below were carried out on a dataset that excluded that participant (and her partner). Those analyses gave the same pattern of results as analyses performed on the full dataset, so only the latter are reported. The design and procedure were the same as for the SELF condition of the Experiment (see below). All recordings were transcribed verbatim by the author, who noted down disfluencies (hesitations and repetitions) and manually marked three time points for every trial: preamble onset, preamble offset, and continuation onset. First, the number of words per continuation was calculated (hesitations

The only differences between the procedure of the pre-test and the procedure of the main experiment relate to the fact that in the main experiment the instructions gave a bit more details about the imagined listeners’ task. First, it was emphasized that the (imagined) listeners would be asked to select the correct shape as quickly as possible. Second, participants were explicitly told to imagine that their descriptions would be mixed with descriptions produced by other participants, and therefore, that each imagined listener would attend to only a subset of their descriptions. The reason for providing this additional information was two-fold. First, we thought it would make the imagined addressees and their task more concrete. Second, we hoped it would discourage speakers from being over-informative on easy trials and from relying on previously established ways of referring to the target shape on hard trials (as common ground with the listeners could not be assumed). Overall, we hoped this would make the manipulation of continuation difficulty more effective. Importantly, the manipulation was successful both in the pre-test and in the main experiment (see text).
and repetitions were counted as separate words for the purpose of this analysis, as we used number of words as a measure of planning difficulty). Hard continuations contained on average more words (M= 13.8, SD= 9.3) than easy continuations (M= 11.2, SD= 10.3), and this difference was significant by participants in a paired t-test (t(15)= 2.29 , p<.05).

Second, we fitted linear mixed-effects models with random by-participant and by-preamble intercepts and random by-participant and by-preamble slopes for the factor Difficulty (i.e., maximal random effects structure) to preamble duration (i.e., preamble offset - preamble onset), after removing a few outliers based on inspection of the distribution (duration above 3 seconds, 0.69%). Preambles were longer before hard (M= 1.732 seconds, SD= .403) than before easy (M= 1.645 seconds, SD= .384) continuations (B = .083 seconds, SE = .027, t=3.04). To the contrary, difficulty had no effect on preamble onset. After removing outliers (onset above 3 seconds, 6.93%), the mean onset was 1.402 seconds (SD = .385) before easy continuations and 1.437 seconds (SD = .413) before hard continuations (|t|<1). In addition, difficulty had no effect on the number of disfluencies: Participants produced 17 disfluencies before easy continuations and also 17 disfluencies before hard continuations\textsuperscript{16}.

**Design**

The factor Difficulty (easy vs. hard) was manipulated within participants and within items. The factor Partner (OTHER vs. SELF) was manipulated within items but between participants. Four lists of experimental materials were created. In every list, the twenty preambles appeared four times each, twice with an easy, twice with a difficult shape, so that

\textsuperscript{15} This cut-off was chosen on the basis of the distribution of preamble onset times. This distribution appeared bimodal, because of a relatively large number of data points concentrated in its right tail, above 3 seconds. To avoid violating distributional assumptions, we opted for discarding these data points.

\textsuperscript{16} Note that for the purposes of this pre-test we only counted hesitations and repetitions as disfluencies. Silent pause longer than 200 ms were not counted as disfluencies as they were not transcribed, unlike in the main experiment (see Recording and Data Analysis below).
both participants produced all preambles both before an easy and before a hard continuation. There were 80 trials in total. Each of the 40 abstract shapes appeared once as an easy target and once as a hard target. However, we made sure that each participant described a given shape only once (either in the easy or the hard version). We then counterbalanced across lists whether a given shape was named in the easy or hard version first, and which participant named it, so that in total each participant named ten shapes in the easy version on their first appearance, ten shapes in the hard version on the first appearance, and the rest of the shapes in the easy or hard version on their second appearance.

On each screen, the target shape was accompanied by two context shapes, one located above and one below it. In the easy version, the context shapes were similar to each other but different from the target shape (one of them served as target on another trial). In the hard version, one of the context shapes was different from the target shape, but the other was similar to it. For each pair, only one of the two similar shapes in each pair served as target (the other only served as distractor); the identity of the target shape was counterbalanced across pairs.

**Procedure**

Participants were recruited individually and paired based on their availability to come into the laboratory for the experiment. Upon arrival, they were introduced to each other and seated in front of the same computer monitor. The experimenter tossed a coin beforehand to decide which participant would be sitting on the left of the screen. The participants were familiarized with the names of the characters in the following manner. First, the characters appeared on screen with the corresponding names, and participants were asked to read the names out loud to aid memory. Each character remained on screen for a variable amount of time, which was controlled by the participant sitting on the left. However, participants were told that they both had to read the names out loud, and that they should try and do so in unison. This was done to limit the duration of the session. Second, the
participants went through the list of characters again (in a different, randomized order), and were asked to name the characters from memory. Again, they were asked to speak at the same time as their partner. The experimenter repeated the correct name in case either participant made a mistake. This procedure was repeated once if the participants did not perform satisfactorily.

After the familiarization phase, the experimenter helped participants wear headset microphones (see Data Recording and Analysis), and then explained the instructions. Participants were told that they would describe pictured scenes containing one of the characters, the name of an action, and a vertical array with three abstract shapes. They were asked to start their descriptions by saying the name of the character and by reading the name of the action out loud (e.g., *the soldier chases*). They could then continue with a description of the target shape (which was always the shape appearing in the middle of the array). To make sure the participants would pay attention to the context shapes as well, the experimenter asked them to imagine they were producing the descriptions for participants in another experiment, and that these people would be looking at the same three shapes, but scrambled in a different order. It was then emphasized that their descriptions should contain as much detail as they thought would be necessary to allow their imagined addressees to correctly select the target shape.

Both participants then produced descriptions for one practice shape (which was in the easy condition), after which the experimenter stressed the importance of producing the descriptions fluently. The instructions were the same for pairs in the SELF and in the OTHER condition, up to this point. Pairs in the SELF condition were then instructed that each scene had to be described by only one of them: The person sitting on the left described the scene when the character was on the left-hand side of the screen, whereas the person on the right described the scene when the character was on the right-hand side of the screen. Pairs in the OTHER condition were instructed to describe the scenes together in the following way. When the character appeared on the left-hand side of the screen, the person
on the left named the character and read out the action, while the person on the right
described the middle shape; and vice versa, when the character appeared on the right-hand
side of the screen. Finally, the experimenter adjusted the sensitivity of the microphones,
started the recording, and left the room. This part of the procedure took approximately 15
minutes.

On each trial, first a fixation cross was displayed for 2 seconds. Following a 500-ms
blank, the scene was displayed and it stayed on screen until the description had been
completed and one of the participants pressed the space bar to move on to the next trial. The
ITI was 1000 ms. Participants were told that they should press the space bar as soon as they
finished describing a picture, and also that they should both be ready to start another
description on the next trial when either of them pressed the space bar. Between the two
blocks, participants were left free to take a break for however long they needed (again, they
negotiated between themselves when to start the second part of the experiment). The
duration of an experimental session was therefore variable. On average, it lasted 25 minutes,
but it ranged from just over 18 minutes for the fastest pairs to just over 44 minutes for the
slowest pair (SD = 7.2 minutes).

**Recording and data analysis**

A 75-ms beep (inaudible to the participants) was used to mark stimulus presentation
and was recorded together with the participants’ responses (on three separate channels,
sampling rate: 48000 Hz) via an M-Audio FireWire 1814 device (inMusic, Cumberland, RI,
www.m-audio.com) in Adobe Audition (Version 4.0). Participants spoke into head-mounted
SM10A microphones (Shure Distribution UK, Waltham Abbey, Essex, UK,
http://www.shure.co.uk).

Beep onsets were automatically tagged using Audacity (Version 1.2.5). Recordings
were then transcribed verbatim by the author, who also recorded the presence of disfluencies
in the preambles as well as in the continuations. Disfluencies that occurred before speech
onset or during articulation of the preamble, but no later than the verb were classified as disfluencies belonging to the preamble. Disfluencies that occurred after the verb but before the start of the continuation were classified as belonging to the continuation. This was done to ensure comparability between the SELF and OTHER condition. Only disfluencies that were classified as belonging to the preamble were analysed. They were further divided into three categories: hesitations (e.g., *uh, uhm the soldier*), repetitions (e.g., *th- the soldier*), and silent pauses (e.g. *the ... soldier*). Pauses were classified as disfluencies if they were longer than 200ms.

The author also marked four points on the audio file for every trial: Preamble Onset (the onset of the character’s name); Preamble Offset (the offset of the action word); Continuation Onset (the onset of the participant’s description of the target shape); Continuation Offset (the offset of the participant’s description of the target shape). These points were then used to derive four measures: Onset (the time elapsed from picture presentation till the participant began speaking), Duration (Preamble Onset–Preamble Offset), Gap (Continuation Onset – Preamble Offset), Continuation (Continuation Offset – Continuation Onset).

We analysed the data using Generalized Linear mixed-effects models (Baayen, et al., 2008), as implemented in the lme4 package (Bates, et al., 2008) in R (R, Version 2.13.1), with a logistic link function (Jaeger, 2008) for categorical binary data (fluent vs. disfluent), and a normal link function for timing data. In all analyses, the fixed effect structure contained two predictors initially, that is Partner (OTHER vs. SELF) and Difficulty (Easy vs. Hard), as well as their interaction. Both predictors were sum-coded. We then selected the best-fitting model using a backward stepwise procedure. We report estimates, standard errors, Wald’s z (for logistic models), or t tests (for models with a normal link function) for the best model.

Random structure was kept maximal (Barr, Levy, Scheepers, & Tily, 2013), which means that we included by-participant random slopes for Difficulty, and by-item random
slopes for Partner, Difficulty, and their interaction. Whenever the maximal random structure
did not converge, we simplified it, starting with the higher order term (the Partner-by-
Difficulty interaction), and proceeded to eliminate random slopes if necessary, first by items,
then by subjects. Estimates of the random effects are reported, as well as correlations
between random effects. For the analyses that focused on the preamble (the main analyses of
interest), each preamble counted as a different item (regardless of which of the two shapes it
was paired with). These analyses refer to the following dependent variables: the likelihood of
disfluencies occurring on the preamble, Onset, and Duration. For the analyses of Gap and
Continuation we also included by-shape random effects, nested within by-preamble random
effects.

4.2.2 Results

A total of 1920 trials were transcribed. First, the number of words (including
hesitations and repetitions) per continuation was calculated. Hard continuations contained on
average more words (M= 13.6, SD= 8.3) than easy continuations (M= 10.3, SD= 7.0), and
this difference was significant by participants in a paired t-test (t(47)=7.74 , p<.001). The
difference between hard and easy continuations was significant also when the SELF and
OTHER condition were considered separately (SELF: hard, M = 13.8, SD = 8.3; easy, M =
10.1, SD = 7.4; paired t-test: t(23)= 5.98, p<.001; OTHER: hard, M = 13.4, SD = 8.5; easy,
M = 10.5, SD = 6.8; paired t-test: t(23)= 4.66, p<.001). The number of pauses was higher for
hard continuations (M = 2.3, SD = 1.8) than for easy continuations (M = 1.4, SD = 1.2), and
this difference was significant by participants in a paired t-test (t(47)= 6.27, p<.001). This
confirms that our manipulation was successful, as hard continuations were indeed more
difficult to describe for our participants than easy continuations.

Before conducting the analyses as described in the Methods section, we discarded
trials that were problematic for a number of reasons: the participants laughed before or while
producing the preamble (0.26%); the participants were distracted when the picture appeared
on screen (because they were talking to each other; 0.63%); the participants did not describe
the target shape, or the wrong participant described it (0.21%); the participants overlapped
their speech (0.31%, all in the OTHER condition). We also discarded all trials on which the
participants used words other than the ones they were instructed to use in their descriptions
of the preambles (4.69%).

4.2.2.1 Likelihood of disfluency while describing the preamble

Preambles were coded as being either fluent or disfluent, regardless of the kind and
number of disfluencies they contained. However, it is clear from Table 4-1 that most
disfluent preambles contained only silent pauses. First, participants in the SELF condition
produced more disfluencies (173) than participants in the OTHER condition (46). Second,
there was little difference in the number of disfluencies produced before hard than before
easy continuations. This was the case in both the SELF and the OTHER condition. These
patterns were confirmed statistically: the best model (see Table B-15 in Appendix B)
included only a main effect of Partner ($\chi^2 (1) = 11.98, p<.001$).
Table 4-1.

*Disfluency counts in Experiment 6, by Difficulty and Partner. The total counts are broken down according to disfluency type (hesitation, repetition, pause; see Recording and Data Analysis).*

<table>
<thead>
<tr>
<th>Disfluency Type</th>
<th>SELF</th>
<th>OTHER</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Easy</td>
<td>Hard</td>
</tr>
<tr>
<td>P</td>
<td>67</td>
<td>79</td>
</tr>
<tr>
<td>H</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>R</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>P+H</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>P+R</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>P+R+H</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>R+H</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Total (except P)</td>
<td>15</td>
<td>12</td>
</tr>
<tr>
<td>Total</td>
<td>82</td>
<td>91</td>
</tr>
</tbody>
</table>

*Note.* This table shows counts of disfluencies that belonged to the preamble only (as specified in Recording and Data analysis). Preambles that contained hesitations or repetitions often contained silent pauses as well. All the observed combinations of more than one kind of disfluency are listed separately as different types in the table, but for the purposes of statistical analyses (see text) each preamble was classified as either fluent or disfluent, irrespective of which type it belonged to. Multiple occurrences of the same kind of disfluency in the same preamble did occur, but they are not listed separately. P = silent pause longer than 200 ms; H = hesitation, R = repetition.
4.2.2.2 Timing analyses

For these analyses, we excluded all trials that contained disfluencies of any kind before or on the preamble.

Onset

Values longer than 3 seconds were considered outliers and removed from further analyses (6.38%).\textsuperscript{17} First, on average participants in the SELF condition took 180 ms longer to start speaking than participants in the OTHER condition (see Table 4-2). Second, participants in the SELF and OTHER condition took no longer to start speaking before hard than before easy continuations. The best model included only the main effect of Partner ($\chi^2(1) = 8.56, p<.005$). This confirms the results of the pre-test, where we also found that Difficulty did not affect Onset (see Table B-16 in Appendix B).

Table 4-2.

Mean Onset (seconds) and standard deviation of Onset by Difficulty and Partner in Experiment 6.

<table>
<thead>
<tr>
<th>Difficulty</th>
<th>Partner</th>
<th>SELF</th>
<th>OTHER</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy</td>
<td>SELF</td>
<td>1.261 (.361)</td>
<td>1.078 (.367)</td>
<td>1.158 (.376)</td>
</tr>
<tr>
<td></td>
<td>OTHER</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hard</td>
<td>SELF</td>
<td>1.248 (.384)</td>
<td>1.069 (.315)</td>
<td>1.147 (.357)</td>
</tr>
<tr>
<td></td>
<td>OTHER</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>SELF</td>
<td>1.254 (.372)</td>
<td>1.074 (.342)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>OTHER</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\textsuperscript{17} As in the pre-test, the distribution of preamble onset times was bimodal, as a relatively large number of observations concentrated above 3 seconds. We chose to remove these observations to avoid violating distributional assumptions.
**Duration**

Values longer than 3 seconds were considered outliers and removed from further analyses (0.06%). First, preamble durations were on average 200 ms longer in the SELF than in the OTHER condition. Second, preamble durations before hard continuations were 19 ms longer than preamble durations before easy continuations in the SELF condition, but 11 ms shorter in the OTHER condition (see Table 4-3). However, the best model (see Table B-17 in Appendix B) included only the main effect of Partner ($\chi^2(1) = 18.78, p<.001$). Neither the main effect of Difficulty nor the interaction of Difficulty and Partner reached significance. Therefore, the finding that preamble durations are longer before hard than before easy continuations (see pre-test) was not replicated in the SELF condition.

Table 4-3.

*Mean Duration (seconds) and standard deviation of Onset by Difficulty and Partner in Experiment 6.*

<table>
<thead>
<tr>
<th></th>
<th>SELF</th>
<th>OTHER</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Difficulty</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Easy</td>
<td>1.453 (.344)</td>
<td>1.268 (.232)</td>
<td>1.353 (.303)</td>
</tr>
<tr>
<td>Hard</td>
<td>1.472 (.329)</td>
<td>1.257 (.223)</td>
<td>1.354 (.296)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>1.462 (.337)</td>
<td>1.262 (.227)</td>
<td></td>
</tr>
</tbody>
</table>

The lack of an effect of Difficulty in the SELF condition is surprising given the magnitude of the effect in the pre-test. One difference between the pre-test and the SELF condition of the main experiment is that the sample of participants tested in the main experiment appeared to be overall faster (for example, mean onset before easy continuations was 1.402 seconds in the pre-test, but only 1.261 seconds in the SELF condition of the main experiment). To explore whether this difference could account for the lack of an overall
effect in the main experiment, we divided the 24 participants in the SELF condition into two groups, based on a median split of their mean Onset (median Onset = 1.329 seconds). Even when a model with Difficulty as the only predictor was fitted to the subset of participants whose mean Onset was above the median, no effect of Difficulty was found (Easy: M = 1.404 seconds, SD = .361; Hard: M = 1.466 seconds, SD = .373; t<1.6).18

Gap

Values longer than 5 seconds were considered outliers and removed from further analyses (1.14%). Inspection of the distribution revealed that Gap was non-normally distributed. To correct for this, we applied Box-Cox transformations (Box & Cox, 1964) using the package geoR (Diggle & Ribeiro Jr., 2007, 2011).19 First, the gap was much longer when it occurred between people (in the OTHER condition), than when it occurred within people (in the SELF condition). Second, gaps were 228 ms longer before easy than before hard continuations in the SELF condition and 230 ms longer before easy than before hard continuations in the OTHER condition (see Table 4). Accordingly, the best fitting model (see Table B-18 in Appendix B) included the main effects of Partner ($\chi^2 (1) = 25.81, p<.001$) and Difficulty ($\chi^2 (1) = 16.49, p<.001$), but no interaction. In combination with the lack of an effect of Difficulty on Onset and Duration, this indicates that the participant who described the continuation was affected by the difficulty of the continuation only after having uttered the preamble. This was the case in both the SELF and the OTHER conditions.

Another difference between the SELF condition and the pre-test relates to the fact that analyses of the pre-test included preambles that contained silent pauses. It is possible that the effect of Difficulty on the duration of the preamble in the pre-test was entirely driven by these silent pauses. In addition, in the SELF condition of the main experiment participants produced numerically more silent pauses before hard than before easy continuations (see Table 1). Therefore, we ran this analysis again on a dataset that included preambles with silent pauses as well as fluent preambles. This additional analysis, however, yielded exactly the same pattern of results; that is, there was not a main effect of Difficulty, or an interaction between Difficulty and Partner.

We thank Ian Finlayson for advice on this matter and for providing us with some custom-made help functions.
Mean Duration (seconds) and standard deviation of Gap by Difficulty and Partner in Experiment 6.

<table>
<thead>
<tr>
<th>Difficulty</th>
<th>SELF</th>
<th>OTHER</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy</td>
<td>.493 (.558)</td>
<td>1.131 (.806)</td>
<td>.833 (.770)</td>
</tr>
<tr>
<td>Hard</td>
<td>.721 (.854)</td>
<td>1.361 (.915)</td>
<td>1.069 (.942)</td>
</tr>
<tr>
<td>Total</td>
<td>.605 (.728)</td>
<td>1.247 (.869)</td>
<td></td>
</tr>
</tbody>
</table>

Continuation

Values longer than 30 seconds were considered outliers and removed from further analyses (0.13%). Participants produced longer descriptions for hard than for easy continuations, which further confirms that our manipulation was effective (see Table 4-5). Continuations were numerically longer in the SELF than the OTHER condition. However, the best model (see Table B-19 in Appendix B) contained only the main effect of Difficulty ($\chi^2 (1) = 25.46, p<.001$).

Mean Duration (seconds) and standard deviation of Continuation by Difficulty and Partner in Experiment 6.

<table>
<thead>
<tr>
<th>Difficulty</th>
<th>SELF</th>
<th>OTHER</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy</td>
<td>4.666(4.821)</td>
<td>4.253(3.615)</td>
<td>4.443(4.215)</td>
</tr>
<tr>
<td>Hard</td>
<td>6.292(4.860)</td>
<td>5.785(5.058)</td>
<td>6.015(4.973)</td>
</tr>
<tr>
<td>Total</td>
<td>5.468(4.905)</td>
<td>5.019(4.460)</td>
<td></td>
</tr>
</tbody>
</table>
4.2.3 Discussion

We asked speakers to produce a preamble and a continuation in one of two conditions. In the SELF condition, a speaker produced the preamble and then the same speaker produced the continuation. In the OTHER condition, a speaker produced the preamble and then the other speaker produced the continuation. We compared the SELF to the OTHER condition to investigate to what extent representing what another speaker is about to say is similar to planning one’s own utterance ahead.

In a pre-test, we showed that speakers who have to produce both the preamble and the continuation (as in the SELF condition), take longer to articulate the preamble before hard than before easy continuations. We interpreted this finding as suggesting that the participants began planning the continuation while still articulating the preamble. More specifically, the finding suggests that they were already planning enough of the continuation to be slowed down more when the continuation was harder than when it was easier. Since we did not control the type of descriptions that participants produced for the target shapes, we do not know whether hard continuations were structurally more complex than easy continuations.

However, the descriptions produced by participants were significantly longer (i.e., they contained more words and/or disfluencies) in the case of hard continuations than in the case of easy continuations. This confirms that target shapes in hard continuations were indeed more difficult to describe than target shapes in easy continuations. Therefore, we argue that participants in the pre-test had to invest more resources in planning hard than easy continuations while still articulating the preamble. Because the amount of resources available for production processes is limited, they took longer to articulate the preamble in the former than in the latter case.

Given this finding, we were interested in testing whether a speaker who only produces the preamble would nevertheless invest resources in representing the continuation that her partner was about to produce. If so, we expected the speaker to take longer to
articulate the preamble before a hard than before an easy continuation in the OTHER condition, though we anticipated this effect might be smaller than in the SELF condition. Unfortunately, we could not replicate the effect of Difficulty on the duration of the preamble in the SELF condition. This meant that the comparison between the SELF and OTHER condition was no longer informative with respect to our original hypothesis.

It is not clear why the results of the pre-test could not be replicated. One possibility is that this relates to the high variability of planning scope. This issue has been recently discussed by Konopka (2012). In her study, speakers retrieved lexical items further ahead of articulation when they produced a sentence whose syntactic structure had been primed, suggesting that planning scope might be related to ease of retrieval and to the amount of processing resources available on a trial-by-trial basis. Moreover, she reported that fast and slow speakers (defined in terms of a median split of their mean utterance onset) differed in planning scope, with slower speakers showing a larger planning scope than faster speakers.

In our case, there was no indication that the Difficulty manipulation affected slow speakers more than fast speakers (see Results, Duration). However, other (uncontrolled) individual differences (e.g., processing capacity, or working memory) might be responsible for the discrepancy in results between the pre-test and the main experiment. Such differences might have been particularly influential in our task as we did not control for the actual length and complexity of the continuations, which meant that participants with lower processing capacity, for example, might have produced overall shorter utterances.

In both the SELF and the OTHER conditions, effects of Difficulty emerged only on the length of the Gap, so only after articulation of the preamble was completed. This suggests that, by this time, speakers in both conditions were not planning enough of the continuation to be slowed down more before a hard than before an easy continuation. However, this does not necessarily mean that speakers were not planning the continuation at all at this stage. In fact, one possibility is that they were, but that they were not planning ahead enough to be differentially affected by hard vs. easy continuations.
Nevertheless, we did find a very clear difference between the SELF and the OTHER conditions. Speakers in the OTHER condition were faster to start speaking (main effect of Partner on Onset) and also faster to articulate the preamble (main effect of Partner on Duration) than speakers in the SELF condition. As a result, gaps were much longer in the OTHER than in the SELF condition\textsuperscript{20}. There are two alternative explanations for this finding. First, it is possible that speakers in the SELF condition invested more resources in planning the continuation than did speakers in the OTHER condition. In fact, speakers who produced the preamble in the OTHER condition might not have planned the continuation at all, given that there was no need for them to describe the target shape on that particular trial. This would therefore constitute evidence against the hypothesis that speakers represent what other speakers are about to say using language production mechanisms.

