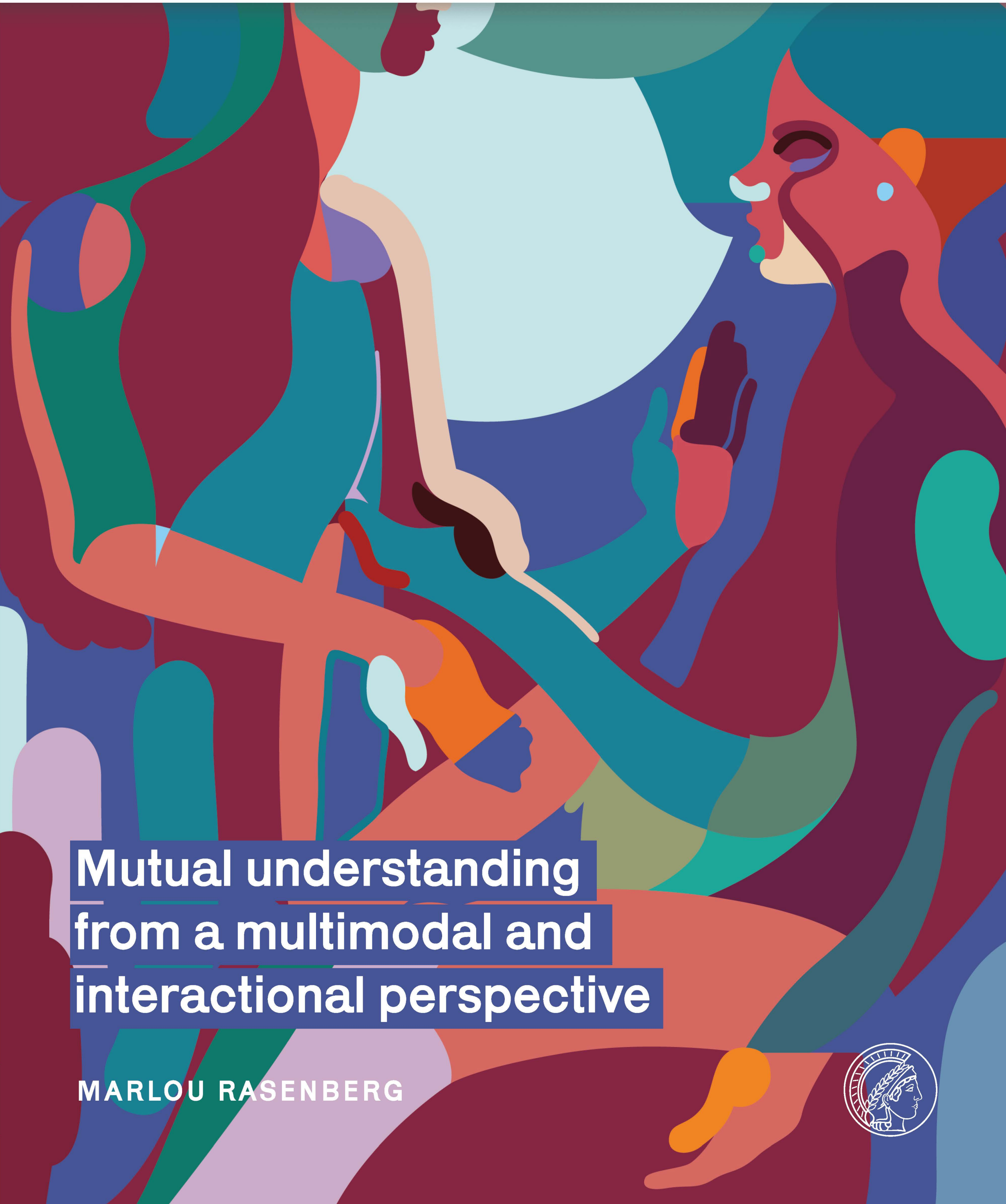
MAX PLANCK INSTITUTE
FOR PSYCHOLINGUISTICS



Mutual understanding from a multimodal and interactional perspective

MARLOU RASENBERG

Mutual understanding from a multimodal and interactional perspective

Marlou Rasenberg

Funding body

This research was funded by the Netherlands Organisation for Scientific Research (NWO) Gravitation Grant 024.001.006 to the Language in Interaction consortium.

International Max Planck Research School (IMPRS) for Language Sciences

The educational component of the doctoral training was provided by the International Max Planck Research School (IMPRS) for Language Sciences. The graduate school is a joint initiative between the Max Planck Institute for Psycholinguistics and two partner institutes at Radboud University – the Centre for Language Studies, and the Donders Institute for Brain, Cognition and Behaviour. The IMPRS curriculum, which is funded by the Max Planck Society for the Advancement of Science, ensures that each member receives interdisciplinary training in the language sciences and develops a well-rounded skill set in preparation for fulfilling careers in academia and beyond. More information can be found at www.mpi.nl/imprs

The MPI series in Psycholinguistics

Initiated in 1997, the MPI series in Psycholinguistics contains doctoral theses produced at the Max Planck Institute for Psycholinguistics. Since 2013, it includes theses produced by members of the IMPRS for Language Sciences. The current listing is available at www.mpi.nl/mpi-series

ISBN: 978-94-92910-45-5
Cover design and artwork: Alexander Voutchkov
Layout and design: Hans Schaapherder | persoonlijkproefschrift.nl
Printed and bound by Ipskamp Drukkers, Enschede

The research reported in this thesis was carried out at the Centre for Language Studies, the Donders Institute for Brain, Cognition and Behaviour, Radboud University and the Max Planck Institute for Psycholinguistics, in Nijmegen, the Netherlands.

Mutual understanding from a multimodal and interactional perspective

Proefschrift ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college voor promoties
in het openbaar te verdedigen op

vrijdag 17 maart 2023
om 12.30 uur precies

door

Marlou Elisa Rasenberg
geboren op 2 oktober 1993
te Made en Drimmelen

Promotor:
Prof. dr. H.A. Özyürek

Copromotor:
Dr. M. Dingemanse

Manuscriptcommissie:
Prof. dr. I.J.E.I. van Rooij (voorzitter)
Prof. dr. G. Brône (KU Leuven, België)
Dr. R. Fussaroli (Aarhus Universitet, Denemarken)
Dr. M.W. Hoetjes
Dr. J.A.M. Holler

Contents

Chapter 1

General introduction

General introduction

We spend a great deal of our lives interacting with others. For example when catching up with family or friends, or when working together to cook a meal, create a theatre play, or do research. It is our capacity for social interaction—the human "interaction engine"—that lies at the basis for societies as we know them today (Levinson, 2006). Crucial for these interactions, and for people to reach their shared goal, is that people understand each other. Without mutual understanding, they won't get far.

This thesis is concerned with the question how people jointly negotiate mutual understanding in social interaction. I will investigate this in the context of referential communication about novel referents; a communicatively challenging setting which enables me to study how people make sure they understand each other. To undertake this research endeavour, I start from the following vantage point:

> Social interaction is collaborative and multimodal.

Though this might seem a trivial statement, taking this perspective has fundamental implications for how to go about studying social interactions, and pushes us to go beyond traditional research foci in (psycho)linguistics and cognitive science.

First, if we ask how people establish mutual understanding through *collaborative interaction*, it draws our attention to the joint work that people do. That is, we consider interactions to be *joint actions*, in which people coordinate their actions to reach shared goals (Clark, 1996; Sebanz et al., 2006). When studying the coordination of language use in social interaction, our analytical focus becomes the utterance as it is embedded in a sequential context—as "simultaneously an effect of something prior and a cause of something next" (Enfield, 2009, p. 223). To study how people jointly work towards mutual understanding, I focus on the interactional practices that people have available, for which I use insights from conversation analysis and interactional linguistics (Clift, 2016; Couper-Kuhlen & Selting, 2017).

Second, if we ask how people establish mutual understanding through *multimodal interaction*, then we should look at how people use different semiotic resources across modalities. When we think of conversations as a form of social interaction, speech usually comes to mind first. Yet people use a whole range of communicative signals, using their body, face and hands. Of special interest here are gestures—specifically hand gestures that co-occur with speech—as these can convey semantic meaning, and serve various communicative functions (Bavelas & Chovil, 2000; Kendon, 2004; McNeill, 1992). To understand how different signals convey meaning in interaction, and how they complement each other, I build on prior work in the field of semiotics and gesture studies.

Finally, if we consider mutual understanding to follow from social interactions which are both collaborative *and* multimodal in nature, different questions come to light. How are different modalities used for particular interactional practices? How does this follow from modality-specific affordances or constraints? By investigating how people collaborate using multiple modalities to reach the joint goal of mutual understanding, I aim to contribute to this body of work.

Collaborative interaction

How do people establish mutual understanding in multimodal interaction? In this thesis, I will focus on two interactional phenomena: alignment and other-initiated repair. I use alignment in the sense of cross-participant repetition of communicative behaviour, for example when a speaker repeats a word or gesture that their conversational partner used earlier in the interaction. Other-initiated repair refers to the joint efforts to resolve trouble with perceiving or understanding a prior turn (Schegloff, 2000; Schegloff et al., 1977), which can be initiated with utterances like "huh?" or "you mean this one ((point))?".

Why focus on these two phenomena? Both are prevalent in social interaction and have been argued to play a fundamental role in the process of establishing mutual understanding (see e.g., Albert & de Ruiter, 2018; Pickering & Garrod, 2004). Yet both have been studied across different research fields, yielding different takes on how they support mutual understanding. To see how our understanding of these phenomena would benefit from disciplinary integration and a multimodal perspective, I believe it is useful to start from a broader view on how mutual understanding is accomplished through and embedded in social interaction.

For the most part, establishing mutual understanding is a tacit process. If you ask me "coffee or tea?" and I reply with "coffee would be nice", it is apparent to both of us that I understood your utterance as an offer. If you subsequently hand me a cup of coffee, this shows that you understood my reply as an acceptance of your offer. As such, displays of understanding are "by-products of bits of talk[1] designed in the first instance to do some action such as agreeing, answering, assessing, responding, questioning and so on" (Schegloff, 1992, p. 1300), they are simply part of the progressive continuation of the unfolding interaction (Mondada, 2011).

But even in such smooth interactions, participants do not take understanding for granted; they "bear the mutual responsibility of assuring that what is said has been heard and understood before the conversation goes on" (Clark & Schaefer, 1987, p. 19). This process entails the *grounding* of contributions, adding them to people's *common ground* (i.e., their

[1] Throughout this thesis, I consider "talk" (and "talk-in-interaction" and "turns-at-talk") in a modality-independent way; I take it to refer to both spoken and signed language, as well as embodied signals such as manual gestures and facial expressions.

mutual knowledge, mutual beliefs and mutual assumptions; Clark, 1996; Clark & Brennan, 1991; Clark & Schaefer, 1987). To this end, participants in interaction seek evidence that their utterance has been understood (Clark, 1996; Clark & Brennan, 1991); they are attentive to the addressee's signals of understanding and non-understanding. Such signals can take different forms, such as a relevant next turn (as in the coffee-example), or feedback responses like "yes" or "mm" (Allwood & Ahlsen, 1999; see also back-channel responses; Yngve, 1970).

How do alignment and other-initiated repair sit in with this notion of signals of understanding and non-understanding? I will start with repair, because—unlike alignment—it is a well-bounded interactional phenomenon, with a cumulative and cohesive research history. Other-initiated repair refers to situations where an addressee experiences problems with perceiving or understanding prior talk, inviting the producer to fix it (Schegloff, 2000; Schegloff et al., 1977). Depending on the nature of the trouble, repair initiations can take various forms (e.g., "come again?", "who?" or "this one ((point))?"), which indicate how the sender can fix the problem (repeat the whole utterance, specify a person reference, or confirm the candidate understanding). Other-initiated repair has been well-attested in terms of its sequential structure (an insert sequence composed of an initiation and proposed solution) and repair initiation formats (ranging from less to more specific; Dingemanse & Enfield, 2015; Drew, 1997; Schegloff, 2004). What all repair initiations have in common is that they are produced to halt the ongoing course of the conversation in order to attend to interactional trouble—thus qualifying as a signal of non-understanding.