However, this finding could be interesting in light of the debate on the scope of utterance planning (see Section 4.1). Note that our results cannot inform the debate over the scope of phonological planning, as we did not control for properties of the phonological forms produced by speakers. Similarly, we did not control for properties of the lemmas selected by speakers, so our results cannot inform the debate over how many lexical items are selected in advance of articulation.

However, we can provide evidence that bears on the scope of grammatical encoding. It is possible that longer Onsets and Durations for preambles in the SELF than in the OTHER condition reflect interference from concurrent grammatical encoding of the constituent(s) making up the description of the target shape. While speakers in the OTHER condition planned only up to the verb, speakers in the SELF condition began planning of the continuation before speech onset, and continued during articulation of the subject noun

\textsuperscript{20} This did not have to be the case. Speakers who produced the continuation in the OTHER condition could also have been faster to start speaking. Importantly, they did not need to produce the preambles, which could have allowed them to focus on planning the continuation from the very beginning of a trial. Despite this, participants in the OTHER condition began their continuation on average at the same time as participants in the SELF condition did (as measured from scene onset; values over 9 seconds removed, 0.95%). SELF: $M = 3.690$ second, $SD = 1.420$; OTHER: $M = 3.705$ seconds, $SD = 1.146$. 

200
phrase and the verb. However, planning of the continuation must have been rather shallow at this stage in the SELF condition, as we found no effect of Difficulty on the preambles. Perhaps, speakers’ advance planning was limited to preparing to describe a transitive event (as opposed to an intransitive event in the OTHER condition), and to the selection of an appropriate sentence frame. This suggests that subcategorization frames for verbs can sometimes be selected before speech onset.

However, there is also an alternative explanation for the differences between the SELF and OTHER conditions. It is possible that speakers in the OTHER condition sped up, compared to speakers in the SELF condition, rather than the latter being slowed down by the additional conceptual planning load. Why would have speakers in OTHER sped up? Perhaps to leave their partners the floor and make sure they can avoid overlapping with them. Interestingly, Vesper et al. (2011) showed that, in a joint Simon task, participants who responded before a partner (with the instruction to respond one after the other as quickly as possible) tended to respond faster than participants who were not followed by a partner. It is possible that a similar tendency took place in our experiment, but we cannot be sure as we did not test a condition in which one speaker produced the preamble, but no speaker produced the completion. Therefore, we ran another experiment to test whether speakers who speak before another tend to speed up to leave the other speaker the floor.

4.3 Experiment 7

The present experiment was designed to investigate whether speakers spontaneously speed up the production of their utterance when they know that another speaker is about to continue their utterance. As mentioned in the Discussion to Experiment 6, Vesper et al. (2011) showed that when participants are asked to coordinate consecutive actions so as to

---

21 This experiment was designed by the author for a third-year Honours Group Project. Data were collected by a group of third-year Psychology students at the University of Edinburgh, working under the supervision of Prof. Martin Pickering and the author. The data were analysed by the author.
leave as little gap as possible between one another’s actions, the participant who acts first (i.e., the leader) tends to speed up her actions.

In particular, they compared the speed of the leader’s responses when she acted on her own (individual task) to the speed of her responses when another participant (the follower) was instructed to act immediately after the leader (joint task). In the joint task, the leader was told that the follower’s action should be performed as quickly as possible after the leader’s action. Participants took turns, so that they acted as leaders on some trials and as followers on other trials. The joint and the individual task, however, were tested in different blocks. The actions were button presses, as the participants performed a spatial compatibility (Simon) task (Vesper et al., 2011; Experiment 3).

Vesper et al. (2011) showed that leaders performed their actions faster in the joint than in the individual task. In addition, the variability in leaders’ response times was lower in the joint than in individual condition. In fact, response times and variability were inversely correlated. Importantly, faster response times correlated with shorter mean asynchronies between the leader’s and the follower’s responses (only positive asynchronies were included in this analysis), and this correlation was mediated by the reduction in variability associated with faster responses. Therefore, the authors concluded that leaders sped up their actions to make them as predictable as possible for followers, hence facilitating coordination.

In this experiment, we investigated whether speakers would similarly speed up production of an utterance when they knew that another speaker would continue their utterance. We asked participants to describe pictured scenes (see Figure 4-2). Scenes comprised two characters involved in a transitive event (e.g., a cook tickling a painter); one of the characters was coloured in red. Participants were tested in pairs and sat next to each other. On each trial, one participant was assigned the role of leader. This participant was instructed to name the red character. The other participant (follower) was assigned a different task depending on condition. In the individual condition (NO), the follower was instructed to remain silent. In the joint condition (OTHER), the follower had to complete the
description of the scene. This meant that the follower produced an action word and the name of the other character.

In addition, we varied the difficulty of the follower’s task. On some trials the leader named the agent of the transitive event (e.g., *the cook*), so that the follower continued with an active verb and an object noun phrase (e.g., *tickles the painter*). On other trials, the leader named the patient of the transitive event (e.g., *the painter*), so that the follower continued with a passive verb and a prepositional (by-) phrase (e.g., *is tickled by the cook*). There is some evidence that the time it takes to initiate a passive sentence is longer than the time it takes to initiate an active sentence, although such evidence comes from sentence recall rather than picture naming tasks (F. Ferreira, 1994; Tsiamtsiouris & Smith Cairns, 2013). Here, we were interested in whether onset times of leaders’ utterances would be affected by the difficulty of the continuations followers were about to produce.

Because of technical limitations, we recorded both participants’ utterances on the same channel. This meant that we could not measure the asynchrony between participants’ responses, as it was impossible to quantify overlap between the participants’ responses precisely. For the same reason, we measured only the onset of the leaders’ utterances (but not their duration). We expected onsets to be faster in the OTHER than in the NO condition. We also expected the variability of onsets to be lower in the OTHER than the NO condition.
Figure 4-2.

Schematic illustration of the design and procedure used in Experiment 7; an item in the active and passive versions is included as an example. In the active version, the cook is in red; in the passive version, the painter is in red.
4.3.1 Method

Participants

Twelve pairs of English-speaking participants (6 female pairs, 6 mixed-gender pairs) volunteered to take part in the experiment. They were recruited by third-year Psychology students at the University of Edinburgh among the students’ acquaintances. Participants in a pair were recruited together and were acquainted with one another. Ten participants were non-native speakers of English.

Materials

The materials used in this experiment were 32 pictures of transitive events (e.g., a cook tickling a painter, see Figure 4-2 and Appendix A, Table A-4). Action names (e.g., tickle) were printed in Times New Roman 18 points bold capital letters, centred underneath the pictures. We prepared two versions of each picture. In one version, the character on the left was coloured in red, the one on the right in grey. In the other version, the character on the left was coloured in grey, the one on the right in red. Also, half of pictures showed the agent of the action (e.g., the cook) on the left, the other half showed the agent on the right. There were sixteen different characters; each character appeared four times across the 32 pictures, twice as agent (once on the right, once on the left) and twice as patient (once on the right and once on the left).

Design

We varied two factors: Type (active vs. passive) and Partner (NO vs. OTHER). Both were varied within-items and within-participants. Note that there were sixteen items in the experiment (as each character counted as an item). The factor Type refers to whether the character named by the leader was the agent (active) or patient (passive) of the depicted transitive event. This was varied by colouring in red either the agent or the patient on the
picture (see Materials). The factor Partner refers to whether the follower remained silent (NO) or continued the description of the event (OTHER). This was varied by assigning participants different instructions. On each trial, before the picture appeared, participants were shown an instruction screen (see Figure 4-2). The instruction screen displayed the participants’ names, followed by the word red, grey, or no. The participant assigned to red was the leader; she named the character in red. The participant assigned to grey or no was the follower. If she was assigned to grey (OTHER), she continued the leader’s utterance with a verb (in the active or passive form) and the name of the grey character. If she was assigned to no (NO), she remained silent. Both participants performed both roles equally often (on different trials). Therefore, there were four possible instruction screen (A: red, B: grey; A: red, B: no; A: grey, B: red; A: no, B: red). The order in which the names were displayed (A in the top half vs. A in the bottom half of the screen) was counterbalanced across pictures.

Each of the 32 pictures was shown a total of 8 times throughout the experiment (2 Type X 2 Partner X 2 Participants), so that each participant responded to each picture four times in total. Repetitions of the same picture were distributed evenly across four blocks of 64 trials each. We made sure that the two repetitions of a picture in each block were of different types (i.e., one was active, the other passive). The pairing of pictures with instruction screens was rotated across blocks, so that each picture appeared with all instruction screens, but never twice with the same instruction screen within one block. The order in which blocks were presented was counterbalanced across pairs (with four possible orders: ABCD, BCDA, CDAB, DABC). Finally, the order of presentation within each block was randomized separately for each pair.

**Procedure**

Participants were tested in pairs. There were six different experimenters in this experiment. They each tested 2 pairs. All experimenters followed the same procedure. First, a coin was flipped to decide which participant would sit on the left of the screen. Then, the
participants were familiarized with the character names. This was done in the same way as in Experiment 6. At this point, the experimenter positioned and tested the microphones (see Recording and Data Analysis). Then, the experimenter explained the instructions, which were also presented on the computer screen in written form. Participants were informed that they would describe pictures displaying a character with red lines and a character with grey lines.

They were instructed that the person whose name appeared next to the word red had to start the description with the name of the red character. If the other person’s name appeared next to the word grey, then they had to complete the description of the event depicted by the picture (an example was provided); if the other person’s name appeared next to the word no, then they should not describe anything. The instructions emphasized speed as well as clarity. The participants then completed a practice run, which was identical to the experiment (except that only 12 trials were displayed). The experimenter stayed in the room while the participants practiced the task and provided feedback and/or clarified the instructions at the end. In addition, the experimenter made it clear that participants should minimize movement while their partner was speaking (to prevent noise from reducing the quality of the recordings). The experimenter then typed in the participants’ names, always starting with the name of the participant sitting on the left.

At this point, participants started the experimental session. These consisted of 256 trials, divided into 4 blocks of 64 trials each, with breaks in-between each block. The experimenter was not present during the session, but he or she entered the room during the breaks to check that the microphones were still correctly positioned in front of the participants. The structure of each trial was as follows: a fixation cross was displayed for 1000 ms, then the instruction screen was displayed for 2000 ms, followed by a blank screen (500 ms), and the picture for 1500 ms. The ITI was 2000 ms. The experimental session lasted around 30 minutes (excluding breaks).
Recording and data analyses

The experiment was implemented in E-Prime (Version 2.0). Leaders’ responses were recorded using two microphones. A USB 980186-0403 microphone (Logitech UK Limited, Berkshire, UK, http://www.logitech.com) was used to record what the participant said onto a .wav file, time-locked to picture onset, which was automatically stored to the computer’s hard disk. The recording timed out 1700 ms after picture onset. An ATR 20 Cardioid low impedance microphone (Audio Technica Limited UK, Leeds, UK, http://eu.audio-technica.com) was connected to the microphone socket of a Serial Response Box (Psychology Software Tools, Sharpsburg, PA, www.pstnet.com), and triggered when the participant began speaking. Because we wanted to collect data from both participants, both microphones were positioned at a roughly equal distance from both participants. For the same reason, participants were seated as close to one another as possible and slightly at an angle, so that their mouths pointed towards the microphones. However, for most pairs the voice key failed to trigger on a high proportion of trial (28% on average). Also, inspection of the recordings confirmed that the voice key was often triggered by the follower and not by the leader. Therefore, we coded onset times manually. This was done by the author using Check Vocal (Protopapas, 2007).

Accuracy and onset times were analysed using linear mixed effects models (Baayen, et al., 2008; Bates, et al., 2008; Jaeger, 2008), as in Experiment 6. The analyses included three predictors: Type, Partner, and Language competence (non-native vs. native speaker of English). All predictors were sum-coded. The random effects structure was kept maximal whenever possible, and simplified only to aid convergence. Criteria for simplification of the random effects structure and for the selection of fixed effects were the same as in Experiment 6. To analyse the variability of leader’s response, we computed the standard deviation of onsets in each of the four cells of the design. Average standard deviations were computed for each participant and item. These data where then entered into a three-way
ANOVA, with Type and Partner as within-participants and within-items factors and Language Competence as a between-participants but within-items factor.

4.3.2 Results

We discarded trials on which the leader did not respond (1.37%). The remaining trials were coded as correct if the participant who produced the subject noun phrase did so fluently (i.e., without hesitating or stuttering) and using the correct name for the character (e.g., calling a cook “cook” and not “chef”). We did not take into account the accuracy of the follower’s utterance.

Accuracy

Overall, 7.03% of the responses were coded as incorrect according to the criteria listed above. Non-native speakers produced more incorrect responses than native speakers (see Table 4-6). Accordingly, the best model (Table B-20, Appendix B) included a main effect of Language Competence ($\chi^2 (1) = 4.17, p<.05$). In addition, the effect of Type contributed marginally to model fit ($\chi^2 (1) = 3.38, p = .07$), suggesting that participants tended to experience more difficulty when naming patients than when naming agents. This tendency, however, was not modulated by their partner’s task.
Table 4-6.

Counts of errors (and percentages) by Partner and Type in Experiment 7.

<table>
<thead>
<tr>
<th>Type</th>
<th>Partner</th>
<th>NO</th>
<th>OTHER</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active</td>
<td>NO</td>
<td>45 (5.91%)</td>
<td>57 (7.53%)</td>
<td>102 (6.71%)</td>
</tr>
<tr>
<td></td>
<td>OTHER</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Passive</td>
<td>NO</td>
<td>57 (7.53%)</td>
<td>56 (7.43%)</td>
<td>113 (7.48%)</td>
</tr>
<tr>
<td></td>
<td>OTHER</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>NO</td>
<td>102 (6.71%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>OTHER</td>
<td></td>
<td>113 (7.48%)</td>
<td></td>
</tr>
</tbody>
</table>

Language Competence

<table>
<thead>
<tr>
<th>Competence</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Native</td>
<td>97 (5.51%)</td>
</tr>
<tr>
<td>Non-native</td>
<td>118 (9.28%)</td>
</tr>
</tbody>
</table>

Onset

We removed onset times shorter than 300 ms (0.04%) and longer than 3000 ms (0.00%). Importantly, leaders were faster (25 ms) to name a character when they knew that the follower would continue the description of the event than when they knew the follower would remain silent. This was the case regardless of whether the character was the agent or the patient of the event (see Table 4-7). The best model (see Table B-21, Appendix B) included only the main effect of Partner ($\chi^2 (1) = 5.25, p< .05$). Difficulty did not reliably affect leaders’ utterances, although there was a numerical trend towards longer onsets for patients than agents.
Table 4-7.

Mean onsets (and standard deviations) in milliseconds by Partner and Type in Experiment 7.

<table>
<thead>
<tr>
<th>Partner</th>
<th>Type</th>
<th>NO</th>
<th>OTHER</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active</td>
<td></td>
<td>799 (202)</td>
<td>772 (197)</td>
<td>785 (200)</td>
</tr>
<tr>
<td>Passive</td>
<td></td>
<td>810 (229)</td>
<td>787 (220)</td>
<td>799 (224)</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>804 (216)</td>
<td>779 (209)</td>
<td></td>
</tr>
</tbody>
</table>

**Variability**

As can be seen from Table 4-7, the onset of leaders’ utterances was on average slightly less variable in the OTHER than in the NO condition. However, the ANOVA revealed no main effect of Partner on average standard deviations, neither by participants (F(1, 22) = .58) nor by items (F(1, 15) = .41, both p’s>.5).

**4.3.3 Discussion**

We investigated whether speakers spontaneously speed up the production of an utterances when they know that another speaker is about to continue their utterance. We found that speakers were indeed faster to start speaking in the OTHER condition than in the NO condition, that is, they were faster when they knew their partner would speak after them than when they knew their partner would remain silent. This replicates Vesper et al’s (2011) finding for verbal (as opposed to manual) responses. In addition, it suggests that participants in the OTHER condition of Experiment 6 might have applied a similar strategy, and that this might explain why onset times and durations for the preambles were shorter in the OTHER compared to the SELF condition.

Note that participants in the present experiment sped up only by 25 ms, whereas participants in the OTHER condition of Experiment 6 were about 180 ms faster than
participants in the SELF condition of the same experiment. A direct comparison between the two experiments is not possible. First, the manipulation was between-participants in Experiment 6 whereas we used a within-participant manipulation in the present experiment. In addition, onset times were overall around 300 ms faster in the present experiment than in Experiment 6, presumably because participants in the present experiment were preparing to produce only a single lexical item. It is possible that speakers in the present experiment did not speed up their responses more because of a floor effect. Alternatively, at least part of the difference between the SELF and OTHER conditions in Experiment 6 could have been caused by additional planning load in the SELF condition, as we discussed earlier (see Section 4.2.3).

In any case, faster onsets in the OTHER than the NO condition of the present experiment indicate that speakers spontaneously sped up when they knew that their partner would speak after them. Importantly, this finding cannot be explained in terms of social facilitation (e.g., Klauer, et al., 2008), as the partner was present in both the OTHER and the NO condition. Then, why did participants speed up in the OTHER condition? One possibility is that, as argued by Vesper et al. (2011), they did so to reduce variability in their responses, and facilitate their partner in coordinating with them. However, we found no evidence that participants did in fact reduce the variability of their responses in this experiment.

The lack of reduction in variability might reflect the fact that a substantial amount of variability in naming responses is related to properties of the response itself (e.g., frequency), and it might therefore not be fully under the speaker’s control. In other words, it might be more difficult for speakers to make themselves predictable than for actors who perform simple button presses. However, speakers in our experiment might have facilitated coordination in a simpler way.

We propose that they sped up to avoid overlapping with their partner. In other words, they left their partner the floor as quickly as possible. This shows that speakers are sensitive to the fact that another speaker is about to speak after them. But this does not
necessarily imply that they sped up intentionally, or strategically, or because they applied turn-taking conventions that normally operate in conversations. In fact, it is possible that the onset of their partner’s utterance functioned as a timeout. It is known that speaker can speed up production considerably when confronted with explicit response deadlines (F. Ferreira & Swets, 2002). Our experiment extends this finding by showing that speakers also speed up spontaneously when the deadline is set implicitly by the utterance produced by another speaker.

Finally, onset times and accuracy were not affected by the type of entity the speaker named, that is, an agent or a patient. In other words, speakers were not affected by the difficulty of the utterance their partner was about to produce (i.e., active or passive). This might suggest that speakers do not represent what other speakers are about to say using language production mechanisms. However, note that leaders did not need to extract conceptual information from the scenes, as they could simply name the red character. This aspect of the task might have discouraged them from processing the whole event, and instead focussed their attention on the coloured character. In Experiment 8, we address this limitation and provide yet another test of the hypothesis that speakers can represent aspects of the content of another speaker’s utterances using production mechanisms.

In addition, in Experiment 7 we tested both native and non-native English speakers. Despite the lack of interactions between Partner and Language competence, the possibility remains that non-native speakers behaved somewhat differently from native speakers in this experiment. Differences between groups might not have reached statistical significance because of the relatively small sample size. Non-native speakers’ ability to control the timing of their utterances might be limited compared to native speakers. This could mean that they were less likely to be affected by their partners’ utterances, and perhaps particularly by predictions of the content of those utterances. This concern was another reason for running Experiment 8, in which we tested only native speakers of English.
4.4 Experiment 8

In this experiment, we tested the hypothesis that speakers represent what other speakers are about to say, and that they do so in a way that is similar to how they represent their own upcoming utterances. Once again, we asked pairs of participants to describe pictured transitive events. Participants described the events using active sentences. Therefore, their descriptions always started with the name of the character that performed the role of agent in the depicted event, followed by a verb describing the action performed by the agent (e.g., The soldier chases...). Let us call this the description preamble, as in Experiment 6.

We then varied the difficulty (length) of the continuation in the following way (see Figure 4-3). Easy continuations contained only one entity: this was always a character who served the role of patient of the depicted action. Therefore, on easy trials the continuation could be described with a simple noun phrase (e.g., …the swimmer). Hard continuations contained the patient character, and two additional entities (either two objects or an object and a character, see Methods). For example, participants could see the patient carrying two objects, in which case they were required to describe the continuation with a noun phrase followed by a prepositional phrase (which itself comprised of two conjoined noun phrases; e.g., …the swimmer with the cane and the vase). Unlike in Experiment 7, participants had to identify the agent of the event to know how to start their descriptions; therefore, they always had to process the whole event conceptually.

---

22 This experiment was designed, participants were tested, and data were analysed by the author in collaboration with Joris van de Cavey, during an internship period he spent at the Department of Psychology while studying towards a Master in Experimental Psychology at Ghent University, Belgium.
Schematic representation of the design and procedure of Experiments 8a and 8b. An item in the easy and hard versions is displayed as an example. Instruction screens with all possible combinations of assignment of instructions to participants (here, Mary and John) are displayed below the pictures. For each instruction screen on the left-hand side, the right-hand side shows the utterances the participants were expected to produce (the part within brackets was produced only when the participants saw the hard version of the example picture; the rest was common to the easy and hard version).
In a pre-test (see Methods), we found that speakers who described the preamble and the continuation spent more time producing the preamble before hard than before easy continuations. (However, they did not take longer to start speaking in the former than in the latter case.) This suggests that speakers were already planning the continuation while still producing the preamble. Specifically, given that continuations began with the same lexical item in the easy as in the hard condition, speakers were planning beyond the first lexical item of hard continuations. As discussed in Section 4.1, we make no assumption regarding the stage at which this additional planning occurred.

In this experiment, we tested three conditions (Partner, varied within participants). In the SELF condition, participants took turns describing preambles and continuations (as in the analogous condition of Experiment 6). In the OTHER condition, one participant was cued to describe the preamble, and then the other participant provided a description of the continuation (as in the analogous conditions of Experiments 6 and 7). In the NO condition, finally, participants took turns describing the preambles, but they ignored the continuations (as in the analogous condition of Experiment 7).

On the basis of our pre-test, we expected participants in the SELF condition to spend more time producing the preamble before hard than before easy conditions. We did not expect continuation Difficulty to have an effect on the duration of the preamble in the NO condition, as participants did not need to plan the continuation in this condition. We were most interested in the OTHER condition. If speakers represent what others are about to say using language production mechanisms (i.e., mechanisms they use to plan their own upcoming utterances), then participants in the OTHER condition should spend more time producing preambles before hard than before easy continuations, in a similar way to participants in the SELF condition.

More precisely, we can distinguish between two versions of this hypothesis (see Experiment 6 and Chapter 3). If speakers represent others’ utterance using the same mechanisms they use when producing language (the shared mechanisms hypothesis), then
we expect the effect of continuation difficulty to be the same in the SELF as in the OTHER condition. However, if speakers do not recruit the full range of production mechanisms when they represent what others are about to say (the overlapping mechanisms hypothesis), the effect of continuation difficulty might be smaller in the OTHER than in the SELF condition. Crucially, though, we would still expect an effect of continuation difficulty in the OTHER condition. Finally, if speakers do not represent others’ utterances, or do not represent others’ utterances using mechanisms that are used in production (the separate mechanisms hypothesis), there should be no effect of continuation difficulty in the OTHER condition (which should be equivalent in this respect to the NO condition).

In addition, we expected participants who produced the preambles (leaders) to speed up production of the preamble in the OTHER condition compared to the NO condition, similarly to participants in Experiment 7. This would show that speakers are sensitive to the fact that another is about to speak, and that they have a preference to avoid overlap. Importantly, in this experiment we recorded the two participants’ utterances on separate channels (see Methods). Therefore, it was possible to test whether faster responses on the part of leaders led to smaller asynchronies between the leaders and followers, as in Vesper et al. (2011).

Further, we tested two versions of this experiment. In one version (reported below as Experiment 8a), the three conditions were administered in a completely randomized order. In another version (reported below as Experiment 8b), the three conditions were administered in separate blocks. Vesper et al. (2011) tested their individual (NO) condition and their joint (OTHER) condition in separate blocks. It is possible that the requirement of changing task rule on a trial-by-trial basis might make it more difficult for leaders to implement a consistent coordination strategy in the OTHER condition.

In addition, in Experiment 8b we varied the order in which the three conditions were administered. An effect of continuation Difficulty on the duration of the preamble in the OTHER condition could potentially be explained by the fact that speakers covertly planned
the continuation even when they did not have to produce it, perhaps because they were affected by what they did in the SELF condition. If this is the case, we would expect to find the effect only when the OTHER condition in intermixed with the SELF condition (as in Experiment 8a) or in those pairs of participants to which the OTHER condition was administered after the SELF condition (in Experiment 8b).