Categorizing alignment as a signal of understanding or non-understanding is less straightforward. A general assumption that dominates the psycholinguistic literature is that alignment holds when two people refer to the same thing with the same linguistic structure, which could be considered a signal of understanding (i.e., by using a linguistic structure in the same way as their partner, people can signal that they understood the partner's prior use). In some experimental studies on alignment this is also the only possible way in which alignment can appear, as people do not interact freely but take turns (with a confederate) to label images (e.g., Branigan et al., 2000, 2007; Cai et al., 2021). This is beneficial in that it allows for controlled quantification and hypothesis testing, as well as laying the groundwork for computational modelling (e.g., Buschmeier et al., 2010; Goudbeek & Krahmer, 2012). However, it overlooks the diverse ways in which cross-participant repetition is realised in interaction (see e.g., Brennan, 1996; Fusaroli et al., 2017; Holler & Wilkin, 2011a; Norrick, 1987; Tannen, 1989). For instance, people can repeat a phrase but use it to refer to a different referent, or they might repeat their partner's phrase to ask for clarification.

The waters get even murkier if we realise that under certain accounts alignment is considered at the level of mental representations, rather than at the level of behaviour (Stolk et al., 2016). A popular account is the "the interactive alignment account" by Pickering and Garrod (2004), which proposes that mutual understanding comes about as a result of the alignment of linguistic representations. The main process through which this happens is *priming*; when participants in interaction hear a certain phoneme, word, or syntactic structure, the corresponding representation becomes activated, yielding cross-speaker alignment of linguistic representations. The alignment of representations "percolates" across linguistic levels (from phonology to semantics), ultimately yielding alignment of higher-level *situation model* representations, meaning that people have a similar understanding of the situation under discussion. This alignment process is considered to be an automatic and "primitive" mechanism underlying dialogue; it is only when it fails that people resort to more explicit interactional strategies such as other-initiated repair (though these have a more prominent role in extended versions of the original model; Gandolfi et al., in press; Pickering & Garrod, 2021).

In sum, the diverse range of empirical and theoretical approaches to alignment has yielded a rather convoluted image of what alignment is and how it contributes to mutual understanding. This stands in stark contrast with other-initiated repair, of which the sequential structure and interactional purpose is clearly defined. As such, the state of the literature on alignment calls for integrative efforts to enable cumulative progress in this research area, while the thorough understanding of repair opens up possibilities to study novel research questions about language use as a joint action. But before going into detail about how I will undertake these missions in this thesis, we will turn to the issue of multimodality, as I will argue that a multimodal perspective could enhance our understanding of both alignment and other-initiated repair.

Multimodal interaction

Language use is inherently multimodal and semiotically diverse. This holds for both spoken and signed languages. Here I will concentrate on spoken language, though I derive insights from and draw parallels with signed languages at various points. In co-present conversation,[2] people use speech, hand and head movements, eye gaze, facial expressions and body posture, all of which are interconnected at the level of individual´s behaviour, and—as we shall see—between individuals.

In this thesis I focus on speech and manual co-speech gestures. Co-speech gestures are spontaneous bodily movements which are produced in a tight semantic, pragmatic and temporal relationship to speech (McNeill, 1992). There are different types of co-speech gestures. *Iconic* gestures are gestures which visually depict aspects of objects (such as

[2] Here I use "co-present" rather than "face-to-face" to discuss social interaction in a general sense, acknowledging the fact that there is cultural variation in how people usually position themselves with respect to each other in social interaction (Ameka & Terkourafi, 2019). Later on, I will use "face-to-face" when discussing interactions that took place in a face-to-face setting, such as those studied in this thesis.

shape or size), spatial relationships, actions or motions. For example, you can depict the act of drinking by bringing your hand towards your mouth, your fingers somewhat apart as if holding a glass. *Deictic* or pointing gestures are used to single out concrete or abstract referents. Beyond these two main categories, other gesture types have been distinguished under various labels (Bavelas et al., 1992; Kendon, 2004; McNeill, 1992). These include *beat* gestures (biphasic movements which match the rhythm of the concurrent speech), *metaphorical* gestures (which depict concrete objects or actions in order to convey abstract concepts), *interactive* and *pragmatic* gestures (which operate on the discourse-level or refer to the addressee), and *emblems* (conventionalised gestures with a specific meaning within a culture). In the research presented in this thesis, I differentiate between "iconic" and "deictic" gestures, and group the remaining types into the category "other".

Speech and gestures can convey meaning in different ways, and as such offer different affordances for joint meaning-making. A key characteristic of speech is compositionality; meaning is derived from the hierarchical composition of discrete elements (words, morphemes). In contrast to this linear-segmented character of speech, iconic gestures convey meaning in a *global-synthetic* manner (McNeill, 1992). Multiple meanings can be conveyed with one, holistic gesture. For example, if you were to gesturally depict the catching of a ball, you are likely to convey meaning with respect to the size of the ball (keeping your hands a certain distance apart; the fingers in a particular shape) and how it was caught (by holding both hands above your head). And with gesture you can do all of this at once, rather than word by word (Slonimska et al., 2020). Another way in which speech and gestures can be compared is through the lens of semiotics, as the different types of signs in the spoken and gestural modality can be used for different methods of signalling (Clark, 1996, 2016; Ferrara & Hodge, 2018; building on the work of Peirce, 1955). That is, conventionalised words (symbols) can be used for *describing*, and iconic gestures (icons) for *depicting* through perceptual resemblances (though both methods can be attained with both modalities, e.g., depicting through onomatopoeic speech, describing through emblematic gestures; Clark, 1996).

Speech and gesture can be used separately, but usually they are combined in *composite utterances* which can take various forms (Clark, 1996; Enfield, 2009; Engle, 1998; Engle & Clark, 1995; Kendon, 2004; McNeill, 1992). Gestures can be *concurrent* or *co-expressive* with speech (Clark, 1996; McNeill, 1992). Importantly, in such combinations, speech and gesture can still highlight or qualify certain meaning aspects. For example, if the ball-catching gesture from above is produced while saying "I caught a ball", the gesture presents meaning (about the size of the ball and the way it is caught) which is *complementary* to what is conveyed through speech (McNeill, 1992). Gestures can also convey meaning which is not conveyed through speech at all (e.g., when gesturing while saying "like this"), which makes the gesture a *component* of the utterance (Clark, 1996). In each of these instances, speech and gesture jointly convey meaning; "the relationship between speech and gesture is a *reciprocal* one—the gestural component and the spoken component *interact* with one another to create a precise and vivid understanding" (Kendon, 2004, p. 174; original emphasis).

Addressees subsequently make sense of these composite utterances; they "[look] for a way in which those co-occurring signs may simultaneously point to a single overall message of the move that a person is making" (Enfield, 2013, p. 689). Evidence from behavioural and neurophysiological studies shows that people process both speech and iconic gestures semantically (using similar brain regions), integrating them to form a coherent message representation (Kelly et al., 2010; Özyürek, 2014). For example, in Kelly et al. (1999), participants were presented with multimodal stimuli, such as the spoken utterance "my brother went to the gym" along with an iconic gesture depicting the shooting of a basketball. They found that even when participants were explicitly asked to recall what was *said*, the participants would often combine the meanings expressed by speech and gesture (e.g., "went to play ball").

In sum, speakers flexibly deploy speech and co-speech gesture to convey meaning, which addressees are able to integrate and make sense of. While there is ample theoretical work on the semiotic principles of spoken and gestural signs (e.g., Capirci et al., 2022; Clark, 1996; Enfield, 2009; Ferrara & Hodge, 2018; Kockelman, 2005), as well as psycholinguistic studies and models of speech-gesture production and comprehension (e.g., de Ruiter, 2007; Kita & Özyürek, 2003; McNeill, 1992; Özyürek, 2018), here we are concerned with the *coordination* of composite utterances in social interaction. This raises the question how people orient to semiotically-diverse signals in co-present interaction, and how this shapes their incremental contributions as they unfold in real time (e.g., C. Goodwin, 2000; Holler, 2022; Mondada, 2011; Stivers & Sidnell, 2005). By focussing on speech and (iconic) co-speech gestures in other-initiated repair and alignment, we can explore how different properties of communicative systems (conventionalised versus non-conventionalised and arbitrary versus iconic[3]) are used and combined in this process of joint meaning-making.

Current thesis

The prior sections have sketched a picture of social interaction as involving various interactional processes (such as other-initiated repair and alignment) and multimodal language use (such as speech and co-speech gestures). Yet I have purposely divided this into separate sections, since there has been little cross-talk between these lines of research (with some notable exceptions, e.g., Chui, 2014; Holler & Wilkin, 2011a; Hömke, 2019;

[3] This is not to say that these are strictly separate categories that map onto speech versus gesture. Instead, it could be useful to think of them as features that crosscut modalities (Capirci et al., 2022; Dingemanse, Blasi, et al., 2015; Okrent, 2002).

Oben & Brône, 2016; Sikveland & Ogden, 2012 and others as we will see in later chapters). This is perhaps in part because research on other-initiated repair builds on the speech-centred work of Schegloff et al. (1977), and psycholinguistic research on alignment has been inspired by the speech-centred interactive alignment account (Pickering & Garrod, 2004). Conversely, foundational work in semiotics and gesture studies (e.g., McNeill, 1992; Peirce, 1955) centred around isolated signals or individual language processing, which is multiple steps removed from actual dialogue. However, interactional processes and semiotic diversity have been studied together in multimodal interactional studies (see e.g., Bavelas et al., 1992; C. Goodwin, 1981, 2000; Kendrick & Holler, 2017; Mondada, 2011; Stivers & Sidnell, 2005; Streeck, 1993, 1994). Here I follow up on this work to contribute to our understanding of how different semiotic resources are recruited in interaction for doing collaborative work with alignment and other-initiated repair.

To this end, I adopt an interactional approach to language use. To understand how this approach differs from other approaches (e.g., in psycholinguistics), the distinction between Interactional and Aggregate approaches to dialogue by Healey et al. (2018) proves useful:

> Aggregate approaches emphasize the use of established models of individual language processing as a way of accounting for dialogue. They use the same basic cognitive mechanisms that explain individual lexical, syntactic, and semantic processing and then model dialogue as aggregations of those mechanisms acting in concert . . . In contrast to this, Interactional approaches propose that dialogue presents processing challenges that are qualitatively different from individual language processing and require different mechanisms to explain them . . . (p. 369)

An interactional approach to dialogue draws our attention to the processes and tools that people use to streamline coordination. People indicate that they are having trouble formulating an utterance, they manage when and who will take the next turn, and—as I will focus on here—they negotiate mutual understanding. In particular, I zoom in on other-initiated repair and alignment, and investigate how they are realised through speech and co-speech gestures. I investigate this in the context of a specific communicative challenge: collaboratively referring to things for which a conventionalised label is not (yet) available (see Methods below). We regularly encounter such challenges in our daily lives. For example when referring to a person whose name you don't know (e.g., "the kid with the baseball cap") or when talking about an abstract piece of art you saw in a museum. Investigating how people use speech and gesture in alignment and other-initiated repair to negotiate referential expressions for novel referents can help us see how people establish mutual understanding in interaction more generally.

This brings me to **the main research aim of this thesis**: to contribute towards our understanding of how people work together to negotiate mutual understanding in multimodal interaction. This larger aim is broken down into a number of concrete objectives: i) situate alignment in multimodal, sequential talk-in-interaction—both conceptually and empirically, ii) investigate the other-initiated repair system from a multimodal perspective, iii) examine the division of multimodal labour between participants in social interaction from a joint action perspective.

Objective i) follows from the discussion of the state of the literature in the section "Collaborative interaction". This section made clear that while we have good understanding of other-initiated repair in terms of its sequential structure and interactional function, there are diverging takes on what alignment is and how it contributes to mutual understanding. To facilitate cumulative progress in the study of alignment, we would benefit from an integrative framework. This is developed in chapter 2, where I show how it can be used as a conceptual tool to contrast empirical approaches and uncover hidden assumptions. I argue that understanding the building blocks of alignment puts us in a better position to empirically study alignment in the complex reality of sequentially-organised multimodal interaction.

As such the framework provides a stepping stone for a careful operationalisation of alignment in multimodal interaction in chapter 3, where I track when and how frequently people use lexical and gestural alignment to collaboratively refer to novel referents. I qualitatively analyse the sequential environments in which alignment comes about, to shed light on *how* alignment of speech and gesture is used for collaborative referring (complementing quantitative corpus and experimental studies on alignment; e.g., Branigan et al., 2007; Reitter & Moore, 2014). The study of alignment in a task-based setting with novel referents corresponds to a more general question: how do people negotiate referential expressions, ultimately resulting in shared symbols for the referents? Prior research on symbol creation, in the lab or in terms of the emergence of natural languages, has singled out alignment (Fay et al., 2018; Lister & Fay, 2017) and gestural or multimodal communication (e.g., Fay et al., 2013, 2014; Levinson & Holler, 2014; Macuch Silva et al., 2020; Sterelny, 2012; Zlatev et al., 2017) as key elements, though they have not yet been studied together. Here I take the next step by studying the role of modality *in* alignment as an interactional resource for collaborative referring to novel referents.

Objective ii) of this thesis is to investigate the other-initiated repair system from a multimodal perspective. Studies aimed at characterizing other-initiated repair as a unified system have so far focused mainly on speech (Dingemanse & Enfield, 2015; Schegloff et al., 1977; but see Manrique, 2016 on Argentine Sign Language), showing systematicity in the use of linguistic resources for the different types of repair initiations (e.g., "Huh?" or "Who?") and the corresponding repair solutions. While there is also quite some research on the role of co-speech gestures and facial expressions in other-initiated repair (e.g., Floyd et al., 2016; Hoetjes, Krahmer, et al., 2015; Holler & Wilkin, 2011b; Hömke, 2019; Jokipohja & Lilja, 2022; Mortensen, 2016), these are often singular examples or quantitative investigations focusing on repair initiations *or* solutions. In chapter 4 I undertake a more holistic investigation of other-initiated repair as a multimodal system. To this end, I quantify the use of co-speech gestures across initiating and responding positions of different repair types, and then qualitatively inspect how the interactional work in these turns is realised with multimodal strategies.

Finally, **objective iii)** pertains to the basic premise that social interaction is collaborative and multimodal. I put this to the test by inspecting how people divide their multimodal efforts (chapter 5). Prior work has shown that human languages are structured according to efficiency principles (for overviews, see Gibson et al., 2019; Levshina & Moran, 2021), and that people engaged in joint actions minimise their joint efforts (Clark & Brennan, 1991; Clark & Schaefer, 1987; Clark & Wilkes-Gibbs, 1986; Santamaria & Rosenbaum, 2011; Török et al., 2019, 2021). Joint actions involve people doing something together to reach a shared goal, such as having a conversation or moving a couch. Whereas prior work on linguistic efficiency and joint efficiency has focused on speech or non-linguistic behaviour, here I consider *multimodal* language use at the level of the dyad. I do so by taking advantage of the rich research history on repair. I use the well-defined sequential structure of other-initiated repair as a microenvironment where I can systematically quantify communicative efforts. In particular, I investigate how people divide their speech and gesture efforts across repair initiations and repair solutions, and test whether the division of multimodal effort across interactants promotes efficiency at the level of the dyad.

Methodology

For the empirical investigations (in chapters 3-5), I use a referential communication task. In this task, participants are presented with arrays of 16 novel figures (called “Fribbles”, after Barry et al., 2014) and they take turns to describe and find the figures over six consecutive rounds (see Figures 1.1 and 1.2).

The task used in this thesis has been developed by Sara Bögels and further refined in collaboration with the Communicative Alignment in Brain and Behaviour (CABB) team.[4] With the earliest forms dating back to the 1960’s (Krauss & Glucksberg, 1969; Krauss & Weinheimer, 1964), referential communication tasks remain a popular method till today, with researchers using variations of the task to answer a broad range of questions about language use and the emergence of referential conventions. In what follows I outline how the variant used here differs from earlier work, and how the task and set-up are well-suited for studying multimodal interactional practices in the context of negotiating mutual understanding.

[4] The Communicative Alignment in Brain and Behaviour (CABB) team is an interdisciplinary research group within the Language in Interaction Consortium (www.languageininteraction.nl), which this thesis project was embedded in. This thesis has been shaped by contributions of the (former) core members of the CABB team (in alphabetical order): Flavia Arnese, Mark Blokpoel, Sara Bögels, Mark Dingemanse, Lotte Eijk, Mirjam Ernestus, Judith Holler, Branka Milivojevic, Asli Özyürek, Wim Pouw, Iris van Rooij, Herbert Schriefers, Ivan Toni, James Trujillo and Marieke Woensdregt.

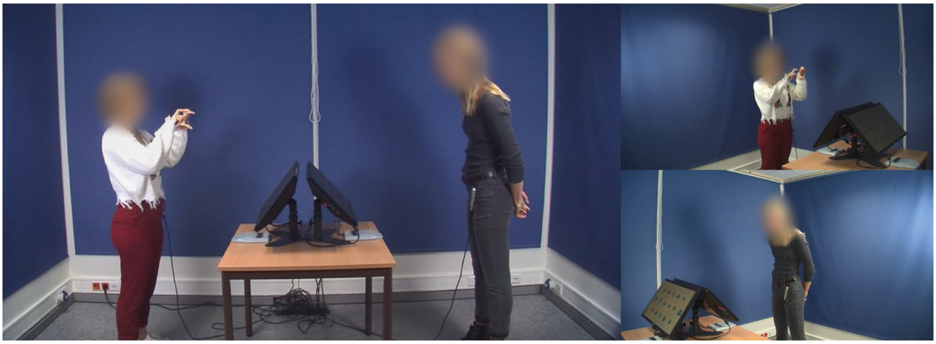


Figure 1.1. Participants performing the referential communication task as shown in stills from the cameras. Microsoft Kinect (V2) devices were used to record motion tracking data; these were positioned right next to and at about the same height as the two cameras capturing the individual participants, in order to record the body movements from roughly the same angle.

Let's first consider the "Fribbles" that were used as stimuli in this thesis, and how they differ from figures used in other referential communication tasks, such as "Tangrams" and "Greebles" (see Figure 1.2). The Fribbles were inspired by Barry et al. (2014), and are made up of a base shape with distinct protrusions attached to it, as is the case for Greebles (Hoetjes, Koolen, et al., 2015; Hoetjes, Krahmer, et al., 2015). For these stimuli types, participants are likely to describe the individual protrusions, especially in initial stages of the interaction. This yields (multimodal) referential expressions which are more compositional compared to the holistic expressions (and depictions) typically used for Tangram figures (e.g., "a person who's ice skating"; Clark & Wilkes-Gibbs, 1986). Of course, this is not a clear-cut distinction, as participants can use geometrical terms to describe individual parts of Tangrams ("triangle", "square", etc.), and abstract descriptions for the whole Fribbles ("robot", "wine glass"), sometimes even as their initial description in the first round. But of importance is that the Fribbles were pretested and designed (by Sara Bögels) to ensure different conceptualisations across items and individuals (Eijk et al., 2022). As such, the Fribbles required participants to negotiate referential expressions, yielding rich multimodal interactions with ample variation.

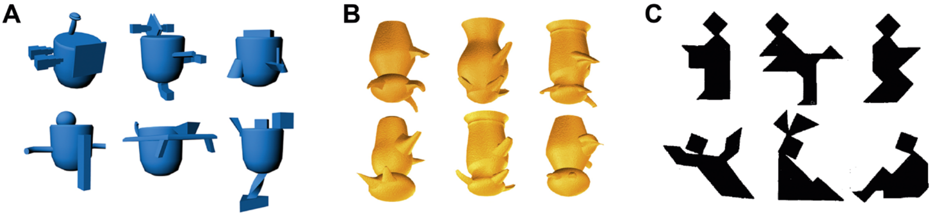


Figure 1.2. Stimuli in referential communication tasks; selection of six items from each type. Panel A: Fribbles used in this thesis research (adapted from Barry et al., 2014). Panel B: Greebles (Hoetjes, Koolen, et al., 2015; Hoetjes, Krahmer, et al., 2015; adapted from Gauthier & Tarr, 1997). Panel C: Tangrams (Clark & Wilkes-Gibbs, 1986; selected from Elffers, 1976).

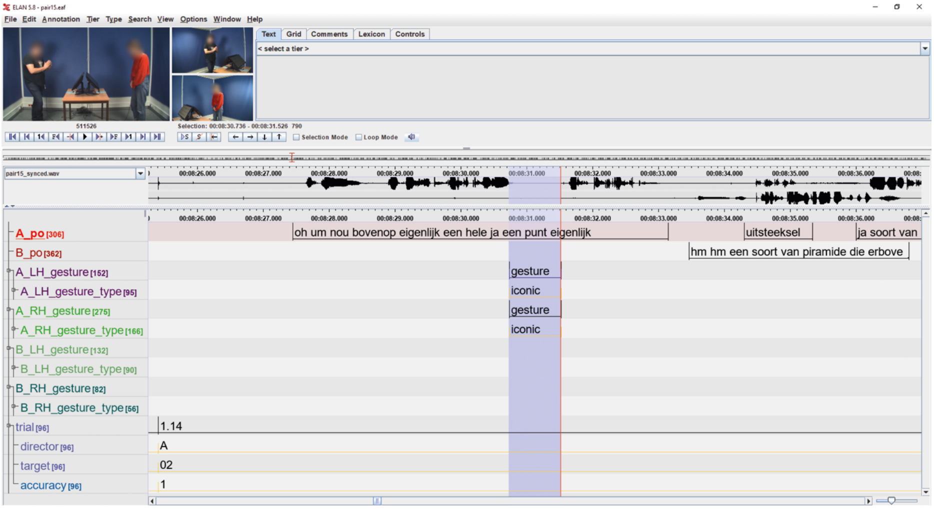
In referential communication tasks, participants are commonly assigned “director” and “matcher” roles for the duration of the whole task, but in the task used for this thesis participants switched roles after each trial. There are six rounds in the task, where each round consists of 16 trials, one for each Fribble. The switching of roles ensured that both participants of a dyad refer multiple times to each Fribble, enabling us to investigate whether they aligned their descriptions. This also more closely resembles naturalistic settings of symbol creation, where different people will refer to the same referent over time (see e.g., Haviland, 2013; L. Horton, 2020). Participants were instructed to “communicate in any way they wanted”; an instruction phrased to be agnostic with respect to the use of speech and gesture.

The lab set-up was created (together with members of the CABB team) in such a way as to support a holistic investigation of language use in this thesis, where I consider how people use both speech and gesture to make meaning together. During the referential communication task, participants faced each other while standing on opposite sides of a table with two monitors (see Figure 1.1). The monitors were slightly tilted so participants could easily view the screen and their partner, and the monitors were positioned at hip height to ensure mutual visibility of upper torso and main gesture area (McNeill, 1992). Head-mounted microphones were used to record the speech of the individual speakers, aiding the segmentation and transcription of overlapping speech. Three HD video cameras were used to capture each participant’s bodily movements from two angles (see Figure 1.1). Besides video recordings, I also collected motion tracking data using two Microsoft Kinect (V2) devices.

In sum, the task-based setting adopted in this thesis yielded interactions in which participants recruit multimodal utterances in relatively free-form in order to negotiate mutual understanding and jointly solve the task. Referring to novel objects which lack conventionalised labels is communicatively challenging, which likely pushes participants to go beyond tacit displays of understanding, resorting to the most forceful means to establish mutual understanding. As such, the design of the referential communication task enables me to quantitatively and qualitatively investigate if, how and when people jointly and multimodally recruit alignment (chapter 3) and other-initiated repair (chapters 4 and 5) in the process of joint meaning-making.

Yet it is important to keep in mind that task-based interactions in the lab are multiple steps removed from naturalistic co-present interactions. What the task adds is a degree of control and comparability that is harder to achieve in free informal interactions. Specifically, the fixed task structure facilitates the coding of communicative behaviour (since we know which target item directors refer to), and ensures comparability of quantitative findings across dyads. But we should be mindful of how task-oriented interactions in the lab differ from spontaneous dialogue (de Ruiter, 2013; Dideriksen et al., 2019; Kendrick, 2017), and interpret the findings accordingly. In this thesis I will do so by discussing how the insights of the quantitative and qualitative analyses relate to findings of studies on naturalistic interactions (chapters 3-5), and by replicating patterns in other-initiated repair sequences found in free co-present interactions (chapter 5).