4.4.1 Method

Participants

Twelve pairs of previously unacquainted participants from the University of Edinburgh student community were paid £6 to participate in Experiment 8a (5 female pairs, 7 mixed gender pairs). Twelve new pairs of participants from the same population were paid to take part in Experiment 8b (4 female pairs, 1 male pair, 7 mixed gender pairs). All participants reported being native speakers of English with no speaking or reading difficulties.

Materials

The materials used in this study were 21 black and white line-drawings showing one of seven simple actions. The corresponding verb (chase, follow, punch, hit, kiss, scold, shoot) was printed underneath the scene in Times New Roman 18 points bold capital letters. Each verb was used in 3 different items. Each item consisted of a unique combination of a verb, an agent (soldier, pirate, cowboy, painter, nun, policeman, waitress, swimmer, doctor, ballerina, monk, or sailor), one or two patients (chosen from the following list: clown, boxer, burglar, swimmer, doctor, ballerina, monk, sailor, pirate, nun, waitress, soldier, policeman), and one or two objects (chosen from the following: apple, axe, balloon, basket, bell, bottle, bow, broom, candle, cane, crown, drum, flag, flower, guitar, hammer, hanger, hat, kite, ladder, lamp, owl, racket, trumpet, umbrella, vase, violin, wand). Each object was used in
only one item. Each character appeared in between one and five items, but in a maximum of three items as agent.

For each of the 21 items we prepared two versions, an easy version and a hard version. The easy version showed the agent, the verb, and one patient (e.g., The soldier chases the swimmer). The hard version showed the same agent and verb as the easy version, but the patient was modified in one of three ways. The hard version of seven items showed the patient carrying two objects. These were intended to elicit descriptions in which the object NP was followed by a PP (PP-items), which in turn comprised of two conjoined noun phrases (e.g., The soldier chases the swimmer with the cane and the vase).

The hard version of seven other items showed the patient carrying only one object; in addition, the patient was accompanied by a second character that did not carry any objects. These items were intended to elicit descriptions in which the object NP was composed of two conjoined noun phrases, the first of which was modified by a PP (NP-PP-NP-items). For example, the hard version of the item “The cowboy punches the doctor” would be described as “The cowboy punches the doctor with the balloon and the swimmer”. (Note that participants were asked to describe the pictures from left to right; in this way, we made sure that the NP-PP constituent was produced first.)

Finally, the hard version of the seven remaining items showed the patient carrying no object and accompanied by a second character carrying one object. This set of items was intended to elicit descriptions in which the object NP was composed of two conjoined noun phrases, the second of which was modified by a PP (NP-NP-PP-items). For example, the hard version of the item “The policeman scolds the sailor” would be described as “The policeman scolds the sailor and the ballerina with the flag”. In this way, we hoped to introduce some structural variation and prevent the hard version from becoming too easy for participants over the course of the experiment. Each verb appeared once with each kind of structure. See Appendix A, Table A-5, for a complete list of experimental materials. Twelve
more pictures were created for use on practice trials. They used a subset of the characters, objects, and verbs used in experimental items, but in different combinations.

**Pre-test.** Fourteen participants from the University of Edinburgh community volunteered to take part in the pre-test. They were all native speakers of English and none of them later took part in Experiment 8a or 8b. Data from three participants were not analysed because of technical problems with the recording (1), the participant having motor coordination problems (1), or the participant providing incomplete descriptions (1). Data from one additional participant were discarded because only 64% (27/42) descriptions were coded as correct. This meant that data from ten participants were entered into the analysis. Participants took part individually and described 42 different target pictures, which were the easy and the hard versions of the 21 experimental items (see Materials). After the pre-test, two of the objects were replaced as participants had difficulties recognizing them in the hard versions (*scarf* was changed to *broom*, *tie* to *drum*). The pictures were displayed for 3500 ms. All other details of the Procedure were the same as in Experiment 8a and 8b (see Procedure below).

Data were analysed using linear mixed-effects models with Difficulty (Hard vs. Easy) as the only, sum-coded predictor, and maximal random effects structure (see Recording and Data Analyses below for additional details); as usual, we used a logistic link function for the likelihood of disfluency data, and a normal link function for timing data. Trials on which the description of the preamble was incorrect (6.67%) or incomplete (0.71%) were discarded. Speakers produced were very few disfluent preambles (9) and there was no indication that they were more likely to be disfluent before hard than before easy continuations (6 disfluent preambles before hard continuations, 3 before easy continuations; Difficulty: $\chi^2(1) = .06, p = .81$). We discarded onsets that were longer than 2500 ms (1.58%) as outliers. Preamble onsets were 18 ms faster before easy ($M = 1.286$ seconds, SD

---

$^{23}$ Cut-off values were chosen based on inspection of the distribution (see below as well).
than before hard continuations (M = 1.304 seconds, SD = .289), but this difference was not significant (Difficulty: \( \chi^2(1) = .15, p = .70 \)), nor was there any significant difference in onsets depending on the type of hard continuation (i.e., between PP-, NP-PP-NP and NP-NP-PP-items; \( \chi^2(3) = 1.19, p = .76 \)).

We discarded durations that were longer than 2500 ms (5.26%) as outliers. Preambles were 147 ms shorter before easy (M = 1.009 seconds, SD = .275) than before hard (M = 1.156 seconds, SD = .350) continuations, a difference that was highly significant (Difficulty: \( \chi^2(1) = 13.04, p < .001 \)). Post-hoc pairwise comparisons between different types of hard continuations revealed no significant differences (PP-items: M = 1.156 seconds, SD = .325; NP-PP-NP-items: M = 1.209 seconds, SD = .424; NP-NP-PP-items: M = 1.098 seconds; SD = .283, all differences, p > .1, corrected for multiple comparisons).

**Design**

Two factors were manipulated, both within participants and within items: Partner (NO, OTHER, SELF) and Difficulty (Easy vs. Hard). Each item was presented once to each participant in every cell of the design, for a total of twelve repetitions of the same item (3 Partner x 2 Difficulty x 2 Participants), and 252 trials in total. These were divided into three blocks of 84 trials each. Presentation order within each block was randomized separately for each pair and the order of blocks was also counterbalanced between pairs. In Experiment 8a, each condition was encountered equally often in each block, and the assignment of items to conditions was rotated across blocks. In Experiment 8b, the presentation of conditions was blocked, and the order of conditions was counterbalanced between pairs. Six different orders were possible, as all permutations of the three conditions were tested (N= NO, O = OTHER, S = SELF): SNO, SON, ONS, OSN, NSO, NOS. Each order was encountered by two pairs of participants in Experiment 8b. In both experiments, participants took turns performing the role of leader according to a randomized sequence.
Procedure

Participants were tested in pairs. Upon coming into the laboratory they were introduced to each other and told that they would perform a task together. First, however, they familiarized themselves with the names of the characters and objects in separate booths, assisted by two different experimenters. The characters were shown to the participant once with the correct names written underneath them. The participants repeated these names out loud to aid memorization; the experimenter controlled the pace with which the characters were presented. Then, participants were shown the characters (in a different, random order) without names, and were asked to name them out loud, with the experimenter providing feedback and correcting mistakes. This procedure was repeated once if needed. Then, the participants went through the same procedure with the object names.

At this point, the participants were brought into the same room and sat next to each other in front of a computer screen. A coin was flipped to decide which participant would sit on the left-hand side of the screen. Then, one of the experimenters (either the author or her collaborator; see Footnote 9) explained the instructions. Participants were informed that they would describe pictures together. They were told that they should describe everything they saw on the pictures (i.e., name all the characters and objects they may carry), from left to right, in linear order. They were also provided with examples of the sentence structures they should use to describe the hard versions of the pictures.

Then, the experimenter explained that before each picture they would see an instruction screen, with their names, or the word no, written next to the words start and continue (see Figure 4-3). The participant whose name was displayed next to start began the description by naming the agent and the verb. The participant whose name was displayed next to continue provided a description of the continuation (i.e., the patient and any objects and/or characters accompanying the patient). Therefore, if the same name appeared next to start and continue, the cued participant described the whole picture (SELF). If two different names appeared, one participant (leader) described the preamble, while the other (follower)
described the completion (OTHER). Finally, if the instruction no appeared next to continue, the cued participant described the preamble and the other participant remained silent (NO).

The experimenter then helped participants putting on the headset microphones and these were tested. The participants went on to complete some practice trials, which used different pictures than the experimental items (see Materials) but were otherwise identical to experimental trials. The experimenters remained in the room during the practice and provided feedback afterwards. The instructions emphasized that participants should try to speak as clearly as possible. In Experiment 8a, instructions for the three conditions were explained at the start, and participants practiced the three tasks together at the start of the session (there were 12 practice trials in total). In Experiment 8b, instructions were explained just before the corresponding block, and the corresponding task was also practiced just before the relevant block (there were 6 practice trials before each block, for a total of 18 practice trials).

Once the practice was completed, the two experimenters started the recording and left the booth. An experimental session lasted about an hour in both experiments. Between blocks, participants were allowed to take a short break. A trial began with a fixation cross being displayed for 1000 ms. Then, after a 500 ms blank, the instruction screen was displayed for 2000 ms. Following the instruction screen, the picture appeared for 3800 ms, which was judged to be a sufficiently long time on the basis of the pre-test. The ITI was 1000 ms.

**Recording and data analysis**

Utterances were recorded using high quality head-mounted SM10A microphones (Shure Distribution UK, Waltham Abbey, Essex, UK, http://www.shure.co.uk). Responses from the two participants were recorded onto two separate channels via an R-44 multiple channel EDIROL portable recorder (Roland Corporation, Hamamatsu, Japan, www.roland.com). A 75-ms beep was produced by E-Prime (Version 2.0) software every
time a picture was displayed (inaudible to the participants) and was recorded onto a third channel. Beep onsets were automatically tagged using the Silence Finder algorithm in Audacity (Version 1.2.5). Preamble Onsets (the onset of the agent’s name) were also automatically tagged (after removing background noise), and then manually checked. In addition, two time points were manually marked on the recordings: Preamble Offset (the offset of the verb), and Continuation Onset (the onset of the description of the continuation). These points were then used to derive three measures: Onset (the time elapsed from picture presentation till the participant began speaking), Duration (Preamble Onset-Preamble Offset), Gap (Continuation Onset – Preamble Offset). Tracks for the two participants in a pair were analysed simultaneously, so that if their speech overlapped this could be noted down as well.

In addition, preambles were checked for accuracy: incomplete descriptions, wrong descriptions (i.e., using the incorrect name for the agent and/or action), and disfluencies (i.e., hesitations and repetitions) were noted down. This procedure was carried out by the author for half of the recordings, and by her collaborator (see Footnote 9) for the other half (in each experiment). Only disfluencies that occurred before speech onset or during articulation of the preamble, but no later than the verb were analysed. This was done to ensure comparability between conditions.

We analysed the data using Generalized Linear mixed-effects models, using the lme4 package (Baayen, et al., 2008; Bates, et al., 2008) in R (R, Version 2.13.1), with a logistic link function (Jaeger, 2008) for categorical binary data (fluent vs. disfluent, and non-overlapped vs. overlapped), and a normal link function for timing data (Onset, Duration, and Gap). For Experiment 8a, the fixed effect structure contained two predictors initially: Partner (NO vs. OTHER vs. SELF) and Difficulty (Easy vs. Hard), as well as their interaction. (For the analysis of the likelihood of overlapping, Difficulty was the only predictor of course, given that this analysis was restricted to the OTHER condition.) Both predictors were sum-coded. We proceeded to select the best-fitting model using a backward stepwise procedure.
If an interaction was significant, the corresponding main effects were retained in the model as well. We report estimates, standard errors, Wald’s Z (for logistic models) or t tests for the best fitting model.

Whenever possible, the random structure was kept maximal (Barr, et al., 2013), which means that we included by-participant and by-items random slopes for Partner, Difficulty, and their interaction. Whenever the maximal random structure did not converge, we simplified it, starting with the higher order term (the Partner-by-Difficulty interaction), and proceeding to eliminate random slopes if necessary, first by items, then by subjects. Estimates of the random effects are reported, as well as correlations between random effects.

To analyse the variability of leaders’ responses, we computed the standard deviation of onsets in each of the six cells of the design. Average standard deviations were computed for each participant and item. These data where then entered into by-participant and by-item two-way repeated measures ANOVA’s, with Type and Partner as within-participants and within-items factors.

We analysed the data of Experiment 8b in the same way. However, because the factor Partner was blocked in this experiment, the fixed effect structure contained Block (First vs. Second vs. Third) as an additional predictor, as well as its two-way interactions with Partner and Type and the three-way interaction. Block was contrast-coded: we defined a contrast comparing the mean of the second and third blocks to the mean of the first block (Block1) and another contrast comparing the mean of the third to the mean of the second Block (Block2). Random structure was kept maximal (Barr, et al., 2013), which means that we included by-items random slopes for Partner, Difficulty, Block, the three two-way interactions and the three-way interaction. By-participants random slopes were limited to Partner, Difficulty, Block, the two-way interaction of Partner and Type, and the two-way interaction of Block and Type; as the assignment of Blocks to conditions was counterbalanced between pairs of participants, it was not appropriate to include random slopes for the Partner-by-Block interaction and the three-way interaction (see Design).
Variability of leaders’ responses was analysed as in Experiment 8a, except that, in addition to Partner and Difficulty, Block was entered into the analyses as a within-item and between-participants factor with three levels (see Design). Finally, we conducted combined analyses of the two experiments, which included Experiment as a between-participant and within-item factor.

### 4.4.2 Results

We removed trials on which the description of the preamble was incomplete or wrong (3.27% in Experiment 8a, 3.41% in Experiment 8b). Combined analyses indicated that participants overlapped their speech more often in the OTHER condition of Experiment 8b (11.08%) than in the OTHER condition of Experiment 8a (4.08%; $\chi^2(1) = 6.62, p < .05$). In both experiments, participants showed a tendency to overlap their speech more often on easy trials than on hard trials (Experiment 8a: easy, 5.49%; hard, 2.66%; Experiment 8b: easy, 13.96%; hard, 8.14%). However, the main effect of Difficulty failed to reach significance both in combined analyses ($p = .12$), and in separate analyses of the two experiments (Experiment 8a: $\chi^2(1) = 1.59, p = .21$; Experiment 8b: $\chi^2(1) = 3.45, p = .06$). In Experiment 8b, the likelihood of overlap did not depend on the block in which the OTHER condition was administered. Interestingly, there was considerable individual variation in the likelihood of overlapping (see Figure 4-4). Overlapping speech trials were removed from all subsequent analyses.
Figure 4-4.

Histograms showing the distribution of proportions of overlapping trials for pairs in Experiment 8a (panel A) and 8b (panel B).
4.4.2.1 Likelihood of disfluency while describing the preamble

Experiment was not a significant predictor of the likelihood of producing a disfluency, neither alone nor in interactions (all \( p \)'s >.1). Partner did not contribute to model fit, neither alone, not in interaction with Difficulty (all \( p \)'s >.4). Finally, Difficulty contributed to model fit only marginally (\( \chi^2(1) = 3.54, p = .06 \)), indicating that in both experiments participants showed a tendency to be disfluent more before hard than before easy continuations (see Table 4-8). Separate analyses were also conducted for Experiment 8a and Experiment 8b, and they confirmed the results of the combined analyses (in addition, the analysis of Experiment 8b showed that Block did not affect the likelihood of participants being disfluent, either alone or in interactions).

Table 4-8.

Counts of disfluent trials (and percentages) by Partner and Difficulty in Experiment 8a (randomized) and 8b (blocked).

<table>
<thead>
<tr>
<th>Partner</th>
<th>NO</th>
<th>OTHER</th>
<th>SELF</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficulty</td>
<td>8a</td>
<td>8b</td>
<td>8a</td>
<td>8b</td>
</tr>
<tr>
<td>Easy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>6</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>1.62%</td>
<td>1.81%</td>
<td>1.29%</td>
<td>2.39%</td>
<td>1.25%</td>
</tr>
<tr>
<td>Hard</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>24</td>
<td>5</td>
<td>12</td>
<td>11</td>
</tr>
<tr>
<td>2.06%</td>
<td>4.85%</td>
<td>1.05%</td>
<td>2.73%</td>
<td>2.26%</td>
</tr>
<tr>
<td>Total</td>
<td>18</td>
<td>33</td>
<td>11</td>
<td>22</td>
</tr>
<tr>
<td>1.84%</td>
<td>3.33%</td>
<td>1.17%</td>
<td>2.56%</td>
<td>1.76%</td>
</tr>
</tbody>
</table>
4.4.2.2 Timing analyses

**Duration**

We discarded durations that were shorter than 300 ms or longer than 3000 ms as outliers (Experiment 8a: 0.11%; Experiment 8b: 0.04%). The three-way interaction of Partner, Difficulty, and Experiment was not significant ($p > .9$), nor was any other interaction involving Experiment. Durations were shorter in the SELF condition than in the other two conditions ($B = -82$ ms, $SE = 17$, $t = -4.83^{24}$; see Table 4-9). However, there was no indication that participants spent less time articulating the preamble when they knew their partner was about to speak (OTHER) than when they knew their partner would remain silent (NO; $B = -14$ ms, $SE = 13$, $t = -1.04$).

Importantly, the best model included the interaction of Partner and Difficulty ($\chi^2(2) = 10.98$, $p < .005$; see Table B-22 in Appendix B)$^{25}$. To further explore the interaction, we fitted the model in Table B-22 again, but used contrast coding for the factor Partner (instead of sum coding). In particular, we defined two contrasts. The first contrast compared the mean of the SELF condition to the mean of the OTHER and NO conditions. The second contrast compared the mean of the OTHER condition to the mean of the NO condition. This analysis showed that the effect of Difficulty was significantly larger in the SELF compared to the other two conditions ($B = 45$ ms, $SE = 14$, $t = 3.22$), but also that the effect was no larger in

---

$^{24}$ We take estimates with $t$ values exceeding $|2|$ to be reliably different from zero, as the number of observations in this study is well over a thousand (Baayen, et al., 2008).

$^{25}$ For completeness, in Appendix B we report the best fitting models for the analyses of Duration in Experiment 8a (Table B-23) and Experiment 8b (Table B-24) separately. The interaction of Partner and Difficulty improved model fit in Experiment 8b ($\chi^2(2) = 6.47$, $p < .05$), but not in Experiment 8a ($\chi^2(2) = 4.46$, $p = .11$). However, given that the three-way interaction of Partner, Difficulty, and Experiment was not significant (see text) and, further, that Block did not interact with any of the other predictors in Experiment 8b (all $p$’s > .06), we decided to focus on the combined analyses of the two experiments. Note that numerically the pattern of results was very similar in the two Experiments. In Experiment 8b, the model included a main effect of Block ($\chi^2(2) = 8.11$, $p < .05$; see Table B-24), as participants became progressively faster over the course of the experiment (Block 1: $M = 1073$ ms, $SD = 274$; Block 2: $M = 1036$ ms, $SD = 272$; Block 3: $M = 1025$ ms, $SD = 306$).
the OTHER compared to the NO condition (B=11 ms, SE = 15, t = .76). This is not consistent with the shared mechanisms or the overlapping mechanisms accounts.

Table 4-9.

Mean preamble Duration (and standard deviation), in milliseconds, by Partner and Difficulty in Experiment 8a (randomized) and Experiment 8b (blocked).

<table>
<thead>
<tr>
<th>Partner</th>
<th>NO 8a</th>
<th>NO 8b</th>
<th>OTHER 8a</th>
<th>OTHER 8b</th>
<th>SELF 8a</th>
<th>SELF 8b</th>
<th>Total 8a</th>
<th>Total 8b</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficulty</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Easy</td>
<td>1102</td>
<td>1067</td>
<td>1075</td>
<td>1046</td>
<td>977</td>
<td>961</td>
<td>1051</td>
<td>1024</td>
</tr>
<tr>
<td></td>
<td>(252)</td>
<td>(282)</td>
<td>(238)</td>
<td>(263)</td>
<td>(255)</td>
<td>(261)</td>
<td>(254)</td>
<td>(273)</td>
</tr>
<tr>
<td>Hard</td>
<td>1120</td>
<td>1088</td>
<td>1107</td>
<td>1078</td>
<td>1044</td>
<td>1033</td>
<td>1090</td>
<td>1066</td>
</tr>
<tr>
<td></td>
<td>(270)</td>
<td>(296)</td>
<td>(243)</td>
<td>(272)</td>
<td>(299)</td>
<td>(311)</td>
<td>(273)</td>
<td>(295)</td>
</tr>
<tr>
<td>Total</td>
<td>1111</td>
<td>1077</td>
<td>1091</td>
<td>1062</td>
<td>1010</td>
<td>997</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(261)</td>
<td>(289)</td>
<td>(241)</td>
<td>(268)</td>
<td>(280)</td>
<td>(289)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Onset

We discarded onsets that were shorter than 300 ms or longer than 3000 ms as outliers (0.11% in Experiment 8a; 0.07% in Experiment 8b). The best model (see Table B-25) included a main effect of Difficulty ($\chi^2(1) = 11.71, p < .001$). Unlike in the pre-test, preambles were initiated later when the continuations were hard than when they were easy. Importantly, this effect was of similar magnitude in the three conditions (see Table 4-10). Hard items were visually more complex than easy items. This might have made it more difficult for speakers to identify the agent and the patient on hard than on easy trials. We return to this issue in the Discussion.
In addition, the model (see Table B-25) included an interaction between Partner and Experiment ($\chi^2(2) = 7.30, p < .05$). Analyses that used contrast coding for the factor Partner (see Duration) showed that the estimate for the contrast between the SELF and the other two conditions was not different between experiments (B= 36 ms, SE = 20, t = 1.81), but the estimate for the contrast between the OTHER and the NO condition was more negative in Experiment 8b than in Experiment 8a (B= -51 ms, SE = 24, t = -2.13).

Inspection of Table 4-10 shows that onsets in the NO and SELF conditions were very similar in the two experiments. Instead, onsets in the OTHER condition were faster in Experiment 8b than in Experiment 8a. This might indicate that participants in Experiment 8b implemented a different strategy than participants in Experiment 8a to aid coordination.

Separate analyses of the two experiments showed that in Experiment 8a onsets in the OTHER condition did not differ from onsets in the other conditions; the main effect of Partner was not significant, and the best model included only a main effect of Difficulty ($\chi^2(1) = 10.19, p < .005$; see Table B-26).

In Experiment 8b, instead, onsets in the OTHER condition were shorter than in the NO condition; the best model included a main effect of Partner ($\chi^2(2) = 20.58, p < .001$), as well as a main effect of Difficulty ($\chi^2(1) = 12.96, p < .001$; see Table B-27). Analyses that used contrast coding for the factor Partner showed that preambles were initiated more slowly in the SELF than in the average of the other two conditions (B= 62 ms, SE = 15, t = 4.15), and, crucially, that preambles were initiated more quickly in the OTHER than in the NO condition (B= -64 ms, SE = 16, t= -4.02). This replicates a similar finding in Experiment 7 and contrasts with the lack of a difference between conditions in Experiment 8a. We discuss potential reasons for this discrepancy in the Discussion.

\[26\text{In addition, the model included a main effect of Block ($\chi^2(2) = 15.86, p < .001$), as participants became progressively faster over the course of the experiment (Block 1: M =1124 ms, SD = 270; Block 2: M = 1068 ms, SD = 261; Block 3: M = 1030 ms, SD = 258). Importantly, Block did not interact with the other two predictors (all p’s}>.1).}\]
Table 4-10.

_Mean preamble Onset (and standard deviation), in milliseconds, by Partner and Difficulty in Experiment 8a (randomized) and in Experiment 8b (blocked)._  

<table>
<thead>
<tr>
<th>Partner</th>
<th>NO</th>
<th>OTHER</th>
<th>SELF</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficulty</td>
<td>8a</td>
<td>8b</td>
<td>8a</td>
<td>8b</td>
</tr>
<tr>
<td>Easy</td>
<td>1058</td>
<td>1051</td>
<td>1049</td>
<td>1003</td>
</tr>
<tr>
<td></td>
<td>(253)</td>
<td>(247)</td>
<td>(287)</td>
<td>(228)</td>
</tr>
<tr>
<td>Hard</td>
<td>1114</td>
<td>1125</td>
<td>1110</td>
<td>1044</td>
</tr>
<tr>
<td></td>
<td>(284)</td>
<td>(270)</td>
<td>(314)</td>
<td>(264)</td>
</tr>
<tr>
<td>Total</td>
<td>1086</td>
<td>1087</td>
<td>1080</td>
<td>1024</td>
</tr>
<tr>
<td></td>
<td>(270)</td>
<td>(261)</td>
<td>(302)</td>
<td>(248)</td>
</tr>
</tbody>
</table>

**Gap**

This analysis was conducted only on the SELF and OTHER conditions, and it was carried out separately for the two experiments, because some of the effects of interest interacted with Block in Experiment 8b (see below).

*Experiment 8a.* On the basis of the distribution, we discarded gaps that were longer than 2500 ms as outliers (0.32%). Gaps were on average 170 ms longer in the OTHER condition than in the SELF condition (see Table 4-11). Gaps were also around 100 ms longer before hard than before easy continuations in the OTHER condition, whereas they were only 7 ms longer before hard than before easy continuations in the SELF condition. Inspection of the distribution revealed that Gap was non-normally distributed. To correct for this, we applied Box-Cox transformations (Box & Cox, 1964) using the package geoR (Diggle &
Ribeiro Jr., 2007, 2011). The interaction of Difficulty and Partner significantly improved ($\chi^2(1) = 18.40, p < .001$) the fit of the model (see Table B-28).

Table 4-11.

Mean Gap (and standard deviation), in milliseconds, by Partner and Difficulty in Experiment 8a.

<table>
<thead>
<tr>
<th>Difficulty</th>
<th>Partner</th>
<th>OTHER</th>
<th>SELF</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy</td>
<td></td>
<td>221 (197)</td>
<td>98 (131)</td>
<td>158 (178)</td>
</tr>
<tr>
<td>Hard</td>
<td></td>
<td>322 (243)</td>
<td>105 (142)</td>
<td>213 (226)</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>272 (227)</td>
<td>102 (137)</td>
<td></td>
</tr>
</tbody>
</table>

Experiment 8b. As in Experiment 8a, we discarded gaps that were longer than 2500 ms as outliers (0.11%). Gaps were on average 159 ms longer in the OTHER condition than in the SELF condition (see Table 4-12). Gaps were also 57 ms longer before hard than before easy continuations in the OTHER condition, whereas they were only 9 ms longer before hard than before easy continuations in the SELF condition. Inspection of the distribution revealed that Gap was non-normally distributed. To correct for this, we applied Box-Cox transformations as above. The best model (see Table B-29) included the interaction of Partner and Difficulty ($\chi^2(1) = 11.41, p < .001$). This finding replicates Experiment 8a, and suggests that in the SELF condition speakers completed the additional planning required to prepare difficult continuations before they finished producing the preamble.

In addition, in Experiment 8b the interaction of Partner and Block also improved model fit ($\chi^2(2) = 6.18, p < .05$). Inspection of Table 4-12 suggests that while gap length did not vary with Block in the SELF condition, in the OTHER condition gaps were longer in the third than in the previous blocks. The interaction of Partner and Block was unexpected.
Quite surprisingly, it shows that pairs who had more experience with the pictures (and with each other’s speech) when they performed the joint task (OTHER) coordinated less fluently (i.e., left longer silent gaps between their utterances).

Table 4-12.

*Mean Gap (and standard deviation), in milliseconds, by Partner and Difficulty and by Partner and Block in Experiment 8b.*

<table>
<thead>
<tr>
<th>Difficulty</th>
<th>Partner</th>
<th>OTHER</th>
<th>SELF</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy</td>
<td>OTHER</td>
<td>220 (241)</td>
<td>86 (120)</td>
<td>148 (198)</td>
</tr>
<tr>
<td>Hard</td>
<td>OTHER</td>
<td>277 (210)</td>
<td>95 (116)</td>
<td>182 (190)</td>
</tr>
<tr>
<td>Block</td>
<td>OTHER</td>
<td>222 (207)</td>
<td>91 (134)</td>
<td>154 (185)</td>
</tr>
<tr>
<td>First</td>
<td>OTHER</td>
<td>227 (225)</td>
<td>92 (99)</td>
<td>155 (182)</td>
</tr>
<tr>
<td>Second</td>
<td>OTHER</td>
<td>298 (242)</td>
<td>88 (120)</td>
<td>186 (215)</td>
</tr>
<tr>
<td>Third</td>
<td>SELF</td>
<td>90 (118)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>SELF</td>
<td>249 (227)</td>
<td>90 (118)</td>
<td></td>
</tr>
</tbody>
</table>

**4.4.2.3 Variability**

This set of analyses was carried out separately for the two experiments, because some of the effects of interest interacted with Block in Experiment 8b (see below).

**Experiment 8a.** There was indication that the variability of leaders’ responses varied with Difficulty. Preamble onsets were numerically more variable before hard than before easy continuations (see Table 4-10), though this difference was significant by participants (F<sub>1</sub>(1,23)=6.54, p<.05), but not by-items (F<sub>2</sub>(1,20)=1.78, p=.20). Preamble durations were also more variable before hard than before easy continuations (see Table 4-
9), and this difference was reliable both by-participants ($F_1(1,23)=5.65$, $p<.05$) and by-items ($F_2(1,20)=5.52$, $p<.05$).

The variability of leader’s responses was also affected by Partner, but in a less consistent way. First, variability of preamble onsets was affected by Partner ($F_1(2,46)=6.59$, $p<.005$; $F_2(2,40)=4.61$, $p<.05$). Post-hoc comparisons revealed that the SELF and NO condition did not differ from each other (both $p$’s >.9), while the OTHER condition differed significantly from the SELF condition ($p_1 <.05$, $p_2 <.05$), and also from the NO condition ($p_1 = .053$, $p_2 <.05$). However, inspection of Table 4-10 revealed that variability was actually higher in the OTHER condition compared to both the SELF and the NO condition. Average standard deviation of Onset did not correlate with average (Box-Cox transformed) Gap (by-participants, $r(22)=.297$, $t=1.46$, $p=.16$).

Second, the effect of Partner on the variability of preamble durations was reliable by items only ($F_1(2,46)=0.92$, $p=.41$; $F_2(2,40)=7.86$, $p<.005$). Post-hoc comparisons for items showed that this difference was mainly due to variability being lower in the OTHER than in the SELF condition (see Table 4-9; $p_2 <.005$). Variability in preamble durations was also numerically lower in the OTHER than in the NO condition, but this difference was only marginally significant (see Table 4-9; $p_2 = .06$). Average standard deviation of Duration did not correlate with average (Box-Cox transformed) Gap (by-participants, $r(22)=-.073$, $t=-.34$, $p=.73$). Overall, therefore, we found little evidence that leaders reduced the variability of their responses to facilitate coordination in the OTHER condition. This replicates the null effect in Experiment 7.

**Experiment 8b.** Although variability was numerically larger before hard than before easy continuations both for preamble onsets (see Table 4-10) and preamble durations (see Table 4-9), standard deviations of leaders’ responses were not reliably affected by Difficulty (SD of Onset: $F_1(1,21)=3.59$, $p=.07$, $F_2(1,20)=6.40$, $p<.05$; SD of Duration: $F_1(1,21)=.27$, $p=.61$, $F_2(1,20)=3.71$, $p=.07$). Importantly, the same held for Partner. Variability was
numerically smaller in the OTHER than in the NO condition, for onsets (see Table 4-10) and for durations (see Table 4-9), but the main effect of Partner failed to reach significance in all the analyses (all p’s > .12).

However, by-items analyses indicated that the effect of Partner varied by Block (see Table 4-13), as we found significant Partner * Block interactions for onsets ($F_1(2,43)=.92$, $p=.41$, $F_2(2,40) = 6.05$, $p<.01$) and durations ($F_1(2,43)=.37$, $p=.70$, $F_2(2,40) = 10.33$, $p<.001$). To explore this interaction, we ran a two-way by-item ANOVA separately for each block, with Partner and Difficulty as within-item factors (alpha was set to .0167 to correct for multiple comparisons).

For onsets, the main effect of Partner was significant only in the third block ($F_2(2,40) = 8.61$, $p<.001$). Post-hoc comparisons showed that variability in the NO condition did not differ from variability in the OTHER condition. The effect was driven by higher variability in the NO condition than in the SELF condition ($p_2<.001$), and higher variability in the OTHER than in SELF condition ($p_2<.05$). For durations, the main effect of Partner was significant in the first and the third block, though again only by-items (Block1: $F_2(2,40) = 12.23$, $p<.001$; Block2: $F_2(2,40) = 4.51$, $p =.02$; Block3: $F_2(2,40) = 5.59$, $p<.001$). Post-hoc comparisons showed that variability in the OTHER condition was reliably lower than variability in the NO condition only in the third block ($p_2<.001$), in which it was also lower in the SELF than in the NO condition ($p_2<.005$). In the first block, instead, variability was lower in the NO condition than in the SELF condition, and also lower in the OTHER condition than in the SELF condition (both $p_2$’s<.001). Overall, therefore, there was little indication that leaders’ responses were less variable in the OTHER than the NO condition, which replicates the null effects in Experiment 8a and in Experiment 7.
Table 4-13.

Average standard deviations of Onset and Duration, by Partner and Block in Experiment 8b.

<table>
<thead>
<tr>
<th>Block</th>
<th>NO</th>
<th>OTHER</th>
<th>SELF</th>
</tr>
</thead>
<tbody>
<tr>
<td>First</td>
<td>248</td>
<td>229</td>
<td>276</td>
</tr>
<tr>
<td>Second</td>
<td>247</td>
<td>260</td>
<td>274</td>
</tr>
<tr>
<td>Third</td>
<td>288</td>
<td>254</td>
<td>220</td>
</tr>
</tbody>
</table>

4.4.3 Discussion

In two experiments, we investigated whether speakers represent aspects of what another speaker is about to say. Specifically, we were interested in the mechanisms that speakers might use to form such predictive representations of the content and timing of others’ utterances. Crucially, we tested three different accounts. The shared mechanisms account proposes that speakers anticipate what others are about to say using the same mechanisms they use when planning their own utterances. The overlapping mechanisms account proposes that speakers anticipate what others are about to say using some, but not all of the mechanisms they use when planning their own utterances. The separate mechanisms account, finally, proposes that speakers anticipate what others are about to say using mechanisms that are in no way related to language production mechanisms.
First, participants in the SELF condition spent more time articulating the preamble before easy than before hard continuations. Importantly, this effect was reliably larger in the SELF than in the NO condition, so it cannot be due to visual differences between easy and hard pictures. Rather, the effect confirms the results of the pre-test. It indicates that speakers who produced both the preamble and the continuation planned the continuation while still articulating the preamble. Interestingly, this allowed them to leave only a short silent gap between the preamble and the continuation and the length of this gap was not affected by the difficulty of the continuation, suggesting that most of the planning for the continuation had been completed while still articulating the preamble.

Preamble onsets were also longer before hard than before easy continuations in the SELF condition (unlike in the pre-test). However, this was the case also for the NO condition (and for the OTHER condition, as well). We interpret this effect of continuation difficulty on onset as reflecting difficulties in event apprehension (Griffin & Bock, 2000), which affected speakers equally in all conditions as in all conditions participants had to conceptually process the whole event in order to identify the agent. It is not clear why this effect did not emerge in the pre-test.

Second, participants in the OTHER condition did not spend more time articulating the preamble when their partner later produced a hard continuation than when their partner produced an easy continuation. There was a numerical difference of about 32 ms in the right direction, but this was not statistically larger than the 20-ms difference observed in the NO condition. This finding is not consistent with either the shared mechanisms account or the overlapping mechanisms account. Instead, it is consistent with the separate mechanisms account. Therefore, we conclude that under the conditions tested in this experiment speakers did not anticipate what their partners were about to say using mechanisms they used to plan their own utterances ahead.

In addition, gaps between preambles and continuations were longer in the OTHER condition than in the SELF condition, and the length of gaps was affected by the difficulty of
the continuation more in the OTHER condition than in the SELF condition. These differences between the two conditions suggest that individual participants (in the SELF condition) were more successful in coordinating the timing of the two parts of their utterances than pairs of participants in the OTHER condition. Incidentally, note that on the basis of our results it is not possible to weight the relative contribution of leaders and followers to the duration of gaps in the OTHER condition. We have focused our attention on leaders, and showed that they did not represent what the followers were about to say using language production mechanisms. But it is also possible that followers were not planning sufficiently ahead while leaders were speaking.

For example, followers might have adopted a strategy by which they did not complete planning of the continuation until they detected that leaders were approaching the end of their utterance. In other words, it is possible that followers used leaders’ utterances as variable-timing response cues. Note that when speakers produce sentences in response to a cue, speech onset times vary with the length and complexity of the sentences (e.g., F. Ferreira, 1991), even when such utterances had been prepared beforehand. Therefore, even if followers pre-planned their utterances, it is possible that they completed planning (down to specification of an articulatory plan) only when they detected it was their turn to speak. This could explain why gaps were affected by continuation difficulty in the OTHER condition. Future experiments could tease apart the relative contribution of leaders and followers by having them coordinate with a partner who is not responsive (e.g., a recording).

Importantly, note that the average gap in OTHER was no longer than 300 ms. This means that followers most likely began planning while the leaders were still speaking (even though they completed planning only after the leaders had stopped speaking). Otherwise, we would have observed gaps of 500 to 600 ms at least, given the available estimates on the

Note that, numerically, the time available to plan the continuation before preamble offset (i.e., average onset plus average duration) was roughly comparable in the OTHER (Experiment 8a: 2171 ms, Experiment 8b: 2086 ms) and in the SELF condition (Experiment 8a: 2107 ms, Experiment 8b: 2103 ms)
time course of word production (Indefrey & Levelt, 2004). This might suggest that followers did not simply wait for leaders to stop speaking, but anticipated when they were about to stop speaking (De Ruiter, et al., 2006) and began planning their utterance in advance.

Moreover, speakers are certainly able to coordinate their utterances with even shorter gaps in natural conversations (Stivers, et al., 2009). Therefore, it is possible that joint language tasks that resemble natural conversation more closely that the task used in this experiment might lead to different results. We will discuss this issue in Chapter 5, at the end of this thesis.

A third aspect of our findings merits discussion. Leaders were faster to start their descriptions of the preambles in the OTHER condition than in the NO condition in Experiment 8b, but they were equally fast in Experiment 8a. The effect in Experiment 8b replicates a similar effect found in Experiment 7. We interpret it in the same way. Leaders sped up in order to avoid overlapping with their partner when they knew that their partner was about to speak. Interestingly, participants overlapped more often in Experiment 8b than in Experiment 8a, which might suggest that speeding up production was a reactive strategy on part of leaders. However, it is not clear why overlap was more likely in Experiment 8b. More research is needed to investigate the conditions under which speakers might adapt their production processes in order to avoid overlapping speech. Moreover, as in Experiment 7, we found little indication that speakers reduce the variability of their utterances to facilitate coordination, and there was no indication that reduced variability in either preamble onsets or durations correlated with shorter gaps. Therefore, we conclude that leaders’ utterances were not more predictable the less variable their onsets or durations (cf. Vesper, et al., 2011).

In conclusion, we showed that speakers are sensitive to the fact that another speaker is about to continue their utterance. However, we found no evidence to support the hypothesis that speakers can represent what another is about to say using mechanisms they use to plan their own utterances in advance of articulation.
4.5 General Discussion

In this chapter we investigated whether speakers can anticipate aspects of what another speaker is about to say using language production mechanisms, that is (some of) the same mechanisms they use when planning their own utterances ahead. Across three experiments, we asked speakers to describe sentence preambles of fixed length and complexity. They were then followed by their partners, who described continuations to those preambles. We varied the difficulty of these continuations and asked whether speakers who produced the preamble would be affected by the difficulty of the continuations despite the fact that another speaker was about to produce them.

In Experiment 6, continuation difficulty was varied by manipulating the similarity between a target shape and a context shape and by asking participants to describe the target shape in such a way that would allow an imagined listener to correctly select the target. In Experiment 7, continuation difficulty was varied by asking participants to describe transitive events with active or passive sentences. Finally, in Experiment 8 (8a and 8b), continuation difficulty was varied by manipulating the number of entities (characters or objects) the participants were required to describe.

In all experiments, we found no indication that speakers were affected by the difficulty of continuations produced by their partners. This suggests that speakers do not anticipate aspects of the content and timing of another speaker’s utterance in such a way that their predictions can affect their current production, at least not under the conditions tested in our consecutive language tasks.

Note that Experiments 6 and 7 must be interpreted carefully. In Experiment 6, pretest data indicated that our planning difficulty manipulation had an effect on speakers who went on to produce the continuations; however, this was not confirmed in the main experiment, despite identical materials and similar instructions, suggesting perhaps that the effectiveness of the manipulation hinges on aspects of planning that are subject to a great
deal of variability. In Experiment 7, we assumed that passive sentences would be harder to plan than active sentences based on some findings in the literature, but we did not confirm this would be the case for our materials and using our task (in which colour cues could be used to identify the subject of the utterances).

Despite these limitations of the first two experiments, Experiment 8 clearly showed that speakers spent more time articulating the preamble before they produced a continuation containing three lexical items than before they produced a continuation with only one lexical item. However, the same manipulation did not affect speakers who produced the preamble before a partner who produced the continuation. Therefore, we conclude that the hypothesis we laid out in Section 1.5.2 is not confirmed. Speakers do not use language production mechanisms to represent aspects of the content and timing of other speakers’ utterances in consecutive naming tasks as the ones we have tested in Chapter 4. We will discuss potential limitations of these tasks in Section 5.5, at the end of this thesis.

Here, we note that speakers were sensitive to the fact that another speaker was about to speak after them. In Experiment 7 and in the blocked version of Experiment 8 (8b), we found that speakers were faster to begin speaking when they knew that their partner would speak after them, than when they knew their partner would remain silent. Faster responses, however, were not less variable, nor did less variability correlate with shorter gaps between speakers. We propose that speakers sped up in an attempt to avoid overlapping speech, rather than as part of a strategy to make themselves more predictable for their partner (cf. Vesper, et al., 2011). Importantly, this finding can be explained within a mechanistic framework, as it is known that speakers can speed up production if faced with response deadlines (F. Ferreira & Swets, 2002). In other words, there is no need to assume that speakers implemented a turn-taking strategy as would be assumed by intentional accounts of coordination (e.g., Clark, 1996).
4.6 Acknowledgements

We thank Joris van de Cavey for his contribution to the design and analysis of Experiment 8. Kerry Goodacre, Ashild Bodal, Jamie Crowther, Alicia Harris, Regina Bieche, Isabella Malicki recruited and tested participants for Experiment 7. Ziggy Campbell helped with audio recording for Experiments 6 and 8.
5. Concluding remarks

We will first present an overview of the findings from the three studies presented in Chapters 2, 3, and 4 respectively (Section 5.1). Then, we will discuss to what extent the findings support the hypotheses laid out in Chapter 1. In particular, we will consider the following issues: Do speakers use production mechanisms to represent and predict others’ utterances? To anticipate, we will conclude that the answer to this question is positive. However, the evidence presented in this thesis is consistent with the idea that speakers can use production mechanisms to represent and anticipate whether another speaker is speaking or about to speak, but not with the idea that speakers routinely use language production mechanisms to represent or anticipate the content and timing of another speaker’s utterance.

We will then discuss what sorts of production mechanisms might be used for representing and predicting others’ utterances. In Chapter 1 we proposed two alternatives: forward-models of the production implementer and a partially activated production implementer. We will suggest that the evidence presented in this thesis is consistent with both alternatives. We will end with an evaluation of the novel joint language tasks used in this thesis, and with some indications for future research into the coordination of utterances between speakers.

5.1 Summary of empirical findings

In the first set of experiments (Study 1, Experiment 1-4), using a simultaneous production paradigm, we showed that speakers take longer to name pictures when they believe that their partner is concurrently naming pictures than when they believe their partner is silent or is concurrently categorizing the pictures as being from the same or from different semantic categories. Study 1 provided, for the first time, evidence that a joint interference effect holds for picture naming responses. Study 1 extended the vast literature on joint
interference effects with manual responses (see Section 1.2.2), and showed that language production is in this respect akin to other forms of actions.

Moreover, Study 1 did not support the dominant account of joint interference effects, namely the co-representation account. According to this account (see Sections 1.2.2 and 1.3.2), participants in joint tasks represent their partner’s responses in a format that is similar to how they represent their own responses. However, in four experiments in Study 1 we found no indication that speakers represent the content of their partners’ utterances, as speakers took longer to name pictures when they believed their partner was concurrently naming pictures, regardless of whether they believed the partner was naming the same picture, or a different picture (Experiments 3 and 4), or the same pictures in the same or in a different order (Experiments 1 and 2).

Therefore, we concluded that our findings are more consistent with another account of joint interference effects, namely the agent-conflict account. According to this account, participants in joint tasks represent whether it is their partner’s turn to respond or not, but not (necessarily) the content of their partner’s response. However, Study 1 also showed that speakers took longer to name pictures when they believed their partner was concurrently naming pictures than when they believed their partner was concurrently categorizing pictures. Semantic categorization of the pictures requires retrieval of the concepts associated with the pictures from memory. In addition, in our experiment, it required the participants to produce an overt verbal response. Therefore, this finding suggests that speakers represent not only whether it is their partner’s turn to respond in any way, but also whether it is their partner’s turn to use language production mechanisms.

In Study 2 (Chapter 3), using a consecutive production paradigm, we showed that speakers find it harder to stop speaking when they know that their partner is about to speak than when they know that their partner is not. This study therefore showed that speakers can anticipate that another speaker is about to start speaking, and that doing so interferes with language production mechanisms (in this case, mechanisms involved in stopping speech).
Interestingly, the study also showed that speakers take longer to stop speaking when they know that they themselves are about to speak again than when they know that their partner is about to speak. This suggests that anticipating that another speaker is about to speak is not the same as planning to speak.

In the final set of experiments (Study 3, Experiments 6, 7, 8a and 8b), we used three variations of the consecutive production paradigm, in which speakers described pictured events together. In particular, in the joint task, one participant (the leader) produced a preamble which included at least the sentence’s subject phrase (and the verb in Experiments 6, 8a, and 8b); the other participant (the follower) produced a continuation of variable length and complexity, which continued the preamble to produce a full description of the pictured event (and a grammatically complete sentence).

In Experiment 6, we found no indication that leaders anticipated aspects of the content and timing of the continuation that followers were about to produce. However, even speakers who went on to produce the continuation themselves (in the solo task) planned very little of the continuations before starting to speak and while producing the preambles. In Experiments 8a and 8b, instead, speakers in the solo task took longer to articulate the preamble before shorter than before longer continuations. Nevertheless, in the joint task leaders were not affected by the length of followers’ continuations. In addition, in Experiment 7 leaders took no longer to start speaking before active than before passive continuations. Overall, this set of results suggests that speakers do not anticipate the content and timing of others’ utterances using the same mechanisms they use to plan their own utterances.

However, Study 3 also confirmed that speakers are sensitive to the fact that another person is about to speak after them, though in a different way to Study 2. Recall that in that study speakers were asked to stop speaking as quickly as possible, and they found it harder when they knew their partner would speak after them, leading to a higher proportion of completed words. In Study 3, instead, leaders were given no instruction to stop speaking and
always completed their descriptions of the preambles. In addition, they were given no explicit instructions to avoid overlap. Nevertheless, they spontaneously developed a tendency to speed up production when they knew their partner was about to continue their utterance (Experiment 7 and 8b). We speculated that they did so in order to avoid overlapping their speech with the followers’ speech.

5.2 Do speakers use production mechanisms to represent, and anticipate, others’ utterances?

In this thesis, we showed that speakers can represent that other speakers are concurrently producing language (Study 1), and anticipate that other speakers are about to produce language (Study 2). Crucially, we showed that the computation of these (predictive) representations affects the way speakers produce an utterance. On the basis of this finding, we argued that the mechanisms involved in the computation of representations of others’ utterances overlap with the mechanisms involved in the production of utterances. This entails that speakers can use at least some production mechanisms to represent and anticipate utterances produced by another speaker.