For this thesis, data has been collected from 20 dyads performing the referential communication task, yielding about 8 hours of audio, video and motion tracking recordings.[5] Chapters 3-5 all make use of the audio-video recordings of this same dataset; in chapter 3 a subset is analysed (focusing on 10 out of 20 dyads and 8 out of 16 Fribbles), and in chapters 4 and 5 the whole dataset is analysed. Speech and co-speech gestures annotations were manually created in the multimedia annotation tool ELAN (see Figure 1.3, and more details in Appendix A), and were used in all three chapters. The motion tracking data is used in chapter 5. To analyse the transcribed speech, gesture annotations and motion tracking data, I bridge insights and methods from linguistics, gesture studies and human movement science.



oh um nou bovenop eigenlijk een hele ja een punt eigenlijk

uitsteeksel

ja soort van

hm hm een soort van piramide die erbove

gesture

iconic

gesture

iconic

1.14

A

02

1

Figure 1.3. Screenshot from an ELAN file. It includes the synchronised videos from three cameras (top left), the audio from two head-mounted microphones (waveforms in the middle), and annotations on various tiers (rows at the bottom of the window). The first two tiers include the transcribed speech for the participant on the left ("A_po") and right ("B_po"; where "po" stands for practical orthography). The speech annotations correspond to Turn Constructional Units; i.e., potentially complete, meaningful utterances (Clayman, 2013; Couper-Kuhlen & Selting, 2017; Schegloff, 2007). The next eight tiers include gesture annotations for the two participants (A, B), for the left and right hand separately (LH, RH). The annotations in the highlighted time window denote a two-handed iconic gesture produced by participant A. The remaining tiers pertain to the task structure and performance. This screenshot is taken in trial 14 of round 1, where participant A was the director, Fribble number 2 the target, and the matcher's selection correct.

[5] In collaboration with the CABB team, the communicative task used here has been further refined, and additional behavioural and neuroimaging tasks have been developed. Using this extended paradigm, another dataset has been collected together with CABB team members (most notably Sara Bögels and Lotte Eijk). This *CABB dataset* (*N* = 71 dyads) has been fully archived and documented, and has been made openly available to the scientific community (Eijk et al., 2022). The *CABB dataset* is not analysed in this thesis.

Further details about the task (in terms of stimuli, recording set-up and procedure) are described in chapter 3. The specific methodological approach for each study is presented in the corresponding chapter. The procedure for the transcription of speech and coding of gestures, alignment and other-initiated repair (as well as inter-rater reliability analyses) are described in the Method section of the individual chapters, and in more detail in Appendix A.

Thesis outline

The individual chapters of this thesis map onto the objectives presented earlier:

i) situate alignment in multimodal, sequential talk-in-interaction—both conceptually (chapter 2) and empirically (chapter 3);
ii) investigate the other-initiated repair system from a multimodal perspective (chapter 4);
iii) examine the division of multimodal labour between participants in social interaction from a joint action perspective (chapter 5).

How each chapter is set up to achieve those objectives is summarised below.

Chapter 2 presents an integrative framework to get a grip on the wealth of multidisciplinary research on alignment. I discuss the most prominent theoretical perspectives on alignment and review operationalisations of lexical and gestural alignment in earlier work. To make sense of the multidimensional nature of alignment and the diverse ways in which it is studied, I identify five key dimensions to formalise the relationship between pairs of behaviour: *time, sequence, meaning, form,* and *modality.* This framework provides us with the conceptual tools to capture the complex nature of alignment, and with common terminology to operationalise alignment in multimodal talk-in-interaction in a transparent and systematic manner.

Chapter 3 investigates how and when alignment in the spoken and gestural modalities comes about in the referential communication task where interactants collaboratively create labels to refer to novel referents. I used the framework presented in chapter 2 to operationalise alignment, and subsequently tracked the emergence of lexical and gestural alignment in a systematic manner (focusing on a subset of the data—10 out of 20 dyads, 8 out of 16 Fribbles—to keep the amount of hand-coded data manageable). For this chapter I build on earlier (psycholinguistic) work on alignment, as well as research on the role of modality in language emergence and development. These literatures show that alignment on the one hand, and gestural or multimodal communication on the other hand play a key role for establishing joint reference to (novel) referents. Here I take the next step by studying the role of modality *in* alignment in the context of a symbol creation setting. I quantitatively analyse when and how frequently lexical and gestural alignment emerges in the interaction. I add qualitative analyses to study how alignment (of speech and/or gesture) is deployed in particular sequential environments to effectively refer to novel referents. I predict that over the course of the interaction participants establish shared symbols for the referents (prediction 1), which I assessed with a naming task that people individually completed before and after the interaction. With respect to the interaction, I predict that multimodal alignment and lexical alignment will emerge more frequently than gestural alignment alone (prediction 2), as converging on a lexical item of a shared spoken language could be a more robust strategy than only aligning on non-conventionalised gestures. I also predict that if people align in both modalities, that this alignment is likely to emerge in speech and gesture simultaneously, or first in gesture and later in speech (prediction 3), because the gestural modality lends itself well for the production of motivated signs for novel referents (through iconicity and indexicality).

In **chapter 4**, I study how people incrementally build up understanding through coordinative, multimodal contributions in other-initiated repair sequences. While other-initiated repair has been studied extensively with respect to speech as well as some facial signals, here I single out the role of manual co-speech gestures as a domain where more progress can be made. Specifically, I argue that we can benefit from a holistic look at how they are distributed in the repair system as a whole, that is, looking at how gestures are used together with speech to initiate repair (i.e., to target trouble in a prior turn) and to resolve repair (i.e., in response to a repair initiation). I aim to answer this by quantifying the use of gestures in three different types of repair formats (open request, restricted request and restricted offer), across two sequential positions (repair initiation and repair solution). I complement this with qualitative analyses, where I inspect how different types of gestures are used across the different repair types and positions. I present examples where I discuss how the multimodal contributions of participants are contingent on their partner's prior turn, thereby showing how people coordinate their use of gesture and speech to solve problems with perceiving or understanding.

Chapter 5 studies the interactional dynamics of other-initiated repair as multimodal phenomenon, by investigating how people distribute the interactional work amongst them. Prior work on linguistic and non-linguistic joint actions has shown that people minimise their *joint* efforts. In dialogue, this has so far been studied by looking at speech efforts only, but the findings on the role of gesture in repair presented in chapter 4 give rise to a new question: how do people divide their *multimodal* efforts such as to minimise the overall effort for the dyad? I set out to replicate earlier findings found for unimodal interaction, and subsequently extend the hypothesis to multimodal language use: the more specific the repair initiation, the more multimodal labour is invested by the person initiating repair (relative to the person resolving the trouble). Furthermore, in accordance with findings in the joint action literature, I expect people to favour the repair initiation type which yields the least amount of multimodal effort for the dyad as a whole. To test these hypotheses, I quantify communicative efforts by combining linguistically informed measures of the speech and state-of-the art kinematic measures (based on Trujillo et al., 2019, 2021) derived from the motion tracking data for the gestures used in the interaction.

Finally, in **chapter 6**, I synthesise the findings of chapters 2-5. Based on the insights derived from this thesis, I argue that it can be useful to come to think of alignment as interactional resource, gesture as coordination device and social interaction as efficient joint action. This is followed by a discussion of methodological contributions and limitations, and directions for future research.

Chapter 2

Alignment in multimodal interaction: An integrative framework

This chapter is based on:
Rasenberg, M., Özyürek, A., & Dingemanse, M. (2020).
Alignment in multimodal interaction:
An integrative framework.
Cognitive Science, *44*(11), e12911.
https://doi.org/10.1111/cogs.12911

Abstract

When people are engaged in social interaction, they can repeat aspects of each other's communicative behaviour, such as words or gestures. This kind of behavioural *alignment* has been studied across a wide range of disciplines and has been accounted for by diverging theories. In this paper, we review various operationalisations of lexical and gestural alignment. We reveal that scholars have fundamentally different takes on when and how behaviour is considered to be aligned, which makes it difficult to compare findings and draw uniform conclusions. Furthermore, we show that scholars tend to focus on one particular dimension of alignment (traditionally, whether two instances of behaviour overlap in form), while other dimensions remain understudied. This hampers theory testing and building, which requires a well-defined account of the factors that are central to or might enhance alignment. To capture the complex nature of alignment, we identify five key dimensions to formalise the relationship between any pair of behaviour: time, sequence, meaning, form, and modality. We show how assumptions regarding the underlying mechanism of alignment (placed along the continuum of *priming* vs. *grounding*) pattern together with operationalisations in terms of the five dimensions. This integrative framework can help researchers in the field of alignment and related phenomena (including behaviour matching, mimicry, entrainment, and accommodation) to formulate their hypotheses and operationalisations in a more transparent and systematic manner. The framework also enables us to discover unexplored research avenues and derive new hypotheses regarding alignment.

Introduction

In social interactions, people coordinate their actions in an effort to incrementally and interactively reach their communicative goals. One component of such joint actions is cross-participant repetition of communicative behaviour. Work across a wide range of fields shows that when people are engaged in communicative interaction, their behaviours may grow to be in tune with each other at several levels: from body postures and eye gaze, to words and gestures. A key research objective within cognitive science is to gain a fuller understanding of this kind of behavioural *alignment* and how this can lead to mutual understanding. To answer this question, we benefit from adopting a broad perspective, by also considering work on related concepts, such as behaviour matching, imitation, mimicry, entrainment, repetition, and accommodation (which may serve other, partially overlapping cognitive or socio-affective functions).

To get a grip on the phenomenon of alignment, we need to start from the vantage point that natural communication is inherently multimodal, comprising both speech and such bodily behaviours as facial expressions, eye gaze, and co-speech gestures. *Co-speech gestures* are meaningful movements (usually of the hands or arms) that accompany speech. A subset of these are so-called *iconic* gestures, which visually depict object attributes, spatial relationships, or actions. Consider the following example from *The Late Late Show* with James Corden (an American late-night talk show). The talk show guests, Mila Kunis (M) and Christian Slater (C), are engaged in a conversation about the dating show *The Bachelorette*. In this show, one particular participant ("Chad") became known for always eating meat on camera.

(1) 1 C: Do you remember how crazy Chad was in that one sea-
2 M: The **meat [eating]**$_{\text{M}}$ **Chad**?
3 C: Yeah the **meat [eating]**$_{\text{C}}$ **Chad** guy

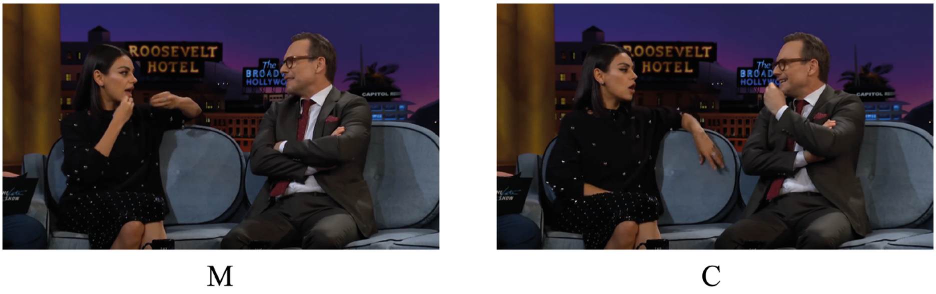


Figure 2.1. Alignment of speech and gestures by the talk show guests Mila Kunis (M) and Christian Slater (C).[6]

[6] Stills are reproduced under fair use from the video "Mila Kunis and Christian Slater are addicted to dating shows" by The Late Late Show with James Corden, 2018 (https://www.youtube.com/watch?v=B2Pcc_CSaK4, 2:40-2:41).

Square brackets indicate the start and end points of a gesture, and the capital letters correspond to the pictures shown in Figure 2.1. In this excerpt, M uses the lexical phrase "meat-eating Chad" along with an iconic co-speech gesture depicting the act of eating. C repeats both the lexical phrase ("meat-eating Chad"), as well as the eating gesture.[7] This kind of lexical and gestural alignment occurs regularly in both natural and task-based interactions, and has been shown to support joint problem-solving and coordination (e.g., Holler & Wilkin, 2011a; Pickering & Garrod, 2004).