Our conclusion rests on two assumptions. The first assumption is the following: if representations of others’ utterances in our joint language tasks were computed using mechanisms unrelated to language production mechanisms, then there would be no reason to expect such representations to affect concurrent language production. The second assumption is that our dependent measures (speech onset times, likelihood of completing a word, utterance durations) are indeed measures of the time course of language production mechanisms. For the validity of the second assumption, we rely on a long tradition of psycholinguistic studies, which have established the use of similar measures as indices for the time course of production processes (see e.g., Levelt et al., 1990 for onset times, Tydgat et al., 2011 for likelihood of completion; Gahl & Garnsey, 2004 for durations).
But is the first assumption justified? Perhaps, representations of others’ utterances are computed using mechanisms that are not related to production mechanisms but that affect some mediating variable, which in turn affects language production. For example, in our tasks participants could infer that their partner was speaking or about to speak based on task instructions, and making this inference could have affected the allocation of attentional resources, which in turn could have affected the time course of language production (e.g., Roelofs & Piai, 2011). This alternative explanation cannot be definitely discarded at this point, but a number of findings argue against it. First, Experiment 2 (Study 1) showed that the effect is at least partly specific to language production, as speakers experienced more interference when they believed their partner was naming pictures than when they believed their partner was categorizing pictures.

Unfortunately, we cannot provide analogous evidence for the effect reported in Study 2, where speakers found it harder to stop speaking when they knew their partner was about to name a picture. We expect speakers would not find it as hard to stop if they knew their partner was about to, for example, categorize the entity depicted on the picture as animate or inanimate. A future study could test this prediction. In addition, in Study 2 representations of another’s utterance influenced language production within the space of stopping production of a single lexical item. This further suggests that the effect of other-representations on production is more likely to be direct and immediate, rather than mediated by some additional processes.

Therefore, we conclude that speakers in our joint language tasks used language production mechanisms to represent that their partners were producing or about to produce language. This supports two of the hypotheses at the basis of the model we introduced in section 1.4.3, namely that speakers represent others’ utterances in a format that is similar to how they represent their own utterances, and that they can form predictive representations of others’ utterances. However, these assumptions need to be qualified, as we discuss in the next section.
5.3 What can speakers represent of others’ utterances?

In Section 1.4.3 (see Figures 1-1 and 1-2), we laid out a comprehensive inventory of the type of information that speakers could potentially represent about others’ utterances. There, we listed representations of both the content and timing of those utterances at all levels of the language production hierarchy, from semantics to phonology, and articulation. Now, it is time to evaluate which of these representations, if any, are actually computed in the language tasks we have implemented.

First, speakers in Study 1 did not represent the lexical content (i.e., the lemmas) of their partner’s utterances, nor did they represent the order in which those lemmas were produced. This suggests that they did not compute $\hat{p}($syn$)$ for their partners’ utterances. Second, in Study 3, speakers did not anticipate aspects of the content and timing of their partners’ utterances that differed between the hard and easy versions of those utterances. Let us focus on Experiment 8, as in that experiment difficulty had a clear effect in the solo task. As discussed in Section 4.4, we cannot determine precisely at what stage in the language production hierarchy our difficulty manipulation exerted its effects in the solo task. With reference to Bock and Levelt’s (1994) model, our manipulation might have affected the functional processing stage, if participants selected at least one of the lemmas for the additional entities depicted in the hard versions while producing the preambles. Alternatively, it might have affected the positional processing stage, if participants constructed a (possibly partial) frame that specified the position in the sentence of at least one of additional entities depicted in the hard versions of the materials.

Either way, speakers in Study 3 did not represent their partners’ utterances in a way that specified how the easy and hard versions differed in terms of grammatical encoding processes. This could be interpreted as suggesting they did not compute $\hat{p}($syn$)$ for their partners’ utterances, which would be consistent with Study 1. But it must be noted that easy and hard continuations differed also in terms of the message that was conveyed, and
obviously in terms of length. Therefore, the results of Study 3 could equally be interpreted as suggesting that speakers did not compute $p'(\text{sem})$ or $f(\text{art})$ for their partners’ utterances.

Note that we do not draw any conclusions as to whether speakers represented their partners’ utterances at levels that are further downstream in the language production hierarchy. This is because, as we discussed in section 1.4.3, it is possible that speakers might compute representations of others’ utterances at different levels independently of one another (unlike when they compute production representations during the planning of their own utterances). Therefore, the fact that speakers did not compute $p'(\text{syn})$ for their partners’ utterances does not imply that they could not have computed $f(\text{art})$. More generally, additional research is needed to conclude that speakers do not represent any aspect of the content and timing of others’ utterances. For example, we have not investigated directly whether speakers compute $p'(\text{phon})$ for others’ utterances.

However, speakers in our joint language tasks did not generally represent the content and timing of others’ utterances. Rather, they only represented whether another speaker was about to produce language. More precisely, they only represented whether another speaker was about to produce language using language production mechanisms. It is important to emphasize the latter. We are not claiming that speakers do not anticipate aspects of the content and timing of others’ utterances. The evidence for prediction in language comprehension (see Section 1.4.2.1), the smoothness of turn-taking, and the existence of cross-person compound contributions all indicate that people do indeed anticipate what others are about to say, and when (and not just whether they are about to say something or not).

But we could not find any evidence to support the idea that representations of the content and timing of others’ utterance are computed using the same mechanisms used in the production of utterances. It appears that production mechanisms are implicated in representing and anticipating others’ utterances, but that the representations computed using language production mechanisms are rather unspecific. Of course, this finding could be
related to limitations of our joint language tasks, and we will discuss these in Section 5.5 below.

### 5.4 What sorts of production mechanisms are used to represent and anticipate others' utterances?

To the extent that speakers do use language production mechanisms to represent and anticipate others’ utterances, what sort of language production mechanisms are these? In section 1.4.3 we noted that there are at least two alternatives. The first alternative is that such representations are computed using the production implementer (the implementation account). This means that computing representations of others’ utterances activates representations that are routinely accessed during the production of utterances. To return to the example in Figure 1-1, forming a representation that another speaker is about to produce the word “kite” could entail activation of the concept KITE, the lemma *kite*, and the phonological representation /kaIt/.

However, the fact that such representations could be activated does not mean that they always are. In fact, unless overt imitation of another’s utterance takes place, we must assume that the production implementer is inhibited to prevent articulation. But there is no theoretical reason to assume, *a priori*, that this inhibition takes place or affects only the very late stages of language production, just before motor plans that control the articulators are executed. To the contrary, it is possible that the flow of activation through the language production hierarchy is inhibited much earlier, potentially even before a lemma is accessed.

As discussed in section 1.4.1.1, listening to speech activates motor areas in the brain; furthermore, listening to syllables modulates the excitability of muscles involved in articulation in a very specific way. This would seem to imply that the implementer is fully activated, to the extent that representing others’ utterances enhances the readiness of specific muscles. But as we noted, such activation is not automatic and does not occur under every condition, but mostly when speech comprehension is difficult. Therefore, it should not be
taken as evidence that the production implementer is routinely activated to the fullest extent when representing others’ utterances.

The second alternative is that representations of others’ utterances are computed using the forward production model (the forward-model account). According to this alternative, no activation of production representations within the implementer is causally implicated in representing others’ utterances. However, activation of the implementer can occur as a consequence of computing predictive representations within the forward production model (as these computations are normally followed by activation of the implementer during language production). Below, we schematically illustrate how the two alternatives could explain the finding that representing whether another speaker is producing language or about to produce language interferes with language production.

According to the implementation account, a speaker represents that her partner intends to speak. In Pickering and Garrod’s (2013) model, this corresponds to a (recovered) production command. The production command activates representations within the production implementer of the speaker to some extent, but these representations are soon inhibited (to prevent an overt response). If, at the same time, the speaker is preparing to produce language, she will have also formed an intention to speak. This means that another production command is already controlling activation of representations within the production implementer.

Interference arises within the implementer. One possibility is that inhibition of representations activated by the production command corresponding to the partner’s intention to speak delays activation of representations activated by the speaker’s own production command. Another possibility is that amongst the representations activated by the two production commands are representations of the agent associated with that command (i.e., the speaker or her partner). For example, the production implementer might contain nodes that represent the speaker (self-node) and other language users (other-node). When the speaker intends to speak, the self-node is activated within the implementer. When the
speaker represents that another intends to speak, the other-node is activated within the implementer. This would allow speakers to keep representations for their own, and other speakers’ utterances separate. The self-node and the other-node might be in competition. Thus, if the other-node is activated, it could conflict with the self-node, dampening its activation. This could, in turn, interfere with production of the speaker’s own utterance.

Importantly, note that both possibilities outlined above account for interference without assuming that speakers represent the content of other speakers’ utterances. That is, both versions of the implementation account explain the findings from Study 1 and 2 without making assumptions that are, at present, not supported by empirical evidence and, furthermore, incompatible with the findings from Study 3.

According to the forward-model account, a speaker represents that her partner intends to speak. However, the (recovered) production command does not activate the production implementer. Rather, it is sent to the forward production model. The output of the forward production model is a prediction of the state the production implementer would be in if the production command was executed. In this case, based on a representation of the partner’s intention to speak, the speaker predicts that her partner will produce an utterance (though she does not predict any aspect of the content of that utterance).

If, at the same time, the speaker is preparing to produce language, she will have also formed an intention to speak herself. This means that another production command has been sent to the production implementer, and also to the forward production model, yielding a prediction of the state the implementer will be in once the command has been fully executed (this prediction could well specify aspects of the content of the speaker’s utterance; whether it does is not crucial for the present argument).

Current production of an utterance can then be affected in two ways. One possibility is that the prediction of her partner’s utterance is combined with the speaker’s prediction of her own utterance. This could cause the prediction of the state of the implementer to be less accurate, and consequently cause larger prediction errors when a representation of the
predicted state of the implementer is compared to a representation of its actual state (which can occur prior to articulation). This might in turn lead to additional adjustments to the production command, which would take time and slow down production.

Another possibility is that the two predictions computed by the speaker (for herself and for her partner) are not combined, but that the prediction that her partner is about to produce language independently affects the speaker’s own production implementer. For example, it might indicate that resources need to be reallocated from execution of the current production command to execution of another production command. This would result in resources being taken away (predictively) from the current production command, thus slowing down its execution.

In summary, we identified two potential mechanisms that could underlie the representation of others’ utterances. While they are both production mechanisms, in the sense that are routinely implicated in the production of utterances, their mode of operation is different. For each mechanism we also identified two possible ways in which recruitment of the mechanism to represent another speaker’s utterance could yield the observed interference effects in language production measures. The fact that various theoretical explanations are potentially compatible with our findings reflects a need for further empirical investigation.

For example, it might be possible to tease apart whether representations of others’ utterances are kept separate from representations of one’s own utterances using joint tasks in which a combined representation would have different properties than two separate representations. To illustrate, take two speakers naming pictures in adjacent soundproof rooms (as in Study 1). One of the speakers utters the pseudo-word *rab*, while at the same time the other utters the word *bit*. If representations are kept separate, the participant who utters *bit* might experience less interference in this condition then in a condition where her partner utters a word (e.g., *rub*). Recall that in Study 1 we showed that interference arises at least in part from representing that one’s partner is engaging in lexical access. But there is no lexical entry for *rab* (for English-speaking participants). However, if representations are
combined, the speaker who utters *bit* might actually represent the word *rabbit*. If this is the case, interference might occur to the same extent as in the *rub* condition.

Teasing apart the implementation and the forward-model account experimentally, however, is much harder. This is because both accounts allow the possibility that the production implementer is activated during the representation of others’ utterances, though they make different hypotheses with regard to the causal role of this activation. To provide positive evidence for the forward-model account, then, one would have to show that representations of others’ utterances are computed, and that they affect concurrent production, in a task where activation of the implementer could not be causally involved, either because the effect occurs too quickly, or because activation of the implementer is inhibited (for example by application of repetitive TMS; see Lesage et al., 2012). The evidence provided in this thesis is equally consistent with both accounts (see Sections 2.6 and 3.4.1).

### 5.5 Evaluation of joint language tasks and future directions

In Section 1.5, we argued that joint language tasks in which traditional psycholinguistic paradigms are distributed across two speakers could enhance our understanding of the mechanisms speakers use to coordinate utterances with one another. The comparison between speakers’ productions in a simultaneous or consecutive joint language tasks, and speakers’ productions in solo language tasks informs us on the extent to which language production mechanisms are implicated in representing, and anticipating, others’ utterances. Specifically, as we have seen, the experimental evidence collected using joint language tasks suggests that speakers do indeed represent others’ utterances using language production mechanisms, though these representations appear to be representations of whether another is speaking or about to speak rather than about when (or for how long) the other is speaking, or of what the other is saying.
5.5.1 Different routes to prediction

Clearly, speakers are able to form representations of others’ utterances that are more fine-grained. It is possible that the mechanism they use to compute these representations is independent of language production mechanisms. For example, it could be an associative mechanism, like the prediction-by-association route hypothesized by Pickering and Garrod (2013). This route to prediction, unlike the simulation route, does not recruit forward production models, or the implementer. Instead, it relies on the presence of regularities in the input. A comprehender who is sensitive to these regularities could learn to anticipate that occurrence of an event increases the probability of another event occurring in the future. The idea that language comprehenders keep track of such probabilities is incorporated in different ways in various account of prediction in language comprehension (e.g., Altmann & Mirković, 2009; N. J. Smith & Levy, 2013).

There might be situations in which the association route is the only route available to the comprehender. For example, late second language learners can sometimes comprehend words containing sounds that they are not able to produce. Moreover, they might anticipate the occurrence of that sound based on context. To illustrate, a late learner of Dutch could have difficulties producing the combination of a [s] sound and a [x] sound, as in the beginning of the word Schiphol. However, based on her experience of listening to Dutch people (and of reading Dutch), she can anticipate the initial sound of the word Schiphol in the context De luchthaven van Amsterdam heet Schiphol, using the association route. Once she hears the initial sound she can compare her perceptual representation of this event to her prediction, and update her prediction accordingly. This route can therefore support a late learner’s comprehension of the Dutch word in this particular instance, and also help her learn to fine-tune her predictions about the occurrences of particular sounds in particular contexts in Dutch.

However, what happens when the critical sound is masked by noise? The learner can still anticipate it based on the association route, but a comparison between her prediction and
perceptual information would no longer allow her to establish whether her prediction was correct or not, because she could not filter out the noise from the signal. A native speaker of Dutch, instead, can anticipate the sound using the simulation route. This means that he can internally generate a prediction of what it should sound like. This will cancel out the signal, so that noise can be extracted from the sensory input and filtered out.

In summary, this example illustrates that the simulation route is most useful when comprehension is more difficult. Therefore, it is possible that language production mechanisms are engaged in representing others’ utterances to a larger extent when others’ representations are more difficult to compute. This would be compatible with recent evidence that articulator-specific activation of the motor system during listening to speech is restricted to cases in which the input is noisy (e.g., D’Ausilio et al., 2012).

But can this observation help explain why speakers did not compute content-specific representations of others’ utterances in our joint language tasks? Perhaps, though note that we investigated the computation of other representations in a context where comprehension of others’ utterances did not take place at all (Study 1), or only took place after the point in time where representations would have been computed (Study 2 and Study 3). Therefore it is not immediately clear how this factor could have affected our results. Nevertheless, it is possible that speakers would have been more prone to anticipate the content of their partners’ utterances in Study 2 and Study 3 if those utterances were harder to comprehend.

5.5.2 “Minimal” joint tasks

Another aspect of our tasks that should be considered is that they are joint tasks only in a “minimal” sense. In Study 1, participants merely believed that they performed a task alongside another speaker. Although this belief was continuously reinforced throughout the experiment because participants observed their partner’s instructions as well as their own, the lack of interaction might have discouraged participants from representing the content and timing of their partners’ utterances using production mechanisms. It is possible that speakers
do so only when it is needed; that is, for example, only when they need to coordinate their utterances with other speakers.

Study 2 and Study 3 investigated precisely a situation in which speakers were interacting and could coordinate their utterances with those of another speaker. Nevertheless, Study 3 failed to provide evidence that speakers do represent what their partners are about to say using production mechanisms. Therefore, it is possible that language production mechanisms are not used by default to support coordination. Indeed, our speakers might have resorted to different coordination strategies, like speeding up the production of their utterances, instead of simulating their partners’ upcoming utterances.

If representing others’ utterances using language production mechanisms is not an automatic process, one could investigate the conditions under which speakers are more likely to employ this mechanism to support coordination between their utterances. Pickering and Garrod (2013) suggested that use of the simulation route might be enhanced by two factors. First, people should be more likely to rely on the simulation route in situations where they produce language than in situations where they don’t (e.g., in dialogue rather than monologue). This is because use of language production mechanisms during the production of one’s own utterances should prime use of the same mechanisms when representing somebody else’s utterances. Our joint language tasks were language production tasks, so clearly language production mechanisms were highly activated.

The second factor mentioned by Pickering and Garrod (2013) is the similarity between interlocutors (both actual and perceived, see Gambi & Pickering, 2013a). Participants in our studies were mostly strangers to each other and were given very little chance to interact prior to the experiment. Therefore, it is possible that they did not perceive their partner as sufficiently similar to themselves. Note that Pickering and Garrod (2013) and Gambi and Pickering (2013) focused on cases in which dissimilarity is apparent, as in child-parent interactions and in interactions between speakers of different dialects. Speakers in our studies came mostly from a relatively homogeneous group, but we did not control for
differences in accents. In addition, speakers took part in a very constrained interaction, and for a limited amount of time. Perhaps, the degree of (perceived or actual) similarity needed to trigger the use of production mechanisms to represent another’s utterances is greater when the chances to interact are more limited.

There are at least two ways in which this hypothesis could be tested. First, one could boost actual similarity by having the participants align prior to entering the joint task. For example, alignment at the syntactic and lexical level should make speakers more likely to use language production mechanisms to represent the content of their partner’s utterances. Phonetic imitation of one’s partner, or the experience of speaking in synchrony with them could also be beneficial, for a number of reasons. Pickering and Garrod (2004) proposed that alignment at one level boosts alignment at other levels, so phonetic alignment could enhance lexical and syntactic alignment. But in addition, phonetic alignment could increase perceived similarity because speakers might be particularly sensitive to similarity at this level, given its status as a powerful social cue to group identity (e.g., Pope, Meyerhoff, & Ladd, 2007). Indeed, a recent study showed that imitating the accent of a speaker increases the perceived social attractiveness of that speaker (Adank, Stewart, Connell, & Wood, 2013).

Incidentally, note that alignment could not only favour the use of language production mechanisms over other mechanisms, but it could also facilitate how accurate representations of others’ utterances computed using language production mechanisms will be (Pickering & Garrod, 2013). Clearly, the more similar two speakers are, the easier it should be for one to simulate the other, as this would be very close to simulating oneself. However, in our experiment speakers had sufficient information to represent their partner’s utterances accurately or at least with sufficient accuracy to be differentially affected by our manipulations. Therefore, it is likely that this second role of alignment would be less relevant in our joint language tasks.

Nevertheless, it is possible that coordination requires more than accurate representations of another speakers’ utterances. Perhaps, speakers would benefit by
synchronizing with their partner in speech rate or breathing patterns (see Section 1.4). Such synchronization could function as a coordination smoother (Vesper et al., 2011) and could be relatively independent of the formation of joint task goals and representations. Interestingly, a recent study of pianists performing piano duets showed that familiarity with the part played by one’s co-performer (which implies that the co-performer’s part can be accurately represented) does not reduce asynchrony; instead, repeated practice with the co-performer leads to better temporal coordination over time (Ragert, Schroeder, & Keller, 2013). Future research should focus both on the role of alignment in facilitating use of production mechanisms, and on the role of alignment (and synchronization) in making production-based representations of others’ utterances more accurate.

Second, one could boost perceived similarity by increasing affiliation between speakers before they begin the joint task. There are several examples of how to do this in the literature on joint action, from manipulating whether participants feel they belong to the same group as their partner (e.g., Mueller, Kuhn, et al., 2011), to increasing their pro-social attitudes via implicit priming (e.g., Colzato, et al., 2012).

But is similarity, whether actual or perceived, sufficient? One key ingredient of natural conversations is the presence of a joint goal. Interlocutors coordinate in order to reach a joint goal, that is, to bring about some change in the world (Sebanz, Knoblich, & Bekkering, 2006). This does not mean that interlocutors can only coordinate intentionally, or that they need to have joint intentions in order to engage in conversation (contra Clark, 1996). To the contrary, interlocutors simply need to represent that to attain the desired outcome they must perform their part in a joint action that involves another agent performing his or her part (Vesper et al., 2010).

For example, one joint goal in natural conversations could be smooth turn-taking (i.e., keep gaps as short as possible while avoiding overlap), of course alongside many other joint goals, both pertaining to the management of the linguistic interaction itself (e.g., agree
on referring expressions when no conventionalized label is available) and to the management of nonverbal aspects of the interaction (i.e., actions; e.g., complete a transaction in a shop).

Did speakers in our joint language tasks have a joint goal? They knew they were doing “something together”, but is this the same as having a joint goal? It is possible that speakers did not represent a joint goal. In other words, they simply represented their task and their partner’s task as separate rather than as two parts of the same task, directed towards a joint goal. Despite this, they automatically represented the other’s task, and moreover, they did so using language production mechanisms. However, the representations they formed were unspecific. Perhaps, speakers represent the content of others’ utterances using production mechanisms only when this is actually needed to achieve coordination towards a joint goal.

One way of testing this hypothesis is to introduce an explicit joint goal for the speakers. For example, speakers in a consecutive joint language task could be asked to produce the two parts of the utterance “as if it was produced by a single speaker”, or they could be directly asked to avoid pauses and overlaps between their utterances. Note that in the former case, smooth turn-taking would be an implicit goal, whereas in the latter case it would be an explicit goal. In either case (i.e., implicit or explicit goal), it doesn’t have to be the case that speakers coordinate intentionally or strategically, as they can have a joint goal without having shared intentions (see above).

A second way of introducing a joint goal for the participants would be to set up a task in which their interaction is directed at achieving a change in the external word. In our joint language tasks, speakers are not trying to bring about a change in the environment. To the contrary, it could be said that their interaction does not serve any function, apart from that of fulfilling experimental instructions. While this is the case for most psycholinguistic studies of monologue, studies of dialogue normally involve a communicative aspect. Interlocutors need to communicate because they possess different pieces of information that need to be shared in order to solve the task.
One possibility, then, is that speakers will represent others’ utterances using language production mechanisms only when they possess different information that needs to be integrated to reach the joint goal of task success. At present, our joint language tasks do not incorporate this aspect, as all information was available to all participants. Interestingly, in Experiment 6, information was available to both speakers in the joint task but not to an imagined listener. Still, speakers did not appear to represent each other’s utterances using the same mechanisms they use to represent their own utterances. However, the relevant information could be communicated by one of them without any contribution from the other.

It is possible that making both participants contribute relevant information could encourage speakers to treat the task as one involving a joint goal, even if the information is available to both of them. For example, using a task similar to the one used in Experiment 6, speakers could be given distractor preambles as well as distractor shapes, and they could be told that imagined listeners would need to identify both the target preamble and the target shape to select the correct scene.

A discussion of the limitations of joint language tasks in their current implementation makes it clear that they differ in fundamental respects from dialogue and natural conversation. First, the interaction between speakers is limited, both quantitatively (in terms of the number of episodes of interaction) and qualitatively (in terms of the degree of constraint imposed on the interaction). Second, it is not clear to what extent speakers in this joint task formed a joint goal. Therefore, one could question the idea, put forward in Section 1.5, that these tasks allow us to uncover the mechanisms used by speakers when they coordinate their utterances in real dialogue.

This is a fair criticism. However, the study of dialogue from a psycholinguistic perspective has been dominated by an approach in which relatively unconstrained interaction is often the norm. Various studies have increased the degree of experimental control, for example by limiting the number and form of participants’ contributions, and by imposing a pre-defined turn-taking structure, or by using confederates. However, these studies still
lacked sufficient experimental control to examine the moment-by-moment coordination of utterances over the time scales normally investigated by monologic psycholinguistic experiments (see Sections 1.4 and 1.5).

To the contrary, joint language tasks are designed to be minimal extensions of these traditional paradigms to a joint task context. As argued in Gambi and Pickering (2013b), one should start with these tasks to understand what mechanisms might be potentially involved in the coordination of utterances, and under what “minimal” conditions (see D. C. Richardson et al., 2012). However, as this section has outlined, future work should focus on progressively enriching these tasks with elements that characterize real dialogue. In this way, it will be possible to investigate directly which of these elements are essential, and what type of mechanisms speakers are more likely to use to coordinate their utterances depending on the nature of the interaction.