Despite the emergence of various theoretical accounts, a comprehensive understanding of the phenomenon of alignment is still lacking. This is partly due to the large variation in methodological approaches. Take again Example 1 above. We can easily identify some form of alignment here in that both participants produce the same lexical phrase ("meat-eating Chad") and similar-looking gestures. This focus on alignment of form ties in with the traditional notion of behavioural alignment. However, in order to have a complete understanding of the phenomenon—when, how, and why it happens—there are other dimensions to consider. For example, some scholars quantify the extent to which the spoken utterances or gestures overlap in form, while others care more about the fact that both speakers used similar words or gestures to collaboratively refer to the same person. Some restrict analyses to alignment in speech *or* gestures, while others look at both. Some only focus on these cases of alignment in adjacent speech turns, while others also look for alignment of behaviours which are further apart in time. Design choices and measurement techniques vary both across and within fields, and they often (implicitly) follow from theoretical presuppositions. This makes it difficult to bring the findings together into an all-encompassing view of why and how alignment comes about for various types of behaviour in interactions.

Given the diversity of work in the interdisciplinary area of social interaction, some notes on terminology are in order. First, scholars have used the term "alignment" in different ways. In the most general sense, in the context of social interaction, alignment could be taken to mean interpersonal coordination between two communicators. The term (interactive) alignment was originally introduced by Pickering and Garrod (2004) to refer to the interpersonal alignment of mental representations underlying linguistic behaviour. However, various scholars have used the same term to simply refer to observable similarities in communicative behaviour itself (e.g., Bergmann & Kopp, 2012; Fusaroli et al., 2017; Howes et al., 2010; Oben & Brône, 2016). Of course, the two senses are related (since inferences about mental representations are often made on the basis of observed behaviour), but in light of theoretical discussions, it is important to keep them apart. Therefore, we differentiate between *behavioural alignment* and *alignment of mental representations*.

[7] However, note the difference in terms of handedness: M produces a two-handed gesture, while C only uses his left hand to illustrate the "eating". Some might therefore argue that these gestures are not aligned (or not "mimicked" or "matched"). Later in the paper we will further discuss such criteria regarding overlap in form.

Most of the empirical work reviewed in this paper is concerned with behavioural alignment. So we use terms like lexical alignment and gestural alignment with the intuitive meaning that two people produce similar lexical items or phrases or co-speech gestures (similar to the discussion of the example above). We make it explicit when we are referring to alignment of mental representations instead.

Second, what we call behavioural alignment here has been studied under a range of terms, and is part of a larger array of phenomena variously labelled behaviour matching, entrainment, accommodation, repetition, imitation, and mimicry. Though all of these terms target contingent behavioural similarities in socially interacting agents, each of them comes with its own disciplinary history. Thus, each carries its own commitments and implications with regard to the kinds of behaviour in focus, the embodied and interactional mechanisms at play, and the cognitive or socio-affective functions involved. While we opt for “alignment” as a widely used and relatively theory-agnostic term, a key contribution of our paper is to provide an integrative framework that can enable cumulative progress regardless of the precise label used.

With many fields now working toward empirical and theoretical accounts of alignment, it is crucial to have a shared framework that allows us to capture the space of possibilities of what can be considered alignment. By systematically tracking five dimensions along which communicative behaviours may relate to each other, we formulate clear and unambiguous terms of comparison that help to sharpen and contrast predictions of different theoretical approaches. We illustrate the utility of this framework by reviewing recent and foundational work on lexical and gestural alignment. Our approach makes visible how methodological choices and operationalisations tend to pattern together with assumptions regarding underlying mechanisms (for instance, *priming* vs. *grounding*), resulting in a situation where some areas of the space of possibilities are much better explored than others. We devote special attention to the interrelation of lexical and gestural alignment as one of the promising areas for future studies.

Theoretical approaches to alignment

Social interaction is an incredibly complex process, which has resulted in a diverse set of empirical and theoretical approaches. In the field of alignment, however, theoretical contributions are usually framed as belonging to one of two prominent “camps”, which could be denoted as *priming* and *grounding* (cf. Oben, 2018; also denoted automatic vs. strategic alignment (Kopp & Bergmann, 2013); and related to the distinction between Aggregate and Interactive approaches (Healey et al., 2018)). According to this dichotomy, priming accounts suggest that alignment involves an automatic, low-level priming mechanism that is confined to the individual’s mind (e.g., Pickering & Garrod, 2004, 2006), whereas grounding accounts argue that alignment follows from interactive, coordinative efforts involved in joint meaning-making (Brennan & Clark, 1996; Holler & Wilkin, 2011a).

The priming versus grounding juxtaposition falls short in two respects. First, because the accounts are not mutually exclusive, and second, because it does not do justice to the wealth of integrative theory on communication more generally (also beyond the specific phenomenon of alignment). Nonetheless, as shall be seen, empirical investigations of alignment often appear to be implicitly guided by either of the two perspectives. We will discuss these perspectives against the backdrop of a more inclusive set of theories on social interaction from various fields relevant to the study of alignment. We find that theories differ from one another on two key aspects: a) the extent to which they presume perspective-taking, and b) the relation they predict between alignment at various levels.

In some theories, cross-participant repetition of communicative behaviour is not considered to be produced "for" the conversational partner or with the partner's perspective in mind. For example, according to *direct mapping accounts*, the partner's behaviour directly activates the corresponding motor representations (through the mirror neuron system), which underlies the production of the same behaviour (e.g., Brass & Heyes, 2005; Dijksterhuis & Bargh, 2001; Heyes, 2011; Rizzolatti et al., 2001).[8] In a similar vein, the *interactive alignment account* (Pickering & Garrod, 2004) entails a parity between the representations used in comprehension and production, and therefore hearing a certain phoneme, word, or syntactic structure (which leads to the activation of the corresponding representation), "primes" the hearer to subsequently use it in his/her own speech production as well. "As part of this process, interlocutors do not model each other's mental states but simply align on each other's linguistic representations" (Pickering & Garrod, 2004, p. 180).

On the other end of the continuum are theories in which communicators carefully keep track of and adjust to their partner's perspective. For example, Clark et al. argue that people explicitly represent the information that is shared (and mutually known to be shared) with the communicative partner; that is, they keep track of their *common ground* (Clark, 1996). Using an object description task, it has been shown that people establish partner-specific shared conceptualisations of objects that become part of common ground (e.g., conceptualising a particular shoe as a "loafer"; Brennan & Clark, 1996). Communicators repeatedly refer to these *conceptual pacts* with the same words when talking to the same partner, thus yielding sustained lexical alignment (or using the original term: "lexical entrainment"), which they abandon when switching to another partner with whom the pact is not shared.

We could conceptualise priming versus grounding as being positioned on either end of a continuum, as they represent opposing ideas on the involvement of perspective-taking

[8] Though such accounts have recently been challenged, for example by arguing that this kind of automatic mimicry is a flexible and socially guided process (Chartrand & van Baaren, 2009; Lakin & Chartrand, 2003), which operates under top-down control of the mentalising system—a system recruited to infer other people's mental states or to make social judgements (Wang & Hamilton, 2012).

in alignment (see also the discussion of "mediated" vs. "unmediated" accounts of alignment in Branigan et al., 2011). However, there are also theories that assume more moderate perspectives. Such theories argue that having to always explicitly represent and fully adopt to the partner's perspective might be too costly in terms of cognitive resources, but is necessary to some extent or under some special circumstances. One such proposal is the idea that (language) processing takes place in two "stages": an early egocentric phase, followed by a later phase in which one might correct for the partner's perspective (e.g., W. S. Horton & Keysar, 1996; Keysar et al., 1998; Lin et al., 2010). In contrast, Brennan, Galati, and Kuhlen (2010) propose that communicators do engage in partner-adapted processing early on. However, they argue that rather than this resulting from a detailed representation of the partner's perspective, communicators make use of a simplified model, such as "my partner knows X" or "my partner does not know X" (so called one-bit partner models).

Coming to the second key aspect, namely the level(s) at which alignment is presumed to take place, it should first be noted that both priming and grounding approaches are concerned with alignment of behaviour (called *repetition* or *entrainment*) as well as alignment of higher-level mental representations (*situation models* or *conceptual pacts*). However, they differ in how they theorise alignment at distinct levels. According to priming accounts, speakers do not only observably align their speaking behaviour, but also the linguistic representations underlying that behaviour: "they have aligned linguistic knowledge to the extent that they have similar patterns of activation of linguistic knowledge" (Pickering & Garrod, 2006, p. 215). Furthermore, the priming mechanism is argued to operate at multiple linguistic levels (from phonetics to semantics), where alignment at one level leads to alignment at other levels, ultimately resulting in alignment of situation models.

In approaches taking a grounding perspective, alignment of linguistic representations is not a requisite for alignment at other levels of representation. Conceptual pacts are formed through a process of *grounding* interactional contributions (Brennan & Clark, 1996; Clark & Brennan, 1991), which can happen in various ways. For example, alignment can be used to signal understanding, thereby grounding a certain referring expression (as could be argued for the "meat-eating Chad" in Example 1). However, alignment of communicative behaviour can also occur in the form of other-initiated repair, thus signalling *misunderstanding* as a means to *get to* higher level alignment, rather than being an indicator of it (e.g., Mills & Healey, 2008), as in the following example from Clark and Wilkes-Gibbs (1986):

(2) A Uh, person putting a shoe on.
B Putting a shoe on?
A Uh huh. Facing left. Looks like he's sitting down.
B Okay.

Furthermore, (purposefully) using different words or gestures can also be a way to establish mutual understanding (e.g., Clark & Wilkes-Gibbs, 1986; Holler & Wilkin, 2011a; Tabensky, 2001). For example, Holler and Wilkin (2011a) describe a situation where participant A referred to a figure with the lexical phrase "an ostrich", to which participant B replied "Yeah, okay that, that looks like a woman to me, kicking her leg up behind her, yeah?" (though interestingly both produced the same gesture along with the speech, as further discussed in the section "Modality"). Using the terminology of Clark and Wilkes-Gibbs (1986): the presentation of participant A was not accepted by participant B, who used the repair strategy *replacement* in an effort to get to a shared conceptualisation of the figure.

Thus, in grounding accounts there is a flexible relationship between behavioural alignment (in various modalities or linguistic levels) and alignment of conceptual representations, while for priming accounts this is presented as causally linked, with alignment percolating across all levels.

In general, in the cognitive sciences, cross-participant repetition of communicative behaviour has been theorised as involving shared representations—be they shared linguistic or conceptual representations as just discussed, or shared motor (Rizzolatti et al., 2001), goal (Bekkering et al., 2000; Wohlschläger et al., 2003) or task representations (Sebanz et al., 2006). Yet there is a class of theories that attempts to account for human interaction without appealing to mental representations, namely dynamical systems theory (for an insightful overview, see Dale et al., 2013). For example, Shockley, Richardson and Dale (2009) propose that interpersonal coordination can be thought of as a "coordinative structure—a self-organized, softly assembled (i.e., temporary) set of components that behave as a single functional unit" (p. 313), which does not necessarily involve higher level cognitive representations. This means that when talking about *alignment,* it is important to first of all distinguish empirically observable alignment of behaviour from the presumed alignment of mental representations. And for the latter, to differentiate between alignment of various kinds of representations (motor, linguistic, etc.), as theories make different claims about their involvement and interrelations in social interaction.

A framework for understanding and investigating alignment

In order to go beyond these existing theoretical approaches, we have to outline the space of possibilities of how alignment is conceptualised and measured across studies. Generally speaking, all studies of alignment compare behaviour from person A with behaviour from person B. These behaviours can be discrete events (e.g., one gesture) or streams of behaviour (e.g., a series of consecutive body movements). When these behaviours are *aligned*, they are considered to be "the same" or "matched" in one way or the other. That is, the units of analysis are cross-participant paired behaviours, where A's behaviour is similar to B's behaviour on one or several dimensions. The term "prime-target pairs" is commonly used in controlled experiments on alignment. We use *paired behaviours* here as a more neutral term that does not presuppose a particular methodological or theoretical approach and is agnostic about the mechanism behind the pairing.

Empirical studies show considerable variation with respect to the dimension(s) they take into consideration, and how they operationalise alignment. Most studies use similarity in form as a criterion for alignment, with various definitions and measures of form overlap. However, the relation between the two instances of behaviour on other dimensions is often taken for granted or ignored. This is problematic, as this is where theoretical approaches might have diverging hypotheses. In order to move forward in the field, we need a tool to sharpen and contrast predictions of different theoretical approaches, and to operationalise experimental studies accordingly.

In an effort to clarify and reveal (oftentimes implicit) differences in what is considered to be aligned, we introduce a common integrative framework to decompose the notion of behavioural alignment into its constituent dimensions. We consider five key dimensions that help characterise the relation between any pair of behaviours: time, sequence, meaning, form, and modality. The framework is presented below, where we outline the dimensions in terms applicable to all kinds and levels of verbal and nonverbal behavioural alignment, be it posture or gesture, phonetics or syntax. For illustrative purposes, we use rectangular shapes as instances of behaviour, which are produced by two interlocutors (A and B), as shown in Figure 2.2.

We consider the five dimensions to be inherent to all kinds of paired behaviours. The relation between any two behaviours (or streams thereof) can always be described and analysed in terms of time, sequence, meaning, form, and modality. In the following, we outline what it means for paired behaviours to be related on these five dimensions, and we explain how the dimensions can be employed in empirical studies (Table 2.1).

First, behaviours have a relation in terms of *time*. The temporal lag between paired behaviours can vary from none (in the case of simultaneous production), to a delay of several (milli)seconds or minutes (e.g., as a result of intervening filler trials; Hartsuiker et al., 2008), and can go up to hours or even days. When dealing with multiple streams of behaviour, one can also observe the temporal relation of those time series, for example, in terms of synchrony or convergence in multiscale clustering (Abney et al., 2014).

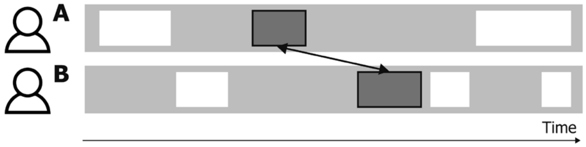


Figure. 2.2. Visualisation of an interaction between two people. Every rectangle represents an instance of behaviour. The behaviour can be of various types (i.e., the rectangles could represent syntactic constructions, lexical choices, mannerisms, co-speech gestures etc.) and units of analysis (e.g., the rectangles could represent discrete events or a stream of behaviour). The arrow indicates a possible comparison between two instances of behaviour.

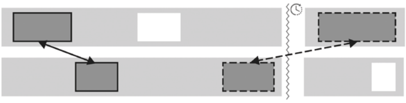
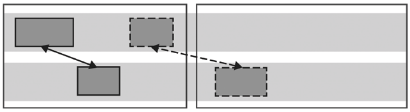
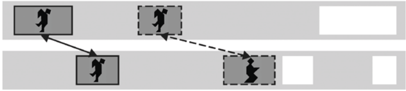
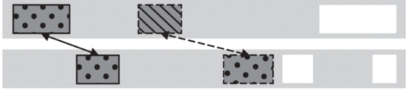
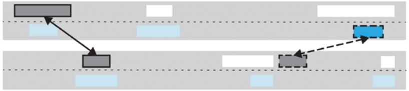
Second, paired behaviours may or may not occur in a conversational *sequence*. A key property of human interaction is that participants take turns, where each turn has a particular sequential relation to a prior turn in the discourse. A clear example of this are “adjacency pairs”; pairs of utterances where the latter is functionally dependent on the first, such as offer-acceptance or question-answer (Schegloff & Sacks, 1973). At a higher level, one or several of such pairs together can constitute a course of action (Levinson, 2013; Schegloff, 2007), such as scheduling a meeting. In a similar vein, task-based interactions can have an experimentally imposed sequential structure in terms of games (e.g., in the Maze task; Garrod & Anderson, 1987) or trials (e.g., in picture description tasks; Branigan et al., 2000; or referential communication tasks; Holler & Wilkin, 2011a). The sequence dimension captures the fact that paired behaviours in a question-answer sequence have a different relation to each other than, say, paired behaviours across experimental trials.

Third, paired behaviours can be related to one another in terms of their *meaning*. Here the type of behaviour plays an important role, as it is more meaningful to talk about whether people mean the same thing with a particular word they utter (Garrod & Anderson, 1987), compared to say, the pitch or foot-wiggles they produce. Furthermore, it is important to note here that we are moving into the domain of alignment of mental representations (as we cannot empirically observe semantics or reference directly), rather than alignment of visible (and directly measurable) characteristics.

Fourth—and this is the most intuitive and well-studied dimension—paired behaviours can be more or less similar in *form*. For example, one could measure the (dis)similarity in the syntactic composition of two utterances (Reitter & Moore, 2014), the extent to which two spoken words have similar acoustic attributes (Pardo et al., 2017), or whether two body movements are contra- or ipsilateral (Bavelas et al., 1988).

Finally, paired behaviours can occur in the same or in a different communicative *modality*. For example, lexical items produced in the spoken modality could be compared to other lexical items in the written modality, or to co-speech gestures (Tabensky, 2001) or facial expressions (Bavelas et al., 2000) produced in the visual modality. The dimension of modality captures the mode in which paired behaviours are produced and interpreted.

Table 2.1. A multidimensional framework for understanding and investigating alignment[a]

Time	The temporal distance between the first and second part of a pair of behaviour can be a short interval (e.g. simultaneous production or a split-second delay) or a long interval (varying from one or multiple turns, several minutes or even hours).	
Sequence	The sequential relation between any pair of aligned behaviour can vary from occurring within a certain sequence (e.g., the behaviour occurs within the same trial, as indicated by the larger rectangles in the figure), to transcending such sequential boundaries.	
Meaning	For levels of behaviour which convey meaning (e.g. lexical items or gestures), any pair of behaviour can vary from conveying the same meaning or referent to conveying different meanings.	
Form	The two parts of a pair of behaviour can vary from being exact copies, to having little or no overlap in form or shape.	
Modality	The two parts of a pair of behaviour can be produced in the same modality (e.g., the two pair parts are both spoken sentences), but can also be produced in different modalities (e.g., the first pair part is a lexical phrase, and the second pair part an iconic gesture).	

[a] The relationship between the two parts of a behaviour-pair can vary on five dimensions, as outlined in this table. For each dimension we visualise two different relationships between instances of behaviour; one with a solid arrow and one with a dashed arrow. For meaning, we use Tangram figures to visualise the referent of speech and/or gestures (cf. Clark & Wilkes-Gibbs, 1986; Holler & Wilkin, 2011a).

Though it is clear that studies have operationalised alignment in different ways, our framework makes visible that sometimes they have focused on different dimensions altogether, or have applied them in fundamentally different ways. Any particular dimension can be used as a *grouping criterion* or as a *measurement variable* in an empirical study. For example, one might restrict analyses to adjacent behaviours or adjacent speech turns, and quantify the extent to which they overlap in terms of form (i.e., sequence as grouping criterion, form as a measurement variable; e.g., Bergmann & Kopp, 2012; Fusaroli et al., 2017). Or one could search for all behaviour of a particular form and modality, and quantify their temporal relations (i.e., form and modality as grouping criterions, time as a measurement variable; e.g., Louwerse et al., 2012). Clearly, although all of these studies can be described as investigating "alignment", the operationalisations are so different that one may question their commensurability. One goal of our framework is to make it more straightforward pinpoint similarities and differences.

Some of the dimensions are interrelated. For example, when two instances of behaviour occur within a certain sequence (e.g., in a question-answer pair), this naturally has consequences for the temporal relation (i.e., the two pair parts are likely to have only a short temporal lag). However, it is possible to experimentally tease them apart, for example, by manipulating the presence of intervening material between a question (prime) and an answer (target), thereby increasing the temporal lag while retaining sequential cohesion (Levelt & Kelter, 1982). Another interdependency becomes apparent when comparing instances of behaviour which are produced in different modalities (e.g., a lexical phrase is compared to an iconic gesture), as here the dimension of form will become less relevant. Due to these interrelations, certain dimensions can become conflated or taken for granted in both empirical and theoretical approaches. Yet it is crucial for work on alignment to treat the dimensions as conceptually distinct from each other and to specify the relationship between two instances of behaviour for each dimension separately. We will corroborate this in the next section, in which we apply the framework to studies on lexical and gestural alignment.

A review based on the framework

This section will illustrate how we can use the five dimensions introduced in the previous section to characterise and compare previous studies on alignment in a systematic manner. We will start each subsection by reviewing the range of empirical possibilities for incorporating that dimension when studying alignment, and we will conclude each section by discussing how these operationalisations relate to the two theoretical approaches (priming and grounding). By doing so, we will show which dimensions are of fundamental importance in various empirical and theoretical accounts, and which dimensions are understudied. We will zoom in on lexical and gestural alignment, though in essence this practice can be applied to work on alignment at all linguistic levels or kinds of behaviour, making the current discussion of relevance to the field as a whole.

We will restrict our focus to studies investigating spontaneous, interactive dialogues (free conversations or task-based), thus excluding studies with interactions which are (partly) scripted, or lack natural turn-taking and feedback (e.g., Kimbara, 2008; Mol et al., 2012). Moreover, we will narrow the focus to studies on lexical alignment at the word level (thus excluding alignment of syntax or phonology, as well as higher level pragmatic levels, such as dialogue acts, e.g., Louwerse et al., 2012) and co-speech gestures (thus excluding bodily behaviour such as posture, e.g., Chartrand & Bargh, 1999). Note that this is not intended to be a complete review of all studies in the field, but instead an illustration of the range of empirical and theoretical approaches for studying alignment, and how they can be positioned in the overall possibility space.

Time

The time dimension can be used as a grouping variable by defining a particular temporal lag between aligned pairs of behaviour. In (priming) experiments, such lags can be experimentally controlled, for instance, by varying the amount of fillers items that appear between prime and target (Mahowald et al., 2016). In corpus studies, alignment could be operationalised as having to occur within a pre-defined temporal window. A useful technique for the latter approach is that of *time-aligned moving averages* (TAMA), where a specific time window (e.g., of 40 seconds) is shifted across the time axis in a stepwise manner. However, this has mostly been used for analyses of prosody (e.g., De Looze et al., 2014), and it is not common for lexical or gestural alignment (but see Oben, 2015).

In contrast, there are studies where for a gesture or lexical pair to count as aligned, there are no restrictions on the amount of time which can intervene. These are typically qualitative studies, which use a descriptive or exploratory approach (e.g., Kimbara, 2006; Tabensky, 2001; Tannen, 1989), though it also applies to some quantitative studies (e.g., Holler & Wilkin, 2011a).

The importance of methodological choices regarding time restrictions might be downplayed, because alignment often occurs with a split-second delay or is intervened by one or a few turns, which means that both approaches will yield a highly similar selection of cases. However, the paired behaviours that are part of the analyses can still differ considerably across studies: whereas in the work by Oben (2015) only gestures that occurred within a window of 40 seconds were considered for alignment, in the study by Holler and Wilkin (2011a) the gestures could be as far apart as several minutes (though the actual time lags are not reported), as long as they were referring to the same referent (see the section “Meaning” below).

Instead of selecting candidate paired behaviours based on a pre-set time window, one can also measure the overall temporal coupling of two streams of behaviours and determine temporal lag more dynamically. For example, Louwerse, Dale, Bard and Jeuniaux (2012) analysed the multimodal interactions of participants engaged in a route communication task (Map Task; cf. Anderson et al., 1991). They investigated the temporal dependencies of “matched” verbal, facial, and gestural behaviours using Cross Recurrence Quantification Analysis (for a discussion of this method, see Fusaroli, Konvalinka, et al., 2014). This yielded average time intervals per behaviour category, such as 25 seconds for deictic (i.e., pointing) gestures. Going beyond such analyses of synchronisation, it also possible to measure convergence in multiscale clustering of behavioural events. To our knowledge, this has not yet been applied to lexical or gestural behaviour, but there is promising work that captures the temporal clustering of speech acoustics using power law distributions (Abney et al., 2014).

How the dimension of time is used often relates in complex ways to one's theoretical assumptions and hypothesised mechanisms. Alignment across large time intervals is less likely to be considered in studies working from a priming approach, as priming effects are hypothesised to decrease over time.[9] From a grounding perspective, a similar prediction can be made for natural interactions, given that topics vary over the course of interactions, thereby decreasing the relevance of certain conceptual pacts and the need to keep repeating certain lexical items or gestures. However, with respect to the grounding perspective, interlocutors have also been shown to repeat words after long temporal lags in free conversations, for example to re-introduce a topic or tie back to a problematic turn which was produced earlier in the conversation (Dingemanse et al., 2014; Sacks, 1992; Schegloff, 2000). Thus, in general, both theoretical accounts would argue that over time the likelihood of encountering behavioural alignment decreases, though for priming this effect would be mechanistic in nature (due to decreased levels of activation), while for grounding it would be more incidental (related to changes in joint projects and topics).