5.5.3 From action to language and back again: What can joint language tasks say about theories of acting together?

So far, we have considered the contribution of joint language tasks to theories of talking together (Section 1.4). Let us now consider whether our joint language tasks can inform theories of acting together (Section 1.3) as well. As argued in Chapter 2, Study 1 provided for the first time evidence for a joint interference effect in picture naming. This study extended our understanding of joint interference effects in at least three ways. First, it showed that such effects occur in tasks where stimulus-response mappings are defined by long-term, well-learned associations between objects and their names, rather than by arbitrary relations provided by the experimental instructions. This suggests that representing another’s task has relatively strong effects on the preparation of one’s own response, although it should be noted that the effect was much weaker when participants named single pictures that were not visually degraded (Experiment 3).
Second, Study 1 confirmed that joint interference effects can be observed when participants are seated in separate rooms and have hardly any access to sensory information about their partner’s responses (see Atmaca, et al., 2011). In our study, participants could perceive their partner only in peripheral vision, while auditory feedback was not available. Our findings therefore speak against the referential spatial coding account of joint interference effects (see Section 1.3.2), as the saliency of the partner’s responses was extremely limited in our study.

Third, Study 1 did not support the co-representation account of joint interference effects. Speakers did not appear to represent their partner’s response, but rather they represented whether it was their partner’s turn to name pictures or not. We concluded that this finding is more consistent with the agent conflict account, but this account needs to be revised to explain why speakers experienced more interference when they believed their partner was concurrently naming pictures than when they believed their partner was categorizing pictures.

Importantly, the latter point serves to illustrate the fact that joint language tasks are particularly well suited to answer detailed questions about the nature of other-representations. This is because they capitalize on existing psycholinguistic models of utterance planning and execution, and these models (unlike models of action planning and execution) specify a clear hierarchy of representations and of processing levels that precede overt execution of an utterance. As we have seen, then, it is possible to vary different properties (e.g., semantic relatedness, length) of the responses participants are asked to execute in joint language tasks, and different manipulations will tap into different stages of the response preparation process. In this way, researchers can investigate which processing stages speakers go through when they form representations of others’ utterances.

Finally, the consecutive production paradigms used in Study 2 and Study 3 demonstrates how joint language tasks can contribute to research on the coordination of complementary actions (Section 1.2.3), an area that is still comparatively understudied but
has seen a surge in interest over the past few years. In particular, consecutive production paradigms provide a test-bed for proposals developed in the joint action literature (e.g., the proposal that people reduce the variability in their actions to aid coordination; Vesper et al., 2010, 2011, 2013). Utterances are very complex, hierarchically organized actions. It is still an open question whether the same mechanisms and strategies that appear to underlie coordination of simpler actions (e.g., button presses) can be usefully extended to the coordination of utterances.

5.6 Conclusion

To conclude, this thesis developed and tested a mechanistic account of how speakers coordinate their utterances. In a nutshell, the account proposes that speakers are able to represent and anticipate other speakers’ utterances using some of the same mechanisms they use when producing language. Partial support for this proposal was provided using language tasks that combined traditional psycholinguistic paradigms like picture naming with joint task manipulations. In such tasks, the knowledge that another speaker was about to produce language at the same time as them or just after them affected the way in which speakers produced utterances. However, speakers were not affected by the content and timing of another person’s utterance. Future work should investigate whether this pattern of findings would generalize to tasks that allow more interaction and set a clear joint goal for the speakers, in an attempt to uncover the mechanisms that support utterance coordination in natural conversations.
6. References


# 7. Appendix A: Experimental materials used in Experiments 1 to 8

## Experimental materials used in Experiments 1-4

Table A-1.

Big pictures, their semantic categories, and the small pictures they were paired with in the unrelated and related conditions.

<table>
<thead>
<tr>
<th>Big</th>
<th>Small - unrelated</th>
<th>Small-related</th>
<th>Semantic category</th>
</tr>
</thead>
<tbody>
<tr>
<td>apple</td>
<td>blouse</td>
<td>banana</td>
<td>food</td>
</tr>
<tr>
<td>bed</td>
<td>dress</td>
<td>chair</td>
<td>furniture</td>
</tr>
<tr>
<td>boat</td>
<td>leg</td>
<td>plane</td>
<td>means of transport</td>
</tr>
<tr>
<td>bowl</td>
<td>cake</td>
<td>vase</td>
<td>container</td>
</tr>
<tr>
<td>bread</td>
<td>guitar</td>
<td>cake</td>
<td>food</td>
</tr>
<tr>
<td>cap</td>
<td>vase</td>
<td>dress</td>
<td>clothing</td>
</tr>
<tr>
<td>car</td>
<td>seal</td>
<td>helicopter</td>
<td>means of transport</td>
</tr>
<tr>
<td>cat</td>
<td>pan</td>
<td>fish</td>
<td>animal</td>
</tr>
<tr>
<td>drum</td>
<td>table</td>
<td>guitar</td>
<td>music instrument</td>
</tr>
<tr>
<td>foot</td>
<td>pig</td>
<td>ear</td>
<td>body part</td>
</tr>
<tr>
<td>frog</td>
<td>banana</td>
<td>seahorse</td>
<td>animal</td>
</tr>
<tr>
<td>glass</td>
<td>waistcoat</td>
<td>cup</td>
<td>container</td>
</tr>
<tr>
<td>goat</td>
<td>cup</td>
<td>pig</td>
<td>animal</td>
</tr>
<tr>
<td>hand</td>
<td>seahorse</td>
<td>eye</td>
<td>body part</td>
</tr>
<tr>
<td>horse</td>
<td>trousers</td>
<td>seal</td>
<td>animal</td>
</tr>
<tr>
<td>jug</td>
<td>chair</td>
<td>bottle</td>
<td>container</td>
</tr>
<tr>
<td>knife</td>
<td>helicopter</td>
<td>pan</td>
<td>utensil</td>
</tr>
<tr>
<td>nose</td>
<td>plane</td>
<td>leg</td>
<td>body part</td>
</tr>
<tr>
<td>onion</td>
<td>ear</td>
<td>carrot</td>
<td>food</td>
</tr>
<tr>
<td>pear</td>
<td>bottle</td>
<td>grapes</td>
<td>food</td>
</tr>
<tr>
<td>saw</td>
<td>eye</td>
<td>pliers</td>
<td>utensil</td>
</tr>
<tr>
<td>shoe</td>
<td>fish</td>
<td>trousers</td>
<td>clothing</td>
</tr>
<tr>
<td>skirt</td>
<td>grapes</td>
<td>waistcoat</td>
<td>clothing</td>
</tr>
<tr>
<td>sock</td>
<td>carrot</td>
<td>blouse</td>
<td>clothing</td>
</tr>
<tr>
<td>sofa</td>
<td>pliers</td>
<td>table</td>
<td>furniture</td>
</tr>
</tbody>
</table>
**Experimental materials used in Experiment 5**

Table A-2.

*The 64 pairs of initial and target pictures used on change trial.*

<table>
<thead>
<tr>
<th>Initial</th>
<th>Target</th>
<th>Initial (continued)</th>
<th>Target (continued)</th>
</tr>
</thead>
<tbody>
<tr>
<td>apple</td>
<td>comb</td>
<td>cricket</td>
<td>bottle</td>
</tr>
<tr>
<td>apple</td>
<td>star</td>
<td>cricket</td>
<td>hat</td>
</tr>
<tr>
<td>lion</td>
<td>comb</td>
<td>pencil</td>
<td>bottle</td>
</tr>
<tr>
<td>lion</td>
<td>star</td>
<td>pencil</td>
<td>hat</td>
</tr>
<tr>
<td>arrow</td>
<td>rake</td>
<td>finger</td>
<td>door</td>
</tr>
<tr>
<td>arrow</td>
<td>spoon</td>
<td>finger</td>
<td>top</td>
</tr>
<tr>
<td>helmet</td>
<td>rake</td>
<td>mountain</td>
<td>door</td>
</tr>
<tr>
<td>helmet</td>
<td>spoon</td>
<td>mountain</td>
<td>top</td>
</tr>
<tr>
<td>barrel</td>
<td>scissors</td>
<td>glasses</td>
<td>mouse</td>
</tr>
<tr>
<td>barrel</td>
<td>sledge</td>
<td>glasses</td>
<td>wall</td>
</tr>
<tr>
<td>needle</td>
<td>scissors</td>
<td>orange</td>
<td>mouse</td>
</tr>
<tr>
<td>needle</td>
<td>sledge</td>
<td>orange</td>
<td>wall</td>
</tr>
<tr>
<td>basket</td>
<td>pear</td>
<td>kettle</td>
<td>bone</td>
</tr>
<tr>
<td>basket</td>
<td>trousers</td>
<td>kettle</td>
<td>sack</td>
</tr>
<tr>
<td>hammer</td>
<td>pear</td>
<td>tiger</td>
<td>bone</td>
</tr>
<tr>
<td>hammer</td>
<td>trousers</td>
<td>tiger</td>
<td>sack</td>
</tr>
<tr>
<td>button</td>
<td>arm</td>
<td>lemon</td>
<td>chair</td>
</tr>
<tr>
<td>button</td>
<td>sofa</td>
<td>lemon</td>
<td>saw</td>
</tr>
<tr>
<td>chicken</td>
<td>arm</td>
<td>mushroom</td>
<td>chair</td>
</tr>
<tr>
<td>chicken</td>
<td>sofa</td>
<td>mushroom</td>
<td>saw</td>
</tr>
<tr>
<td>camel</td>
<td>axe</td>
<td>monkey</td>
<td>scooter</td>
</tr>
<tr>
<td>camel</td>
<td>tie</td>
<td>monkey</td>
<td>steps</td>
</tr>
<tr>
<td>sandwich</td>
<td>axe</td>
<td>rabbit</td>
<td>scooter</td>
</tr>
<tr>
<td>sandwich</td>
<td>tie</td>
<td>rabbit</td>
<td>steps</td>
</tr>
<tr>
<td>candle</td>
<td>nose</td>
<td>onion</td>
<td>vase</td>
</tr>
<tr>
<td>candle</td>
<td>sausage</td>
<td>onion</td>
<td>whistle</td>
</tr>
<tr>
<td>ladder</td>
<td>nose</td>
<td>ruler</td>
<td>vase</td>
</tr>
<tr>
<td>ladder</td>
<td>sausage</td>
<td>ruler</td>
<td>whistle</td>
</tr>
<tr>
<td>cannon</td>
<td>foot</td>
<td>spider</td>
<td>boat</td>
</tr>
<tr>
<td>cannon</td>
<td>pliers</td>
<td>spider</td>
<td>mouth</td>
</tr>
<tr>
<td>window</td>
<td>foot</td>
<td>suitcase</td>
<td>boat</td>
</tr>
<tr>
<td>window</td>
<td>pliers</td>
<td>suitcase</td>
<td>mouth</td>
</tr>
</tbody>
</table>
### Experimental materials used in Experiment 6

Table A-3.

The 20 preambles and the abstract shapes paired with each of them.

<table>
<thead>
<tr>
<th>Preamble</th>
<th>Shapes</th>
</tr>
</thead>
<tbody>
<tr>
<td>The soldier chases</td>
<td><img src="image1" alt="Shapes" /></td>
</tr>
<tr>
<td>The pirate follows</td>
<td><img src="image2" alt="Shapes" /></td>
</tr>
<tr>
<td>The cowboy punches</td>
<td><img src="image3" alt="Shapes" /></td>
</tr>
<tr>
<td>The painter hits</td>
<td><img src="image4" alt="Shapes" /></td>
</tr>
</tbody>
</table>
The nun sees

The policeman chooses

The waitress touches

The cook pushes
The painter shoots

The policeman follows

The sailor chases

The pirate punches
The waitress hits

The ballerina pushes

The cowboy touches

The nun shoots
The cook chooses

The sailor welcomes

The ballerina welcomes

The soldier sees
### Experimental materials used in Experiment 7

Table A-4.

*Agents, actions, and patients used in the 32 pictured scenes.*

<table>
<thead>
<tr>
<th>Agent</th>
<th>Action</th>
<th>Patient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nun</td>
<td>Follow</td>
<td>Doctor</td>
</tr>
<tr>
<td>Boxer</td>
<td>Chase</td>
<td>Nun</td>
</tr>
<tr>
<td>Cowboy</td>
<td>Hit</td>
<td>Burglar</td>
</tr>
<tr>
<td>Doctor</td>
<td>Punch</td>
<td>Pirate</td>
</tr>
<tr>
<td>Waitress</td>
<td>Shoot</td>
<td>Ballerina</td>
</tr>
<tr>
<td>Burglar</td>
<td>Kill</td>
<td>Soldier</td>
</tr>
<tr>
<td>Priest</td>
<td>Touch</td>
<td>Cook</td>
</tr>
<tr>
<td>Sailor</td>
<td>Tickle</td>
<td>Cowboy</td>
</tr>
<tr>
<td>Policeman</td>
<td>Follow</td>
<td>Sailor</td>
</tr>
<tr>
<td>Doctor</td>
<td>Chase</td>
<td>Nun</td>
</tr>
<tr>
<td>Cowboy</td>
<td>Hit</td>
<td>Clown</td>
</tr>
<tr>
<td>Burglar</td>
<td>Punch</td>
<td>Cook</td>
</tr>
<tr>
<td>Waitress</td>
<td>Shoot</td>
<td>Boxer</td>
</tr>
<tr>
<td>Painter</td>
<td>Kill</td>
<td>Ballerina</td>
</tr>
<tr>
<td>Sailor</td>
<td>Tickle</td>
<td>Burglar</td>
</tr>
<tr>
<td>Cook</td>
<td>Touch</td>
<td>Waitress</td>
</tr>
<tr>
<td>Nun</td>
<td>Follow</td>
<td>Pirate</td>
</tr>
<tr>
<td>Swimmer</td>
<td>Kill</td>
<td>Policeman</td>
</tr>
<tr>
<td>Boxer</td>
<td>Chase</td>
<td>Soldier</td>
</tr>
<tr>
<td>Pirate</td>
<td>Punch</td>
<td>Swimmer</td>
</tr>
<tr>
<td>Ballerina</td>
<td>Shoot</td>
<td>Doctor</td>
</tr>
<tr>
<td>Soldier</td>
<td>Hit</td>
<td>Priest</td>
</tr>
<tr>
<td>Clown</td>
<td>Tickle</td>
<td>Pirate</td>
</tr>
<tr>
<td>Priest</td>
<td>Touch</td>
<td>Cowboy</td>
</tr>
<tr>
<td>Policeman</td>
<td>Follow</td>
<td>Sailor</td>
</tr>
<tr>
<td>Painter</td>
<td>Hit</td>
<td>Clown</td>
</tr>
<tr>
<td>Cook</td>
<td>Tickle</td>
<td>Painter</td>
</tr>
<tr>
<td>Clown</td>
<td>Punch</td>
<td>Boxer</td>
</tr>
<tr>
<td>Ballerina</td>
<td>Shoot</td>
<td>Waitress</td>
</tr>
<tr>
<td>Soldier</td>
<td>Kill</td>
<td>Swimmer</td>
</tr>
<tr>
<td>Swimmer</td>
<td>Chase</td>
<td>Policeman</td>
</tr>
<tr>
<td>Pirate</td>
<td>Touch</td>
<td>Priest</td>
</tr>
</tbody>
</table>
Experimental materials used in Experiment 8

Table A-5.

*Easy and hard versions of the 21 items used in Experiment 8 (the same materials were used in Experiment 8a and Experiment 8b).*

<table>
<thead>
<tr>
<th>Item</th>
<th>Easy</th>
<th>Hard</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The soldier chases the swimmer</td>
<td>The soldier chases the swimmer with the</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cane and the vase</td>
</tr>
<tr>
<td>2</td>
<td>The pirate follows the burglar</td>
<td>The pirate follows the burglar with the</td>
</tr>
<tr>
<td></td>
<td></td>
<td>axe and the guitar</td>
</tr>
<tr>
<td>3</td>
<td>The painter hits the ballerina</td>
<td>The painter hits the ballerina with the</td>
</tr>
<tr>
<td></td>
<td></td>
<td>racket and the flower</td>
</tr>
<tr>
<td>4</td>
<td>The pirate punches the clown</td>
<td>The pirate punches the clown with the</td>
</tr>
<tr>
<td></td>
<td></td>
<td>apple and the bell</td>
</tr>
<tr>
<td>5</td>
<td>The painter shoots the policeman</td>
<td>The painter shoots the policeman with</td>
</tr>
<tr>
<td></td>
<td></td>
<td>the kite and the bottle</td>
</tr>
<tr>
<td>6</td>
<td>The pirate kisses the doctor</td>
<td>The pirate kisses the doctor with the</td>
</tr>
<tr>
<td></td>
<td></td>
<td>trumpet and the broom</td>
</tr>
<tr>
<td>7</td>
<td>The cowboy scolds the ballerina</td>
<td>The cowboy scolds the ballerina with the</td>
</tr>
<tr>
<td></td>
<td></td>
<td>drum and the wand</td>
</tr>
<tr>
<td>8</td>
<td>The policeman chases the monk</td>
<td>The policeman chases the monk with the</td>
</tr>
<tr>
<td></td>
<td></td>
<td>basket and the sailor</td>
</tr>
<tr>
<td>9</td>
<td>The sailor follows the soldier</td>
<td>The sailor follows the soldier with the</td>
</tr>
<tr>
<td></td>
<td></td>
<td>lamp and the nun</td>
</tr>
<tr>
<td>10</td>
<td>The waitress hits the swimmer</td>
<td>The waitress hits the swimmer with the</td>
</tr>
<tr>
<td></td>
<td></td>
<td>crown and the boxer</td>
</tr>
<tr>
<td>11</td>
<td>The cowboy punches the doctor</td>
<td>The cowboy punches the doctor with the</td>
</tr>
<tr>
<td></td>
<td></td>
<td>balloon and the swimmer</td>
</tr>
<tr>
<td>12</td>
<td>The waitress shoots the clown</td>
<td>The waitress shoots the clown with the</td>
</tr>
<tr>
<td></td>
<td></td>
<td>violin and the pirate</td>
</tr>
<tr>
<td>13</td>
<td>The ballerina kisses the waitress</td>
<td>The ballerina kisses the waitress with the</td>
</tr>
<tr>
<td></td>
<td></td>
<td>owl and the soldier</td>
</tr>
<tr>
<td>14</td>
<td>The monk scolds the soldier</td>
<td>The monk scolds the soldier with the</td>
</tr>
<tr>
<td></td>
<td></td>
<td>bow and the doctor</td>
</tr>
<tr>
<td>15</td>
<td>The swimmer chases the burglar</td>
<td>The swimmer chases the burglar and the</td>
</tr>
<tr>
<td></td>
<td></td>
<td>nun with the hanger</td>
</tr>
<tr>
<td>16</td>
<td>The doctor follows the nun</td>
<td>The doctor follows the nun and the</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sailor with the candle</td>
</tr>
<tr>
<td>17</td>
<td>The cowboy hits the sailor</td>
<td>The cowboy hits the sailor and the</td>
</tr>
<tr>
<td></td>
<td></td>
<td>burglar with the ladder</td>
</tr>
<tr>
<td>18</td>
<td>The painter punches the clown</td>
<td>The painter punches the clown and the</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ballerina with the hammer</td>
</tr>
<tr>
<td>19</td>
<td>The nun shoots the burglar</td>
<td>The nun shoots the burglar and the</td>
</tr>
<tr>
<td></td>
<td></td>
<td>clown with the hat</td>
</tr>
<tr>
<td>20</td>
<td>The nun kisses the monk</td>
<td>The nun kisses the monk and the doctor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>with the umbrella</td>
</tr>
<tr>
<td>21</td>
<td>The policeman scolds the sailor</td>
<td>The policeman scolds the sailor and the</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ballerina with the flag</td>
</tr>
</tbody>
</table>
8. Appendix B: Full linear mixed effect analyses for
Experiments 1-4 and 6-8 and additional analyses for
Experiment 5.

Best-fitting linear mixed-effects models for Experiments 1-4, and
the combined analyses of Experiments 3 and 4.

Table B-1.

*Best fitting model for the accuracy data in Experiment 1.*

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Estimate</th>
<th>SE</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-3.10</td>
<td>.18</td>
<td>-16.97</td>
</tr>
<tr>
<td>Partner1: naming vs. NO</td>
<td>.24</td>
<td>.11</td>
<td>2.23</td>
</tr>
<tr>
<td>Partner2: SAME vs. DIFFERENT</td>
<td>-.23</td>
<td>.08</td>
<td>-2.75</td>
</tr>
<tr>
<td>Relatedness: Related vs. Unrelated</td>
<td>-.31</td>
<td>.15</td>
<td>-2.05</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random effect</th>
<th>Explained variance estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subjects: intercept</td>
<td>.48</td>
</tr>
<tr>
<td>Items: intercept</td>
<td>.48</td>
</tr>
<tr>
<td>Items: Relatedness</td>
<td>.56</td>
</tr>
</tbody>
</table>

Table B-2.

*Best fitting model for the voice onset time data in Experiment 1.*

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Estimate</th>
<th>SE</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>874</td>
<td>24</td>
<td>36.72</td>
</tr>
<tr>
<td>Partner1: naming vs. NO</td>
<td>14</td>
<td>5</td>
<td>2.79</td>
</tr>
<tr>
<td>Partner2: SAME vs. DIFFERENT</td>
<td>1</td>
<td>4</td>
<td>.17</td>
</tr>
<tr>
<td>Relatedness (related vs. unrelated)</td>
<td>16</td>
<td>5</td>
<td>3.36</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random effect</th>
<th>Explained variance estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subjects: intercept</td>
<td>11980</td>
</tr>
<tr>
<td>Items: intercept</td>
<td>3150</td>
</tr>
</tbody>
</table>

*Note.* Variance estimates have been rounded up to the nearest ten in all the analyses of
naming latencies reported in this paper (the analyses were run on latencies measured in
seconds, and then the estimates were converted to milliseconds for expository purposes).
Table B-3.

**Best fitting model for the accuracy data in Experiment 2.**

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Estimate</th>
<th>SE</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-3.16</td>
<td>.17</td>
<td>-18.16</td>
</tr>
<tr>
<td>Relatedness: Related vs. Unrelated</td>
<td>.24</td>
<td>.13</td>
<td>1.86</td>
</tr>
</tbody>
</table>

Random effect

<table>
<thead>
<tr>
<th>Explained variance estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subjects: intercept</td>
</tr>
<tr>
<td>Items: intercept</td>
</tr>
<tr>
<td>Items: Relatedness</td>
</tr>
</tbody>
</table>

Table B-4.

**Best fitting model for the voice onset time data in Experiment 2.**

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Estimate</th>
<th>SE</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>884</td>
<td>24</td>
<td>36.77</td>
</tr>
<tr>
<td>Partner1: naming vs. CAT</td>
<td>12</td>
<td>5</td>
<td>2.47</td>
</tr>
<tr>
<td>Partner2: SAME vs. DIFFERENT</td>
<td>3</td>
<td>4</td>
<td>.70</td>
</tr>
<tr>
<td>Relatedness (related vs. unrelated)</td>
<td>19</td>
<td>5</td>
<td>3.48</td>
</tr>
</tbody>
</table>

Random effect

<table>
<thead>
<tr>
<th>Explained variance estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subjects: intercept</td>
</tr>
<tr>
<td>Subjects: Size</td>
</tr>
<tr>
<td>Items: intercept</td>
</tr>
<tr>
<td>Items: Relatedness</td>
</tr>
</tbody>
</table>

Table B-5.