In addition to considerations regarding (the lack of) restrictions on the *maximum* time interval, the *minimum* time interval is also relevant. Words or gestures are sometimes produced simultaneously by two speakers, for example, when they interrupt each other or co-produce an utterance (cf. Holler & Wilkin, 2011a; Tannen, 1989), which is a well-documented phenomenon in conversation analysis (e.g., Lerner, 2002). Yet besides a methodological challenge, such cases are also a challenge for theoretical accounts based on priming as the underlying mechanism (if the particular word or gesture had not yet been produced prior to that moment). Such cases might be better explained from the grounding perspective, coupled with an account of incremental and predictive sentence processing.

Sequence

Paired behaviours do not merely stand in a temporal relation to one another; often they also occur within or across larger conversational sequences. The sequence dimension has been used as both a grouping and measurement criterion in studies on lexical and gestural alignment. At the outset, sequence can be employed to define which part of an interaction will be included in the analysis. For example, Chui (2014) qualitatively investigated gestural

[9] Though the term priming generally refers to a short-term, automatic effect, a case has also been made for the existence of so-called "long-term" priming (cf. Pickering & Garrod, 2004), which has been shown to even persist over the course of a week (Kaschak et al., 2011). See also Reitter et al.'s model (2011) which differentiates between short-term priming and long-lasting adaption.

alignment in 12 short stretches of talk in free interaction, in which people communicated about the meaning of a referent. Thus, analyses were restricted to co-speech gestures that were produced in a specific conversational sequence. Alternatively, in quantitative studies, conversations have also been studied as one large chunk, without the differentiation into sequences. For example, Bergmann and Kopp (2012) compared all iconic and deictic gestures from 25 dyads engaged in a spatial communication task (alternating direction-giving and sight description), yielding a total of 3,993 cross-participant gesture comparisons for the analyses.

Once the to-be-analysed data have been selected, a possible approach is to look at paired behaviours which are in a specific sequential relation to each other. For example, alignment can be analysed on the speech turn level; that is, one compares the behaviour in turn *x* from speaker A and in the following turn *y* from speaker B. Thus in this case adjacent speech turns are taken as the unit of analysis,[10] where the aligned lexical item or gesture can occur in any position within those turns (e.g., Fusaroli et al., 2017, for lexical alignment in free interaction and Map Task interactions). It is also possible to look at adjacent behaviour independent of speech turns, for example, by comparing a gesture that depicts a particular object with the next gesture that is produced (by the other speaker) to depict that same object in a spot-the-differences game (Oben & Brône, 2016). Hence, behaviours are "grouped" based on their sequential relation—in this case, adjacency. Note, as mentioned earlier, that this is related to the dimension time, because sequential adjacency usually implicates a relatively short temporal distance between the two pair parts.

It is also possible to completely abstract away from sequential structure, for instance by simply comparing all instances of a kind of behaviour category (such as iconic gestures) from both interlocutors (cf. Bergmann & Kopp, 2012; Louwerse et al., 2012) or by restricting analyses to predefined time windows (Oben, 2015). Such approaches lend themselves to large-scale quantification at the cost of losing sight of fine-grained sequential dependencies in the data.

Sequence can also be used as a measurement criterion, by taking any set of paired words or paired co-speech gestures, and investigate or "measure" their sequential relation. For instance in Chui (2014), co-speech gestures were investigated in terms of their sequential position, by dividing each stretch of talk into three different "phases"—a presentation, collaboration, and acceptance phase (see also Holler & Wilkin, 2011a, for a similar approach). Identifying the sequential relation of paired behaviours is mostly done by those who see alignment as an interactive grounding process. Qualitative work in this tradition has shown that immediate repetition of words in the following turn could be used to initiate repair, express surprise, answer a question, or accept a formulation, to name a

[10] Note that we use the term "adjacency" here in the simple sense of adjacent or neighbouring; not to be confused with the conversation analytic term *adjacency pairs*, as referred to earlier.

few (Dingemanse et al., 2014; Norrick, 1987; Rossi, 2020). Quantitative work has confirmed this for the sequential environment of interactive repair: there is a significantly larger likelihood of finding alignment in adjacent turns in *repair* sequences (consisting of a problematic turn followed by a repair-initiation) compared to other adjacent turns (Fusaroli et al., 2017). Such repair sequences, which are quite frequent, show that some forms of alignment can be the result of explicit coordination. In contrast, from a priming perspective that sees alignment as low-level and automated, sequential structures of the discourse would be deemed irrelevant.

Analytical approaches are shaped by research traditions and theoretical stances. In fact, we could derive opposing predictions from the two theoretical accounts: whereas based on priming accounts we would expect equal amounts of alignment across turn pairs irrespective of sequential organisation (as long as the temporal distance is the same), based on grounding accounts we could expect higher amounts of alignment in turns that stand in a specific sequential relation to each other (e.g., repair sequences). Besides hypotheses related to repair or adjacency, from a grounding perspective one could also expect to find more alignment *within* a project or course of action rather than across such sequential boundaries, while from a priming perspective one would again hypothesise equal amounts (as long the temporal distance is matched). Different levels of behaviour may be differentially susceptible to sequential organisation. Here we have an interesting test bed for contrasting or conciliating priming and grounding approaches, with ample opportunities for new research.

Meaning

The meaning dimension captures the observation that paired behaviours which have a clear relation in terms of time, sequence, and/or form might not always overlap in terms of their meaning. Especially in challenging communicative situations, such as a Maze Task, identical words are sometimes used to denote different things (Garrod & Anderson, 1987; Mills & Healey, 2008). With respect to co-speech gestures, it is evident that they are highly context-dependent and two similar gestures can mean completely different things in distinct contexts. Hence, this dimension is an important characteristic of lexical and gestural alignment, but in contrast to the other dimensions, generalises less well to alignment of linguistic behaviour at lower levels and bodily behaviours (most of which do not convey semantic meaning).

Seeing meaning as a separate dimension also helps to differentiate lexical and gestural alignment from the notions of *semantic alignment* (e.g., Dideriksen et al., 2019) and *semantic co-ordination* (e.g., Garrod & Anderson, 1987). Lexical and gestural alignment are generally understood as the repetition of words or gestures independent of the meaning conveyed; in terms of our framework, form is privileged over meaning. A possible empirical approach in line with this notion of alignment is to search for cross-participant repetition of (lemmatised) words in transcripts (cf. Fusaroli et al., 2017; and the Python package ALIGN by Duran et al., 2019) or to measure the form similarity of gestures (e.g., Bergmann & Kopp, 2012). While degree of form overlap is empirically observable, for comparing meanings we must rely on inferences and contextual anchoring. One approach is to manually code for semantic relations between lexical behaviours, for example, by categorizing phrases into "families" of confidence expressions in a joint perceptual task (Fusaroli et al., 2012) or into "mental model" of maze configurations in a maze game (Garrod & Anderson, 1987). More automated measures of semantic relations have also been employed recently, such as the use of word embeddings in a high-dimensional semantic space (Dideriksen et al., 2019; Duran et al., 2019) or conceptual/semantic recurrence quantification analysis (Angus & Wiles, 2018).

The meaning dimension can also be used as an additional grouping or selection variable in studies on lexical and gestural alignment. For example, in a study by Holler and Wilkin (2011a), iconic or metaphoric gestures are only considered to be aligned (in their terms: "mimicked") when they have some similarity in their form *and* represent the same meaning. Similarly, in Oben and Brône (2016), words or gestures are only considered to be aligned when they refer to the same referent in a spot-the-difference game. This is in sharp contrast with, for example, Louwerse et al. (2012), where alignment is operationalised as mere formal similarity in some time window, without reference to meaning (e.g., two gestures are considered to be "matched" when they are both *deictic* gestures, irrespective of the referent that was pointed to).

There are various ways to examine the semantic overlap between instances of behaviour, which is often far from trivial. In qualitative studies on lexical or gestural alignment in free conversation (e.g., Kimbara, 2006; Tabensky, 2001; Tannen, 1989), researchers rely on the discourse context to know whether the interlocutors are referring to the same thing or just happen to use the same word or gesture to denote something else. Task-based approaches have the benefit that the researchers can experimentally control and keep track of the referents that the participants verbally or gesturally refer to. Examples are Brennan and Clark (1996), Clark and Wilkes-Gibbs (1986), and Holler and Wilkin (2011a); in these studies, participants refer to objects on cards, over multiple rounds, which enables the researchers to track the referring expressions to particular objects over longer distances of time. However, there is rarely an exhaustive correspondence between the semantics of words or gestures and the referent they signify. This is because participants can talk about the *same referent*, yet lexically or gesturally single out different semantic properties; for example, when using the word "straight" or a gesture to depict the orientation versus shape of (a part of) an object. And conversely, words or gestures about *different referents* could still be semantically related. For example, in matching tasks with Tangram figures, participants might lexically align on basic-level categories such as heads, arms, etc., which they apply to all stimulus items (Bangerter et al., 2020).

In the monolingual spoken or written settings most often studied in psycholinguistics, the meaning and form dimensions of alignment can be hard to disentangle. We have highlighted here the potential of multimodal interaction for investigating semantic convergence and divergence. Multilingual interaction (Byun et al., 2019; Costa et al., 2008; Gries & Kootstra, 2017; Schneider et al., 2020) offers another promising and understudied environment in which these dimensions can be teased apart to varying degrees.

The meaning dimension draws the clearest line between the priming and grounding approaches. Priming approaches argue that a low-level, automatic mechanism results in form overlap in behaviour, which can lead (or "percolate") to alignment of semantic representations or vice versa, without semantics as a necessary guiding factor. Grounding approaches, on the other hand, regard instances of behaviour as means to negotiate and calibrate mutual understanding, so they expect alignment to occur when there is a semantic or referential link between the instances of behaviour.

Form

The form dimension in our framework reflects the fact that some degree of form similarity is the sine qua non of most notions of behavioural alignment in the literature. Lexical and gestural alignment can occur in various ways. For example, with respect to lexical alignment, interlocutors can repeat their partner's words or phrase literally, or repeat with variation, such as turning a statement into a question or vice versa (Fusaroli et al., 2017; Tannen, 1989). Though some also consider rephrasing or paraphrasing to be forms of "repetition" (e.g., Tabensky, 2001; Tannen, 1989) or "linguistic alignment" (Fusaroli et al., 2012), most studies adopt a more conservative notion of lexical alignment, requiring the repetition of a particular base word or lemma, thus excluding synonyms or paraphrases (cf. Fusaroli et al., 2017; Howes et al., 2010; Oben & Brône, 2016). However, studies vary considerably in the units of comparison; whereas some work with complete speech turns (Fusaroli et al., 2017), others only include content words (Brennan & Clark, 1996), or even a more restricted subset such as high-frequency words (Nenkova et al., 2008), nouns and verbs (Bangerter et al., 2020), or only nouns (Oben, 2015).[11] Obviously the degree of detected alignment can differ dramatically as a function of which terms are included in the comparison.

Similar to lexical alignment, "gestural rephrasing" has also been considered as a form of "repetition" in the gestural modality (e.g., Tabensky, 2001). However, most studies on gestural alignment require at least some degree of form resemblance, though studies vary with respect to how this is measured. Studies focusing on gestural form similarity generally

[11] The exact operationalisation of such constructs is not straightforward either. For example, there has been ample debate about what constitutes a speech turn, and how they can be recognised in conversations (Selting, 2000). Referring expressions or noun phrases can also be problematic units of analyses in natural interaction, due to the frequent occurrence of ellipsis and grammatically incomplete utterances.

analyse *iconic* co-speech gestures, which are spontaneous, idiosyncratic gestures where plenty of variations in form are possible. This is in contrast to *deictic* gestures (i.e., pointing gestures), *emblems* (such as the thumbs-up gesture), and *interactive* gestures (such as *beats* or *palm-up open-hand* gestures), which have more conventionalised forms. Most studies on gestural alignment are based on manual coding, where form overlap has been operationalised in terms of *mode of representation* (or representation technique, e.g., the hands can draw the outline of an object, enact a certain action, etc.; Streeck, 2008), specific form features or a combination of those. Recent advances in the field point to the promise of automated measures for quantifying the kinematic resemblance of gestures in terms of their velocity, size, distance, etc. (Pouw & Dixon, 2020).