**Best fitting model for the voice onset time data in Experiment 3.**

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Estimate</th>
<th>SE</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>702</td>
<td>18</td>
<td>39.87</td>
</tr>
<tr>
<td>Partner1: naming vs. NO</td>
<td>7</td>
<td>4</td>
<td>1.88</td>
</tr>
<tr>
<td>Partner2: SAME vs. DIFFERENT</td>
<td>3</td>
<td>3</td>
<td>.91</td>
</tr>
<tr>
<td>Relatedness (related vs. unrelated)</td>
<td>3</td>
<td>5</td>
<td>.59</td>
</tr>
<tr>
<td>Size (small vs. big)</td>
<td>21</td>
<td>16</td>
<td>1.34</td>
</tr>
<tr>
<td>Relatedness: Size</td>
<td>22</td>
<td>10</td>
<td>2.19</td>
</tr>
</tbody>
</table>

Random effects structure

<table>
<thead>
<tr>
<th>Explained variance estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject: intercept</td>
</tr>
<tr>
<td>Subject: Size</td>
</tr>
<tr>
<td>Item: intercept</td>
</tr>
<tr>
<td>Item: Relatedness</td>
</tr>
</tbody>
</table>
Table B-6.  
*Best fitting model for the accuracy data in Experiment 4.*

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Estimate</th>
<th>SE</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-3.88</td>
<td>.21</td>
<td>-18.18</td>
</tr>
<tr>
<td>Partner1: naming vs. NO</td>
<td>.05</td>
<td>.13</td>
<td>.39</td>
</tr>
<tr>
<td>Partner2: SAME vs. DIFFERENT</td>
<td>-.08</td>
<td>.11</td>
<td>-.73</td>
</tr>
<tr>
<td>Size: small vs. big</td>
<td>.10</td>
<td>.40</td>
<td>.24</td>
</tr>
<tr>
<td>Partner1: Size</td>
<td>-.63</td>
<td>.27</td>
<td>-2.35</td>
</tr>
<tr>
<td>Partner2: Size</td>
<td>.47</td>
<td>.21</td>
<td>2.18</td>
</tr>
</tbody>
</table>

Random structure

<table>
<thead>
<tr>
<th>Explained variance estimate</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject: intercept</td>
<td>.39</td>
</tr>
<tr>
<td>Subject: Size</td>
<td>.98</td>
</tr>
<tr>
<td>Item: intercept</td>
<td>1.10</td>
</tr>
</tbody>
</table>

Table B-7.  
*Best fitting model for the voice onset time data in Experiment 4.*

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Estimate</th>
<th>SE</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>816.00</td>
<td>21</td>
<td>39.56</td>
</tr>
<tr>
<td>Partner1: naming vs. NO</td>
<td>15.00</td>
<td>5</td>
<td>3.35</td>
</tr>
<tr>
<td>Partner2: SAME vs. DIFFERENT</td>
<td>-.01</td>
<td>4</td>
<td>-.13</td>
</tr>
</tbody>
</table>

Random effects structure

<table>
<thead>
<tr>
<th>Explained Variance estimate</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject: intercept</td>
<td>8090</td>
</tr>
<tr>
<td>Subject: Size</td>
<td>690</td>
</tr>
<tr>
<td>Item: intercept</td>
<td>4220</td>
</tr>
</tbody>
</table>
Table B-8.  
*Best fitting model to latency data in Experiments 3 and 4 (combined analysis).*

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Estimate</th>
<th>SE</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>759</td>
<td>13</td>
<td>57.56</td>
</tr>
<tr>
<td>Partner1: naming vs. NO</td>
<td>11</td>
<td>3</td>
<td>3.51</td>
</tr>
<tr>
<td>Partner2: SAME vs. DIFFERENT</td>
<td>1</td>
<td>3</td>
<td>.47</td>
</tr>
<tr>
<td>Relatedness: related vs. unrelated</td>
<td>4</td>
<td>4</td>
<td>1.00</td>
</tr>
<tr>
<td>Size: small vs. big</td>
<td>22</td>
<td>16</td>
<td>1.39</td>
</tr>
<tr>
<td>Experiment: 4 vs. 3</td>
<td>112</td>
<td>5</td>
<td>21.74</td>
</tr>
<tr>
<td>Relatedness: Size</td>
<td>18</td>
<td>9</td>
<td>2.07</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random effects structure</th>
<th>Explained variance estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject: intercept</td>
<td>2140</td>
</tr>
<tr>
<td>Subject: Size</td>
<td>1100</td>
</tr>
<tr>
<td>Item: intercept</td>
<td>2790</td>
</tr>
<tr>
<td>Item: Relatedness</td>
<td>530</td>
</tr>
<tr>
<td>Item: Experiment</td>
<td>890</td>
</tr>
</tbody>
</table>

*Note.* By-participant random slopes for Size ($\chi^2 (3) = 13.91, p<.005$); by-items random slopes for Relatedness ($\chi^2 (3) = 19.57, p<.001$) and Experiment ($\chi^2 (3) = 45.42, p<.001$). These slopes were selected using a forward method since very complex models would not converge due to the number of factors involved. The model did not include correlations among random slopes and intercepts, unlike previous analyses.

**Supplementary analyses for Initial Onset, Target Onset, and Interval in Experiment 5**

For the analyses of Target Onset and Interval (in the SELF and OTHER conditions), we excluded all trials in which the initial or the target word was named incorrectly or produced disfluently (14.7% in SELF, 15.0% in OTHER). For Initial Onset and Interval, we also excluded all trials with skipped initial responses (but we included interrupted responses, as we were interested in the effect of Response Type; see Table 3-2). In the OTHER condition, it was possible to have negative values for Interval (overlap between the two participants’ responses). Such cases (6.5%) were also excluded from the analyses of Target Onset and Interval in the OTHER condition for the sake of comparability between conditions. For each condition separately, we then removed all trials that were more than 2.5 SD from the grand mean or more than 3 SD from the by-subject mean from the analyses (initial onset: 3.0% in SELF, 2.7% in OTHER, 3.0% in NO; target onset: 3.4% in SELF, 2.9% in OTHER; initial-target interval: 4.9% in SELF, 2.7% in OTHER).
The starting point was the model including the factors Condition, Degradation, Response Type, all the possible two-way interactions, and the three-way interaction. We set the SELF condition as the reference level, and we defined two contrasts, one comparing the weighted average of OTHER and NO to SELF (Condition 1), and the other comparing OTHER against NO (Condition 2). We used weighted contrast coding for all predictors. So, for example, the contrast for the two-level factor Response Type was not (-.5, .5), but rather (-.5*n_i/N; .5*n_c/N), where n_i is the count of interrupted responses, n_c is the count of completed responses and N = n_i + n_c (completed responses were taken as the reference level). In this way, the intercept corresponds to the weighted grand mean and the estimates for the main effects are equal to twice the difference between levels of the corresponding factor.

Table B-9.

*Initial Onset by Initial Response Type and Degradation in the three conditions.*

<table>
<thead>
<tr>
<th>Degradation</th>
<th>Completed</th>
<th>Interrupted</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>SELF</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Degraded</td>
<td>703 (112)[396]</td>
<td>726 (117)[29]</td>
<td>704 (112)[425]</td>
</tr>
<tr>
<td>Intact</td>
<td>698 (117)[383]</td>
<td>783 (119)[32]</td>
<td>704 (118)[415]</td>
</tr>
<tr>
<td>Total</td>
<td>700 (114)[779]</td>
<td>756 (121)[61]</td>
<td>704 (115)[840]</td>
</tr>
<tr>
<td>OTHER</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Degraded</td>
<td>718 (109)[301]</td>
<td>733 (103)[97]</td>
<td>721 (107)[398]</td>
</tr>
<tr>
<td>Intact</td>
<td>714 (107)[289]</td>
<td>774 (128)[111]</td>
<td>730 (116)[400]</td>
</tr>
<tr>
<td>Total</td>
<td>717 (108)[590]</td>
<td>755 (118)[208]</td>
<td>726 (112)[798]</td>
</tr>
<tr>
<td>NO</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Degraded</td>
<td>709 (134)[196]</td>
<td>711 (110)[122]</td>
<td>710 (125)[318]</td>
</tr>
<tr>
<td>Intact</td>
<td>698 (120)[210]</td>
<td>698 (90)[106]</td>
<td>698 (111)[316]</td>
</tr>
<tr>
<td>Total</td>
<td>703 (127)[406]</td>
<td>705 (101)[228]</td>
<td>704 (118)[634]</td>
</tr>
</tbody>
</table>

*Note.* Mean values in ms (standard deviation within round brackets)[cell count in square brackets].
Table B-10.

Initial Onset analyses: complete model.

<table>
<thead>
<tr>
<th>Fixed effect</th>
<th>Coef.</th>
<th>SE</th>
<th>t</th>
<th>χ²</th>
<th>p</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept (weighted grand mean)</td>
<td>718.46</td>
<td>7.41</td>
<td>96.99</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Response Type (Interrupted – Completed)</td>
<td>89.55</td>
<td>12.55</td>
<td>7.13</td>
<td>48.78</td>
<td>&lt;.001</td>
<td>1</td>
</tr>
<tr>
<td>Degradation (Intact – Degraded)</td>
<td>3.89</td>
<td>8.88</td>
<td>0.44</td>
<td>0.002</td>
<td>.97</td>
<td>1</td>
</tr>
<tr>
<td>Cond1: (OTHER,NO) – SELF</td>
<td>2.67</td>
<td>24.73</td>
<td>0.11</td>
<td>2.75</td>
<td>.25</td>
<td>2</td>
</tr>
<tr>
<td>Cond2: OTHER - NO</td>
<td>41.86</td>
<td>31.38</td>
<td>1.33</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Response Type * Degradation</td>
<td>113.64</td>
<td>43.29</td>
<td>2.62</td>
<td>4.93</td>
<td>&lt;.05</td>
<td>1</td>
</tr>
<tr>
<td>Response Type * Condition1</td>
<td>-79.79</td>
<td>59.63</td>
<td>-1.34</td>
<td>7.48</td>
<td>&lt;.05</td>
<td>2</td>
</tr>
<tr>
<td>Response Type * Condition2</td>
<td>123.16</td>
<td>53.88</td>
<td>2.29</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Degradation*Condition1</td>
<td>-33.79</td>
<td>34.49</td>
<td>-0.98</td>
<td>4.34</td>
<td>.11</td>
<td>2</td>
</tr>
<tr>
<td>Degradation*Condition2</td>
<td>69.03</td>
<td>42.53</td>
<td>1.62</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RespType<em>Degr</em>Cond1</td>
<td>-391.35</td>
<td>212.00</td>
<td>-1.85</td>
<td>5.97</td>
<td>.05</td>
<td>2</td>
</tr>
<tr>
<td>RespType<em>Degr</em>Cond2</td>
<td>279.9</td>
<td>183.22</td>
<td>1.53</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Random effect

<table>
<thead>
<tr>
<th></th>
<th>Variance</th>
<th>Corr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject: Intercept</td>
<td>3322.21</td>
<td>NA</td>
</tr>
<tr>
<td>Item: Intercept</td>
<td>752.77</td>
<td>NA</td>
</tr>
</tbody>
</table>

The onset of the initial picture name (see Tables B-9 and B-10) was delayed for interrupted initials compared to completed initials. There was also some indication of the effect being larger in the SELF (756 vs. 700 ms) and OTHER (755 vs. 716 ms) conditions than in the NO condition (703 vs. 705 ms), as indicated by a Response Type by Condition interaction. Finally, the effect of Response Type was larger before intact than before degraded targets in the SELF (85 vs. 23 ms) and the OTHER (60 vs. 15 ms) conditions (see significant Response Type by Degradation interaction and the marginal three-way interaction of Response Type, Degradation, and Condition). Hartsuiker et al. (2008) also reported longer onset times for interrupted than for completed initials (though only in their Experiment 2), and suggested that it was due to the stopping process being more likely to stop word-internally when the initial word is initiated later.
Table B-11.  
*Interval by Initial Response Type and Degradation in the SELF and OTHER conditions.*

<table>
<thead>
<tr>
<th>Degradation</th>
<th>Completed</th>
<th>Interrupted</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>SELF</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Degraded</td>
<td>102 (126)[316]</td>
<td>191 (159)[21]</td>
<td>107 (130)[337]</td>
</tr>
<tr>
<td>Intact</td>
<td>81 (121)[352]</td>
<td>112 (93)[27]</td>
<td>83 (119)[379]</td>
</tr>
<tr>
<td>Total</td>
<td>91 (123)[668]</td>
<td>146 (131)[48]</td>
<td>94 (124)[716]</td>
</tr>
<tr>
<td>OTHER</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intact</td>
<td>235 (224)[226]</td>
<td>303 (228)[88]</td>
<td>254 (227)[314]</td>
</tr>
<tr>
<td>Total</td>
<td>268 (232)[441]</td>
<td>310 (221)[175]</td>
<td>280 (230)[616]</td>
</tr>
</tbody>
</table>

*Note.* Mean values in ms (standard deviation within round brackets)[cell count in square brackets].

Table B-12.  
*Interval analyses: complete model.*

<table>
<thead>
<tr>
<th>Fixed effect</th>
<th>Coef.</th>
<th>SE</th>
<th>t</th>
<th>$\chi^2$</th>
<th>p</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept (weighted grand mean)</td>
<td>184.46</td>
<td>13.05</td>
<td>14.14</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Response Type (Interrupted – Completed)</td>
<td>52.62</td>
<td>32.76</td>
<td>1.61</td>
<td>1.11</td>
<td>.29</td>
<td>1</td>
</tr>
<tr>
<td>Degradation (Intact – Degraded)</td>
<td>-88.27</td>
<td>18.18</td>
<td>-4.86</td>
<td>19.86</td>
<td>&lt;.001</td>
<td>1</td>
</tr>
<tr>
<td>Condition (SELF-OTHER)</td>
<td>-342.65</td>
<td>41.37</td>
<td>-8.28</td>
<td>48.90</td>
<td>&lt;.001</td>
<td>1</td>
</tr>
<tr>
<td>Response Type*Degradation</td>
<td>-86.66</td>
<td>116.16</td>
<td>-0.75</td>
<td>0.06</td>
<td>.80</td>
<td>1</td>
</tr>
<tr>
<td>Response Type * Condition</td>
<td>151.73</td>
<td>125.16</td>
<td>1.21</td>
<td>1.19</td>
<td>.28</td>
<td>1</td>
</tr>
<tr>
<td>Degradation*Condition</td>
<td>90.50</td>
<td>72.62</td>
<td>1.25</td>
<td>2.28</td>
<td>.13</td>
<td>1</td>
</tr>
<tr>
<td>RespType<em>Degr</em>Cond</td>
<td>-818.38</td>
<td>446.52</td>
<td>-1.83</td>
<td>3.38</td>
<td>.07</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random effect</th>
<th>Variance</th>
<th>Corr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject: Intercept</td>
<td>4827.6</td>
<td>NA</td>
</tr>
<tr>
<td>Item: Intercept</td>
<td>4615.6</td>
<td></td>
</tr>
<tr>
<td>Item: Condition</td>
<td>9915.1</td>
<td>-1.00</td>
</tr>
</tbody>
</table>

The Interval between the offset of the initial name and the onset of the target name (in the SELF and OTHER conditions) was longer before degraded than before intact targets and longer in the OTHER than in the SELF condition (see Tables B-11 and B-12). The three-
way interaction of Response Type, Degradation, and Condition was marginal, suggesting that the effect of Degradation is larger before interrupted (79 ms) than before completed (11 ms) responses in the SELF condition but not in the OTHER condition (14 vs. 69 ms). Hartsuiker et al. (2008) reported a non-significant trend in the same direction in their Experiment 1.

In addition, in the SELF condition we observed a numerical trend for longer intervals after interrupted (146 ms) than after completed (91 ms) initial names. Hartsuiker et al. (2008) reported a significant difference in the same direction. According to them, this is because participants have more time to plan the target name while still articulating the initial name when they complete than when they do not. The difference was smaller in our experiment (55 ms) than in Hartsuiker et al. (150 ms), perhaps because intervals after interruptions were shorter in our experiment than theirs (146 vs. 216 ms). This might depend, in turn, on our participants’ reduced propensity to interrupt (see Section 3.4).

Table B-13.
*Target Onset by Initial Response Type and Degradation in the SELF and OTHER conditions.*

<table>
<thead>
<tr>
<th>Degradation</th>
<th>Completed</th>
<th>Interrupted</th>
<th>Skipped</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SELF</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>785 (187)[678]</td>
<td>760 (153)[52]</td>
<td>878 (184)[75]</td>
<td>792 (187)[805]</td>
</tr>
<tr>
<td><strong>OTHER</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Degraded</td>
<td>959 (217)[212]</td>
<td>852 (191)[87]</td>
<td>887 (216)[59]</td>
<td>921 (216)[358]</td>
</tr>
<tr>
<td>Intact</td>
<td>892 (216)[225]</td>
<td>853 (215)[87]</td>
<td>807 (207)[67]</td>
<td>868 (216)[379]</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>924 (219)[437]</td>
<td>852 (202)[174]</td>
<td>844 (214)[126]</td>
<td>894 (217)[737]</td>
</tr>
</tbody>
</table>

*Note.* Mean values in ms (standard deviation within round brackets)[cell count in square brackets].
Table B-14.

**Target Onset analyses: complete model.**

<table>
<thead>
<tr>
<th>Fixed effect</th>
<th>Coef.</th>
<th>SE</th>
<th>t</th>
<th>Χ²</th>
<th>p</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept (weighted grand mean)</td>
<td>851.36</td>
<td>15.16</td>
<td>56.14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RespType1: (Skipped, Interrupted) – Completed</td>
<td>-61.38</td>
<td>29.76</td>
<td>-2.06</td>
<td>11.56</td>
<td>&lt;.01</td>
<td>2</td>
</tr>
<tr>
<td>RespType2: Skipped - Interrupted</td>
<td>94.24</td>
<td>39.00</td>
<td>2.42</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Degradation (Intact – Degraded)</td>
<td>-91.18</td>
<td>24.53</td>
<td>-3.72</td>
<td>15.29</td>
<td>&lt;.001</td>
<td>1</td>
</tr>
<tr>
<td>Condition (SELF-OTHER)</td>
<td>-186.80</td>
<td>44.98</td>
<td>-4.15</td>
<td>19.99</td>
<td>&lt;.001</td>
<td>1</td>
</tr>
<tr>
<td>Response Type1 * Degradation</td>
<td>34.32</td>
<td>104.57</td>
<td>0.33</td>
<td>1.45</td>
<td>.48</td>
<td>2</td>
</tr>
<tr>
<td>Response Type2 * Degradation</td>
<td>-82.77</td>
<td>151.05</td>
<td>-0.55</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Response Type1*Condition</td>
<td>446.57</td>
<td>115.59</td>
<td>3.86</td>
<td>19.50</td>
<td>&lt;.001</td>
<td>2</td>
</tr>
<tr>
<td>Response Type2*Condition</td>
<td>247.74</td>
<td>151.53</td>
<td>1.63</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Degradation*Condition</td>
<td>64.23</td>
<td>85.92</td>
<td>0.75</td>
<td>0.67</td>
<td>.41</td>
<td>1</td>
</tr>
<tr>
<td>RespType1<em>Degr</em>Cond</td>
<td>60.95</td>
<td>407.63</td>
<td>0.15</td>
<td>1.23</td>
<td>.54</td>
<td>2</td>
</tr>
<tr>
<td>RespType2<em>Degr</em>Cond</td>
<td>639.09</td>
<td>591.67</td>
<td>591.67</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random effect</th>
<th>Variance</th>
<th>Corr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject: Intercept</td>
<td>6449.8</td>
<td></td>
</tr>
<tr>
<td>Subject: Degradation</td>
<td>5209.1</td>
<td>.61</td>
</tr>
<tr>
<td>Item: Intercept</td>
<td>6540.6</td>
<td></td>
</tr>
<tr>
<td>Item: Degradation</td>
<td>8159.6</td>
<td>-.27</td>
</tr>
</tbody>
</table>

Finally, the onset of the target picture name was delayed for degraded with respect to intact targets in both the SELF (809 vs. 777ms) and the OTHER condition (921 vs. 868ms), indicating that the manipulation was effective in both conditions (see Tables B-13 and B-14). Target onset latencies were also significantly longer in the OTHER (894 ms) than in the SELF condition (792 ms). Target onsets varied as a function of initial Response Type, but differently in the two conditions. In the SELF conditions, latencies were much longer after skipped initials (878 ms) than after completed (785 ms) or interrupted (760 ms) initials. In the OTHER condition, instead, the target was named faster when the initial was interrupted (852 ms) or skipped (844 ms) than when it was completed (924 ms).
Best-fitting linear mixed-effects models for Experiments 6-8.

Table B-15.  
*Best fitting model for the likelihood of producing a disfluency before or during the preamble in Experiment 6.*

<table>
<thead>
<tr>
<th>Fixed effect</th>
<th>Estimate</th>
<th>SE</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-3.024</td>
<td>.274</td>
<td>-11.03</td>
</tr>
<tr>
<td>Partner (SELF vs. OTHER)</td>
<td>1.840</td>
<td>.485</td>
<td>3.79</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random effect</th>
<th>Variance</th>
<th>Correlations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant: Intercept</td>
<td>1.925</td>
<td></td>
</tr>
<tr>
<td>Participant: Difficulty</td>
<td>.164</td>
<td>.189</td>
</tr>
<tr>
<td>Preamble: Intercept</td>
<td>.430</td>
<td></td>
</tr>
<tr>
<td>Preamble: Difficulty</td>
<td>.477</td>
<td>.075</td>
</tr>
<tr>
<td>Preamble: Partner</td>
<td>.384</td>
<td>-.175</td>
</tr>
<tr>
<td>Preamble: Difficulty*Partner</td>
<td>.871</td>
<td>.415</td>
</tr>
</tbody>
</table>

Table B-16.  
*Best fitting model for Onset in Experiment 6.*

<table>
<thead>
<tr>
<th>Fixed effect</th>
<th>Estimate</th>
<th>SE</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.177</td>
<td>.037</td>
<td>32.12</td>
</tr>
<tr>
<td>Partner (SELF vs. OTHER)</td>
<td>.201</td>
<td>.065</td>
<td>3.08</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random effect</th>
<th>Variance</th>
<th>Correlations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant: Intercept</td>
<td>.046</td>
<td></td>
</tr>
<tr>
<td>Participant: Difficulty</td>
<td>.002</td>
<td>.214</td>
</tr>
<tr>
<td>Preamble: Intercept</td>
<td>.007</td>
<td></td>
</tr>
<tr>
<td>Preamble: Difficulty</td>
<td>.004</td>
<td>-.654</td>
</tr>
<tr>
<td>Preamble: Partner</td>
<td>.005</td>
<td>-.993</td>
</tr>
<tr>
<td>Preamble: Difficulty*Partner</td>
<td>.001</td>
<td>.438</td>
</tr>
</tbody>
</table>
Table B-17.  

*Best fitting model for Duration in Experiment 6.*  

<table>
<thead>
<tr>
<th>Fixed effect</th>
<th>Estimate</th>
<th>SE</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.377</td>
<td>.032</td>
<td>42.38</td>
</tr>
<tr>
<td>Partner (SELF vs. OTHER)</td>
<td>.246</td>
<td>.052</td>
<td>4.74</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random effect</th>
<th>Variance</th>
<th>Correlations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant: Intercept</td>
<td>.035</td>
<td></td>
</tr>
<tr>
<td>Participant: Difficulty</td>
<td>.001</td>
<td>-1.000</td>
</tr>
<tr>
<td>Preamble: Intercept</td>
<td>.007</td>
<td></td>
</tr>
<tr>
<td>Preamble: Difficulty</td>
<td>&lt;.001</td>
<td>1.000</td>
</tr>
<tr>
<td>Preamble: Partner</td>
<td>&lt;.001</td>
<td>.034 .034</td>
</tr>
</tbody>
</table>

*Note.* Preamble: Difficulty*Partner was removed from the random effects structure to aid convergence.

Table B-18.  

*Best fitting model for Gap in Experiment 6.*  

<table>
<thead>
<tr>
<th>Fixed effect</th>
<th>Estimate</th>
<th>SE</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-.372</td>
<td>-.672</td>
<td>4.49</td>
</tr>
<tr>
<td>Difficulty (Hard vs. Easy)</td>
<td>.213</td>
<td>.164</td>
<td>4.92</td>
</tr>
<tr>
<td>Partner (SELF vs. OTHER)</td>
<td>-.920</td>
<td>-.715</td>
<td>5.78</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random effect</th>
<th>Variance</th>
<th>Corr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant: Intercept</td>
<td>1.078</td>
<td></td>
</tr>
<tr>
<td>Participant: Difficulty</td>
<td>.496</td>
<td>.219</td>
</tr>
<tr>
<td>Preamble: Intercept</td>
<td>†</td>
<td></td>
</tr>
<tr>
<td>Preamble: Difficulty</td>
<td>&lt;.001</td>
<td>†</td>
</tr>
<tr>
<td>Preamble: Partner</td>
<td>&lt;.001</td>
<td>†    -1.000</td>
</tr>
<tr>
<td>Picture: Intercept</td>
<td>.269</td>
<td></td>
</tr>
<tr>
<td>Picture: Difficulty</td>
<td>.444</td>
<td>.207</td>
</tr>
<tr>
<td>Picture: Partner</td>
<td>.104</td>
<td>.963 .462</td>
</tr>
<tr>
<td>Picture: Difficulty*Partner</td>
<td>.579</td>
<td>.291 .996 .537</td>
</tr>
</tbody>
</table>

*Note.* Preamble: Difficulty*Partner was removed from the random effects structure to aid convergence. All values refer to the Box-Cox transformed variable ($\lambda = .3408133$). For the fixed effects estimates, we also report values transformed back into the original scale (seconds) within brackets. Note that estimates for Box-Cox transformed variables tend to be biased (Gurka, Edwards, Muller, & Kupper, 2006). † Parameters that could not be estimated.
Table B-19.
*Best fitting model for Continuation in Experiment 6.*

<table>
<thead>
<tr>
<th>Fixed effect</th>
<th>Estimate</th>
<th>SE</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>5.160</td>
<td>.521</td>
<td>9.91</td>
</tr>
<tr>
<td>Difficulty (Hard vs. Easy)</td>
<td>1.651</td>
<td>.291</td>
<td>5.67</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random effect</th>
<th>Variance</th>
<th>Corr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant: Intercept</td>
<td>11.897</td>
<td></td>
</tr>
<tr>
<td>Participant: Difficulty</td>
<td>2.438</td>
<td>.424</td>
</tr>
<tr>
<td>Preamble: Intercept</td>
<td>.118</td>
<td></td>
</tr>
<tr>
<td>Preamble: Difficulty</td>
<td>.101</td>
<td>1.000</td>
</tr>
<tr>
<td>Preamble: Partner</td>
<td>.004</td>
<td>-1.000</td>
</tr>
<tr>
<td>Preamble: Difficulty*Partner</td>
<td>.087</td>
<td>-1.000</td>
</tr>
<tr>
<td>Picture: Intercept</td>
<td>.500</td>
<td></td>
</tr>
<tr>
<td>Picture: Difficulty</td>
<td>.290</td>
<td>-.598</td>
</tr>
<tr>
<td>Picture: Partner</td>
<td>.360</td>
<td>.437</td>
</tr>
<tr>
<td>Picture: Difficulty*Partner</td>
<td>1.862</td>
<td>- .378</td>
</tr>
</tbody>
</table>

Table B-20.
*Best fitting model for accuracy in Experiment 7.*

<table>
<thead>
<tr>
<th>Fixed effect</th>
<th>Est.</th>
<th>SE</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-3.038</td>
<td>.204</td>
<td>-14.89</td>
</tr>
<tr>
<td>Type (passive vs. active)</td>
<td>.391</td>
<td>.175</td>
<td>2.24</td>
</tr>
<tr>
<td>Language (native vs. non-native)</td>
<td>-.808</td>
<td>.350</td>
<td>-2.31</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant: Intercept</td>
<td>.406</td>
<td></td>
</tr>
<tr>
<td>Participant: Partner</td>
<td>.078</td>
<td>-.274</td>
</tr>
<tr>
<td>Participant: Type</td>
<td>.120</td>
<td>-.998</td>
</tr>
<tr>
<td>Participant:Partner*Type</td>
<td>.100</td>
<td>.166</td>
</tr>
<tr>
<td>Item: Intercept</td>
<td>.332</td>
<td></td>
</tr>
<tr>
<td>Item: Partner</td>
<td>.098</td>
<td>1.000</td>
</tr>
<tr>
<td>Item: Type</td>
<td>.002</td>
<td>-1.000</td>
</tr>
<tr>
<td>Item: Language</td>
<td>.994</td>
<td>.218</td>
</tr>
<tr>
<td>Item: Partner*Type</td>
<td>.133</td>
<td>-1.000</td>
</tr>
<tr>
<td>Item: Type*Language</td>
<td>2.141</td>
<td>.289</td>
</tr>
<tr>
<td>Item: Partner*Language</td>
<td>.659</td>
<td>-.301</td>
</tr>
</tbody>
</table>

Note. The by-item three-way interaction was dropped from the random effects structure to aid convergence.
Table B-21.

Best fitting model for onset in Experiment 7.

<table>
<thead>
<tr>
<th>Fixed effect</th>
<th>Estimate</th>
<th>SE</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>783</td>
<td>22</td>
<td>35.37</td>
</tr>
<tr>
<td>Partner (OTHER vs. NO)</td>
<td>-25</td>
<td>10</td>
<td>-2.47</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random effect</th>
<th>Variance</th>
<th>Correlations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant: Intercept</td>
<td>8531</td>
<td></td>
</tr>
<tr>
<td>Participant: Partner</td>
<td>953</td>
<td>-.433</td>
</tr>
<tr>
<td>Participant: Type</td>
<td>108</td>
<td>.770</td>
</tr>
<tr>
<td>Participant:Partner*Type</td>
<td>251</td>
<td>.894</td>
</tr>
<tr>
<td>Item: Intercept</td>
<td>2468</td>
<td></td>
</tr>
<tr>
<td>Item: Partner</td>
<td>171</td>
<td>-.428</td>
</tr>
<tr>
<td>Item: Type</td>
<td>18</td>
<td>.808</td>
</tr>
<tr>
<td>Item: Partner*Type</td>
<td>165</td>
<td>-.992</td>
</tr>
</tbody>
</table>

Note. The by-item three-way interaction was dropped from the random effects structure to aid convergence. Similarly for the two-way interactions of Language and Type and Language and Partner; we chose to remove these interactions because they were theoretically less interesting.
Table B-22. Best fitting model for Duration in the combined analyses of Experiments 8a and 8b.

<table>
<thead>
<tr>
<th>Fixed effect</th>
<th>Estimate</th>
<th>SE</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1062</td>
<td>25</td>
<td>42.81</td>
</tr>
<tr>
<td>Partner1 (NO vs. grand mean)</td>
<td>69</td>
<td>19</td>
<td>3.69</td>
</tr>
<tr>
<td>Partner2 (SELF vs. grand mean)</td>
<td>-110</td>
<td>22</td>
<td>-5.02</td>
</tr>
<tr>
<td>Difficulty (hard vs. easy)</td>
<td>41</td>
<td>8</td>
<td>5.39</td>
</tr>
<tr>
<td>Partner1*Type</td>
<td>-41</td>
<td>17</td>
<td>-2.46</td>
</tr>
<tr>
<td>Partner2*Type</td>
<td>60</td>
<td>19</td>
<td>3.22</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random effect</th>
<th>Variance</th>
<th>Corr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant: Intercept</td>
<td>18455</td>
<td></td>
</tr>
<tr>
<td>Participant: Partner1</td>
<td>12863</td>
<td>.012</td>
</tr>
<tr>
<td>Participant: Partner2</td>
<td>18857</td>
<td>.068</td>
</tr>
<tr>
<td>Participant: Difficulty</td>
<td>430</td>
<td>.771</td>
</tr>
<tr>
<td>Participant:</td>
<td>335</td>
<td>-.451</td>
</tr>
<tr>
<td>Partner1*Difficulty</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant:</td>
<td>3376</td>
<td>-.022</td>
</tr>
<tr>
<td>Partner2*Difficulty</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item: Intercept</td>
<td>4689</td>
<td></td>
</tr>
<tr>
<td>Item: Partner1</td>
<td>220</td>
<td>-1.000</td>
</tr>
<tr>
<td>Item: Partner2</td>
<td>363</td>
<td>.660</td>
</tr>
<tr>
<td>Item: Difficulty</td>
<td>278</td>
<td>.291</td>
</tr>
<tr>
<td>Item: Experiment</td>
<td>90</td>
<td>.965</td>
</tr>
</tbody>
</table>

Note. We removed the following random effects to aid convergence. By-item random effects: Partner*Difficulty*Experiment, Partner*Difficulty, Difficulty*Experiment, Partner*Experiment.
Table B-23.

*Best fitting model for Duration in Experiment 8a.*

<table>
<thead>
<tr>
<th>Fixed effect</th>
<th>Est.</th>
<th>SE</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1084</td>
<td>30</td>
<td>35.66</td>
</tr>
<tr>
<td>Partner1:(NO vs. grand mean)</td>
<td>85</td>
<td>16</td>
<td>5.29</td>
</tr>
<tr>
<td>Partner2: (SELF vs. grand mean)</td>
<td>-130</td>
<td>26</td>
<td>-5.05</td>
</tr>
<tr>
<td>Difficulty (hard vs. easy)</td>
<td>37</td>
<td>9</td>
<td>3.89</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant: Intercept</td>
<td>18310</td>
<td></td>
</tr>
<tr>
<td>Participant: Partner1</td>
<td>2837</td>
<td>.175</td>
</tr>
<tr>
<td>Participant: Partner2</td>
<td>12871</td>
<td>.237</td>
</tr>
<tr>
<td>Participant: Difficulty</td>
<td>469</td>
<td>.148</td>
</tr>
<tr>
<td>Participant:Partner1*Difficulty</td>
<td>1557</td>
<td>.549</td>
</tr>
<tr>
<td>Participant:Partner2*Difficulty</td>
<td>7013</td>
<td>.013</td>
</tr>
<tr>
<td>Difficulty</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item: Intercept</td>
<td>3990</td>
<td></td>
</tr>
<tr>
<td>Item: Partner1</td>
<td>464</td>
<td>-1.000</td>
</tr>
<tr>
<td>Item: Partner2</td>
<td>274</td>
<td>1.000</td>
</tr>
<tr>
<td>Item: Difficulty</td>
<td>152</td>
<td>1.000</td>
</tr>
<tr>
<td>Item:Partner1*Difficulty</td>
<td>118</td>
<td>1.000</td>
</tr>
<tr>
<td>Item:Partner2*Difficulty</td>
<td>5630</td>
<td>-.213</td>
</tr>
</tbody>
</table>
Table B-24.
Best fitting model for Duration in Experiment 8b.

<table>
<thead>
<tr>
<th>Fixed effect</th>
<th>Estimate</th>
<th>SE</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1048</td>
<td>33</td>
<td>32.21</td>
</tr>
<tr>
<td>Partner1 (NO vs. grand mean)</td>
<td>70</td>
<td>27</td>
<td>2.60</td>
</tr>
<tr>
<td>Partner2 (SELF vs. grand mean)</td>
<td>-116</td>
<td>25</td>
<td>-4.65</td>
</tr>
<tr>
<td>Difficulty (hard vs. easy)</td>
<td>42</td>
<td>11</td>
<td>3.71</td>
</tr>
<tr>
<td>Block1 (other vs. first)</td>
<td>-38</td>
<td>26</td>
<td>-1.44</td>
</tr>
<tr>
<td>Block2 (third vs. second)</td>
<td>-8</td>
<td>17</td>
<td>-0.45</td>
</tr>
<tr>
<td>Partner1*Difficulty</td>
<td>-40</td>
<td>24</td>
<td>-1.67</td>
</tr>
<tr>
<td>Partner2*Difficulty</td>
<td>58</td>
<td>24</td>
<td>2.39</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random effect</th>
<th>Variance</th>
<th>Corr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant: Intercept</td>
<td>19014</td>
<td></td>
</tr>
<tr>
<td>Participant: Partner1</td>
<td>7524</td>
<td>-.103</td>
</tr>
<tr>
<td>Participant: Partner2</td>
<td>5036</td>
<td>.009</td>
</tr>
<tr>
<td>Participant: Difficulty</td>
<td>860</td>
<td>.921</td>
</tr>
<tr>
<td>Participant: Block1</td>
<td>12791</td>
<td>.470</td>
</tr>
<tr>
<td>Participant: Block2</td>
<td>1858</td>
<td>.079</td>
</tr>
<tr>
<td>Item: Intercept</td>
<td>5212</td>
<td></td>
</tr>
<tr>
<td>Item: Partner1</td>
<td>1626</td>
<td>-.047</td>
</tr>
<tr>
<td>Item: Partner2</td>
<td>2676</td>
<td>.052</td>
</tr>
<tr>
<td>Item: Difficulty</td>
<td>352</td>
<td>-.143</td>
</tr>
<tr>
<td>Item: Block1</td>
<td>353</td>
<td>-.996</td>
</tr>
<tr>
<td>Item: Block2</td>
<td>117</td>
<td>.482</td>
</tr>
</tbody>
</table>

Note. We removed the following random effects to aid convergence. By-participant random effects: Block*Difficulty, Partner*Difficulty; by-item random effects: Partner*Difficulty*Block, Partner*Difficulty, Difficulty*Block, Partner*Block.
Table B-25.

*Best fitting model for Onset in the combined analyses of Experiments 8a and 8b.*

<table>
<thead>
<tr>
<th>Fixed effect</th>
<th>Estimate</th>
<th>SE</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1081</td>
<td>25</td>
<td>42.71</td>
</tr>
<tr>
<td>Partner1 (NO vs. grand mean)</td>
<td>17</td>
<td>14</td>
<td>1.24</td>
</tr>
<tr>
<td>Partner2 (SELF vs. grand mean)</td>
<td>42</td>
<td>14</td>
<td>3.02</td>
</tr>
<tr>
<td>Difficulty (hard vs. easy)</td>
<td>63</td>
<td>16</td>
<td>3.85</td>
</tr>
<tr>
<td>Experiment (b vs. a)</td>
<td>-20</td>
<td>41</td>
<td>-.50</td>
</tr>
<tr>
<td>Partner1*Experiment</td>
<td>27</td>
<td>27</td>
<td>1.01</td>
</tr>
<tr>
<td>Partner2*Experiment</td>
<td>48</td>
<td>26</td>
<td>1.81</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random effect</th>
<th>Variance</th>
<th>Corr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant: Intercept</td>
<td>20852</td>
<td></td>
</tr>
<tr>
<td>Participant: Partner1</td>
<td>5848</td>
<td>-.076</td>
</tr>
<tr>
<td>Participant: Partner2</td>
<td>5273</td>
<td>.221</td>
</tr>
<tr>
<td>Participant: Difficulty</td>
<td>466</td>
<td>.310</td>
</tr>
<tr>
<td>Participant:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Partner1*Difficulty</td>
<td>252</td>
<td>-.810</td>
</tr>
<tr>
<td>Participant:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Partner2*Difficulty</td>
<td>1214</td>
<td>.748</td>
</tr>
<tr>
<td>Item: Intercept</td>
<td>4617</td>
<td></td>
</tr>
<tr>
<td>Item: Partner1</td>
<td>6</td>
<td>1.000</td>
</tr>
<tr>
<td>Item: Partner2</td>
<td>488</td>
<td>.365</td>
</tr>
<tr>
<td>Item: Difficulty</td>
<td>4665</td>
<td>.333</td>
</tr>
<tr>
<td>Item: Experiment</td>
<td>259</td>
<td>-.824</td>
</tr>
</tbody>
</table>

*Note.* We removed the following random effects to aid convergence. By-item random effects: Partner*Difficulty*Experiment, Partner*Difficulty, Difficulty*Experiment, Partner*Experiment.
Table B-26.
Best fitting model for Onset in Experiment 8a.

<table>
<thead>
<tr>
<th>Fixed effect</th>
<th>Estimate</th>
<th>SE</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1104</td>
<td>32</td>
<td>34.84</td>
</tr>
<tr>
<td>Difficulty (hard vs. easy)</td>
<td>59</td>
<td>16</td>
<td>3.60</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random effect</th>
<th>Variance</th>
<th>Corr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant: Intercept</td>
<td>20948</td>
<td></td>
</tr>
<tr>
<td>Participant: Partner1 (NO vs. grand mean)</td>
<td>67</td>
<td>-1.000</td>
</tr>
<tr>
<td>Participant: Partner2 (SELF vs. grand mean)</td>
<td>337</td>
<td>-.064</td>
</tr>
<tr>
<td>Participant: Difficulty</td>
<td>56</td>
<td>.512</td>
</tr>
<tr>
<td>Participant: Partner1*Difficulty</td>
<td>291</td>
<td>-.943</td>
</tr>
<tr>
<td>Participant: Partner2*Difficulty</td>
<td>2996</td>
<td>-.650</td>
</tr>
<tr>
<td>Item: Intercept</td>
<td>5560</td>
<td></td>
</tr>
<tr>
<td>Item: Partner1</td>
<td>93</td>
<td>1.000</td>
</tr>
<tr>
<td>Item: Partner2</td>
<td>743</td>
<td>-1.000</td>
</tr>
<tr>
<td>Item: Difficulty</td>
<td>5496</td>
<td>.368</td>
</tr>
<tr>
<td>Item: Partner1*Difficulty</td>
<td>4221</td>
<td>.481</td>
</tr>
<tr>
<td>Item: Partner2*Difficulty</td>
<td>3108</td>
<td>-.650</td>
</tr>
</tbody>
</table>
Table B-27.  
*Best fitting model for Onset in Experiment 8b.*

<table>
<thead>
<tr>
<th>Fixed effect</th>
<th>Est.</th>
<th>SE</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1081</td>
<td>32</td>
<td>33.43</td>
</tr>
<tr>
<td>Partner1 (NO vs. grand mean)</td>
<td>23</td>
<td>18</td>
<td>1.24</td>
</tr>
<tr>
<td>Partner2 (SELF vs. grand mean)</td>
<td>83</td>
<td>20</td>
<td>4.15</td>
</tr>
<tr>
<td>Difficulty (hard vs. easy)</td>
<td>68</td>
<td>17</td>
<td>4.04</td>
</tr>
<tr>
<td>Block1 (other vs. first)</td>
<td>-75</td>
<td>17</td>
<td>-4.34</td>
</tr>
<tr>
<td>Block2 (third vs. second)</td>
<td>-44</td>
<td>15</td>
<td>-3.02</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant: Intercept</td>
<td>20814</td>
<td></td>
</tr>
<tr>
<td>Participant: Partner1</td>
<td>1917</td>
<td>.279</td>
</tr>
<tr>
<td>Participant: Partner2</td>
<td>4701</td>
<td>-.302</td>
</tr>
<tr>
<td>Participant: Difficulty</td>
<td>1473</td>
<td>.267</td>
</tr>
<tr>
<td>Participant: Block1</td>
<td>3774</td>
<td>.776</td>
</tr>
<tr>
<td>Participant: Block2</td>
<td>900</td>
<td>-.077</td>
</tr>
<tr>
<td>Participant: Partner1*Difficulty</td>
<td>563</td>
<td>-.545</td>
</tr>
<tr>
<td>Participant: Partner2*Difficulty</td>
<td>4399.89</td>
<td>.010</td>
</tr>
<tr>
<td>Item: Intercept</td>
<td>3748.03</td>
<td></td>
</tr>
<tr>
<td>Item: Partner1</td>
<td>1622.03</td>
<td>-.061</td>
</tr>
<tr>
<td>Item: Partner2</td>
<td>899.89</td>
<td>.267</td>
</tr>
<tr>
<td>Item: Difficulty</td>
<td>3515.28</td>
<td>.330</td>
</tr>
<tr>
<td>Item: Block1</td>
<td>885.96</td>
<td>-.598</td>
</tr>
<tr>
<td>Item: Block2</td>
<td>374.18</td>
<td>-.596</td>
</tr>
</tbody>
</table>

*Note.* We removed the following random effects to aid convergence. By-participant random effects: Block*Difficulty; by-item random effects: Partner*Difficulty*Block, Partner*Difficulty, Difficulty*Block, Partner*Block.
Table B-28.

*Best fitting model for Gap in Experiment 8a.*

<table>
<thead>
<tr>
<th>Fixed effect</th>
<th>Estimate</th>
<th>SE</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-2057.10(-119)</td>
<td>65.90</td>
<td>-31.21</td>
</tr>
<tr>
<td>Partner: (SELF vs. OTHER)</td>
<td>-854.70(-112)</td>
<td>98.75</td>
<td>-8.66</td>
</tr>
<tr>
<td>Difficulty (hard vs. easy)</td>
<td>274.82(35)</td>
<td>48.81</td>
<td>5.63</td>
</tr>
<tr>
<td>Partner*Difficulty</td>
<td>-481.25(-77)</td>
<td>96.01</td>
<td>-5.01</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random effect</th>
<th>Variance</th>
<th>Correlation</th>
<th>s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant: Intercept</td>
<td>89498.2</td>
<td>.392</td>
<td></td>
</tr>
<tr>
<td>Participant: Partner</td>
<td>199184.4</td>
<td>-.270</td>
<td>.290</td>
</tr>
<tr>
<td>Participant: Difficulty</td>
<td>29313.4</td>
<td>.930</td>
<td>-.596</td>
</tr>
<tr>
<td>Participant: Partner*Difficulty</td>
<td>108234.8</td>
<td>.576</td>
<td>.347</td>
</tr>
<tr>
<td>Item: Intercept</td>
<td>8068.3</td>
<td>-.168</td>
<td></td>
</tr>
<tr>
<td>Item: Partner</td>
<td>11165.9</td>
<td>.419</td>
<td>.111</td>
</tr>
<tr>
<td>Item: Difficulty</td>
<td>5161.0</td>
<td>.347</td>
<td>.283</td>
</tr>
<tr>
<td>Item: Partner*Difficulty</td>
<td>21981.8</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* All values refer to the Box-Cox transformed variable \((\lambda = .308748)\). For the fixed effects estimates, we also report values transformed back into the original scale (milliseconds) within brackets. Note that estimates for Box-Cox transformed variables tend to be biased (Gurka, et al., 2006).
Table B-29.

*Best fitting model for Gap in Experiment 8b.*

<table>
<thead>
<tr>
<th>Fixed effect</th>
<th>Estimate</th>
<th>SE</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-2216.02</td>
<td>55.12</td>
<td>-40.20</td>
</tr>
<tr>
<td>Partner (SELF vs. OTHER)</td>
<td>-797.77</td>
<td>102.38</td>
<td>-7.79</td>
</tr>
<tr>
<td>Difficulty (hard vs. easy)</td>
<td>241.77</td>
<td>47.00</td>
<td>5.12</td>
</tr>
<tr>
<td>Block1 (other vs. first)</td>
<td>116.08</td>
<td>130.01</td>
<td>.89</td>
</tr>
<tr>
<td>Block2 (third vs. second)</td>
<td>208.99</td>
<td>97.10</td>
<td>2.15</td>
</tr>
<tr>
<td>Partner*Difficulty</td>
<td>-242.36</td>
<td>71.10</td>
<td>-3.41</td>
</tr>
<tr>
<td>Partner*Block1</td>
<td>-166.60</td>
<td>222.02</td>
<td>-.75</td>
</tr>
<tr>
<td>Partner*Block2</td>
<td>-604.39</td>
<td>234.41</td>
<td>-2.58</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random effect</th>
<th>Variance</th>
<th>Corr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant: Intercept</td>
<td>48105.18</td>
<td></td>
</tr>
<tr>
<td>Participant: Partner</td>
<td>97374.78</td>
<td>-.800</td>
</tr>
<tr>
<td>Participant: Difficulty</td>
<td>22454.47</td>
<td>-.503</td>
</tr>
<tr>
<td>Participant: Block1</td>
<td>196968.61</td>
<td>-.137</td>
</tr>
<tr>
<td>Participant: Block2</td>
<td>1697.61</td>
<td>-.269</td>
</tr>
<tr>
<td>Item: Intercept</td>
<td>8661.89</td>
<td></td>
</tr>
<tr>
<td>Item: Partner</td>
<td>12110.60</td>
<td>-.744</td>
</tr>
<tr>
<td>Item: Difficulty</td>
<td>197.65</td>
<td>-.076</td>
</tr>
<tr>
<td>Item: Block1</td>
<td>13005.60</td>
<td>-.065</td>
</tr>
<tr>
<td>Item: Block2</td>
<td>10.03</td>
<td>.116</td>
</tr>
</tbody>
</table>

*Note.* We removed the following random effects to aid convergence. By-participant random effects: Block*Difficulty, Partner*Difficulty; by-item random effects: Partner*Difficulty*Block, Partner*Difficulty, Difficulty*Block, Partner*Block. The values reported refer to the Box-Cox transformed variable ($\lambda = .01480653$). Note that estimates for Box-Cox transformed variables tend to be biased (Gurka, et al., 2006).
9. Appendix C: Publications