Several studies on gestural alignment use mode of representation as a grouping variable. Oben and Brône (2016) used overlap in mode of representation as their primary criterion for considering gestures to be aligned (thus ignoring such features as motion or position), while Holler and Wilkin (2011a) used it along with the requirement to have the same overall shape/form (where some variability in handshape or position was accepted, but not in handedness). As an example of how mode of representation is used as a criterion, see Figure 2.3 below:

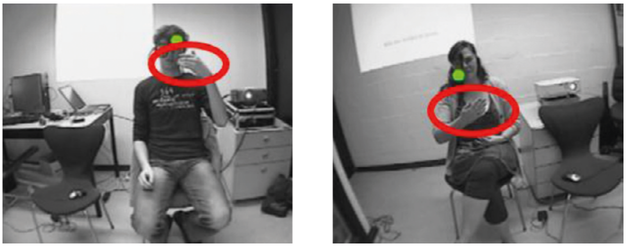


Figure 2.3. Gestures with overlap in "modelling" as the mode of representation, reproduced with permission from Oben and Brône (2016).

Here, both participants gesturally depict the target object DOOR, where the hand is a "model" for the object. The gestures differ in terms of handedness, finger orientation, and the tension in the handshape. However, Oben and Brône "still consider it to be an instance of gestural alignment because the representation technique is identical (i.e. modelling)" (2016, p. 37).

The other approach is to compare gestures on a number of form features. For example, Bergmann and Kopp (2012) investigated gestural alignment separately for mode of representation and other form features (handedness, handshape, palm- and finger orientation, and wrist movement type). Chui (2014) coded whether gestures overlapped in terms of handedness, handshape, position, motion, and orientation. Of the 12 gesture pairs in the analyses that were identified as "mimicked", 11 pairs showed overlap in four or five form features, and one pair in three features. See for example the following gesture pair (Figure 2.4):

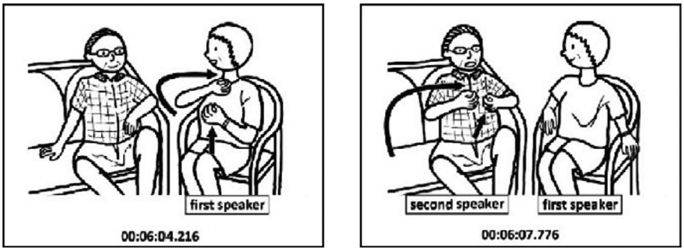


Figure 2.4. Gestures with overlap in handedness, handshape, position and motion, reproduced with permission from Chui (2014).

Here, both speakers gesturally depict a musical instrument: both use two hands (overlap in handedness), with the fingers curled into fists (overlap in handshape), facing each other in front of the chest (overlap in position), moving one hand to enact the idea of moving a bow (overlap in motion). However, as Chui notes, there is some deviance in the orientation of the lower hand (as the second speaker rests his arm on the sofa). She concludes that “in considering the five features together, the deviance in the hand/finger orientation, but the high consistency in the other four features did not affect the conclusion of the analysis that the two gestures were highly similar gestures for the same referent” (p. 73).

Both priming and grounding perspectives use form as their main criterion for considering behaviour to be aligned. However, there are important differences in the role form overlap plays from a theoretical point of view. Whereas priming is considered to naturally result in form overlap (due to activation of motor plans or linguistic representations), grounding perspectives consider more explicit coordination to (also) play a role. Furthermore, as discussed in the section “Theoretical approaches to alignment”, under grounding accounts there is no necessary relationship between alignment of behaviour in terms of form on the one hand, and mutual understanding, on the other—as understanding can also be achieved through repetitions with variations in form, or even through the production of completely different words or gestures altogether rather than aligning. Consequently, those working from a priming perspective might apply stricter form criteria for selecting paired behaviours than those working from the grounding perspective.

Modality

Behaviours can differ in the mode in which they are produced and perceived. For instance, they may be auditory-vocal behaviours, like spoken words, or visual-gestural behaviours, like signs and gestures (Meier et al., 2002). A prevalent assumption in the work on alignment is that for any pair of behaviour which is considered aligned, the behaviour is produced within the *same* modality. That is, the relation between the two pair parts of behaviour is considered to be a unimodal one. However, from a theoretical point of view, the two parts of aligned behaviour can also be in a *cross-modal* relation to each other, as long as they are aligned on one or more of the other five dimensions. This is less intuitive, presumably because of the (implicit) assumption that behaviour should be similar in form to at least some degree, which is difficult when produced in different modalities. However, we argue that two instances of behaviour, which are in a certain sequential, temporal, and/or meaning relation to each other, can still be considered aligned.

Though not considered in the original model of Pickering and Garrod (2004), there is evidence that iconic gestures can prime semantically related words (e.g., Yap et al., 2011), which would be a form of cross-modal alignment coming about through priming. From a grounding perspective, cross-modal alignment could be employed for communicative purposes, since lexical and gestural representations have been shown to be linked at the conceptual level (Mol et al., 2012). Both approaches thus build on the assumption that the matching of public behaviour in interaction ultimately must be related to some sort of convergence in private conceptualisations (at least for communicative speech and gesture). However, this implies that an instance of cross-modal alignment can only be identified on the assumption that we can identify a common conceptual thread to what people are communicating about—which can be challenging, especially in free conversation. To our knowledge, there is only one study which has investigated lexical and gestural alignment with such a cross-modal approach. Tabensky (2001) investigated free conversations and reports interesting cases of what could be denoted as cross-modal alignment: certain semantic information, which was initially conveyed verbally by one person, can be repeated by means of gestures by the other person, and vice versa. Take the following example (English translation, simplified transcription) from a conversation between two speakers (D and N) about buying a house:

(2) 1 D a flat is- unless it measures a hundred and eighty square meters
2 N yeah like a duplex or something

D aims to convey the size of a big apartment; he produces the lexical phrase "a hundred and eighty square meters", and simultaneously makes a gesture by opening and separating his hands sideward, while also raising his chin. Tabensky argues that this gesture conveys additional semantic information, which is not expressed in speech: that is, the gesture conveys both width and height. His conversation partner N takes up the information from the two modalities, and subsequently repeats both idea units in a new lexical phrase: "a duplex" (i.e., a spacious apartment on two levels). In Tabensky's words, she was "verbally re-encoding the sum of information she has just been offered by way of two simultaneous modes of communication" (2001, p. 221).

Work on alignment from a cross-modal perspective is scarce, and cases of cross-speaker gesture-speech alignment have been overlooked in studies restricting their analyses to alignment in either gesture or speech. However, this is not to say that alignment has not been approached from a multimodal perspective at all. It has been explored in a different way, as researchers have investigated how alignment within one modality relates to alignment within another modality. Specifically, they aim to find out whether alignment of various types of behaviour or linguistic levels are driven by the same underlying mechanisms and serve similar functions, or are in fact independent phenomena at different levels of processing. For example, the interactive alignment model "assumes interrelations between all levels" (p. 183) and proposes that "interlocutors will tend to align expressions at many different levels at the same time" (Pickering & Garrod, 2004, p. 175). Though their model is centred on speech, it could be extended to include co-speech gestures. There are two empirical studies which have investigated such interrelations for speech and gesture—Louwerse et al. (2012) and Oben and Brône (2016)—which we will discuss in turn.

In Louwerse et al (2012), many kinds of behaviour in multiple modalities (linguistic expressions, facial expressions, manual gestures, and noncommunicative postures) were found to be aligned in form and time. The authors furthermore argue that "the mechanisms underlying this widespread synchronization seem to have a unitary character, given the simultaneous modulation of the synchrony in our results" (p. 1423). In Oben and Brône's (2016) study, participants engaged in a spot-the-difference game, in which they had to refer to various objects in animated videos. Lexical and gestural alignment were operationalised as adjacent references to the objects produced by the two speakers, which overlap in root form (for words) or mode of representation (for gestures). They found no correlation between the two kinds of alignment; "target objects that are often lexically aligned are not systematically gesturally aligned as well" (p. 41). Furthermore, they found that lexical and gestural alignment can be explained by different factors: lexical alignment is predicted by the number of times one's conversational partner has used a word, whereas for gestural alignment temporal overlap in referring to an object (i.e., whether or not a gesture was produced simultaneously or with a lag) is the most important factor. Thus, in contrast to Louwerse et al. (2012), Oben and Brône (2016) conclude that lexical and gestural alignment seem to be governed by different rules.

With the exceptions of these two studies, most investigations into lexical alignment have adopted a strictly unimodal perspective, where nonverbal aspects of interactions were not taken into account (note that commonly the task setting was such that participants could not see each other; e.g., Brennan & Clark, 1996; Garrod & Anderson, 1987). On the other hand, studies of gestural alignment generally do elaborate on the relation between gestures and the accompanying speech, yet lack a systematic investigation of lexical alignment. For example, Holler and Wilkin (2011a) descriptively distinguish between various ways in which gestural alignment relates to speech. They note that gestural alignment is often accompanied by lexical alignment (e.g., consistently referring to a figure as "the ice skater", along with a physical re-enactment), resulting in so-called conceptual pacts. Yet such coinciding lexical alignment does not always occur, as gestural alignment can also be sufficient on its own to effectively refer to an object or to express acceptance of that reference. Holler and Wilkin report cases of strong gestural convergence which “carry most of the communicational burden”, thereby eliminating the need for lexical alignment, and allowing for less precision and more cross-speaker variation in verbal referring expressions. For example, members of a dyad could interchangeably refer to a figure as either having “arms” or “things” sticking out, yet be consistent in the use of the accompanying gesture (two arms representing the position of the figure’s arms). These observations are in line with the findings of Tabensky (2001) and Chui (2014), who found that interlocutors can repeat a certain gesture while producing a verbal description which diverges from their speech partner’s, thus putting the gesture into a new relationship to speech.

In terms of theory-based hypotheses, priming accounts expect alignment to be linked across (linguistic/conceptual) levels, which might generalise to links across multiple modalities. Hence, similar to how lexical and semantic alignment seem to “boost” syntactic alignment (Branigan et al., 2000; Cleland & Pickering, 2003; Mahowald et al., 2016), lexical and gestural alignment could also be predicted to go hand in hand. However, according to grounding accounts, the relation between modalities might be flexibly adapted to the specific communicative needs at hand—for example by aligning in the manual modality, while purposefully misaligning lexically, or vice versa. So, from a grounding perspective, cross-modal alignment may, but need not, occur, and the division of labour between gesture and speech may be manipulated for communicative or coordinative effect.

Review summary

By unpacking notions of alignment into five distinct dimensions, each of them independently motivated and grounded in empirical work, we have characterised the space of possibilities in which operationalisations of alignment can be situated and compared**.** We distinguished two prominent theoretical perspectives (*priming* and *grounding*), and showed how their assumptions regarding the underlying mechanisms of alignment pattern together with methodological choices and empirical foci. A summary of the two perspectives is presented in Table 2.2.

Broadly speaking, studies that are premised on the notion that communication is (at least partly) driven by automatic, lower-level processes (the priming approach) tend to consist of quantitative analyses to compare instances of behaviour irrespective of their sequential relation, prioritise form resemblance (rather than meaning overlap), and restrict analyses to one modality. In contrast, the line of work in which communication is regarded as an interactive, collaborative undertaking (the grounding approach) is more likely to involve qualitative analyses, with a focus on semantic information conveyed by the potentially aligned behaviour, paired with a consideration of the (multimodal) discourse context and its sequential structure.

Table 2.2. Schematic summary of relations between empirical and theoretical approaches

	Priming	Grounding
Underlying mechanism	Automatic Non-intentional Low-level	Controlled Intentional Higher level
Data collection	Controlled experiments Task-based interactions	Naturalistic interactions Task-based interactions
Modes of analysis	Quantitative	Qualitative
Dimensions prioritised	Time Form	Sequence Form Meaning

It bears repeating that priming and grounding merely represent two points of attraction in a larger space of possibilities. We tabulate them here to bring to light what is perhaps a growing tendency in current strands of work to align with one or the other and favour distinct sets of mechanisms, methods, and analyses.[12] However, as our framework shows, it is possible (and indeed perhaps desirable) to carry out fundamental work on behavioural alignment while taking inspiration from across these perspectives. The five omnirelevant dimensions of alignment that make up our integrative framework are designed to facilitate such research.

As our summary shows, the five dimensions differ in terms of their relative importance. The form dimension seems to be most prominent in the literature, understandably since this is the most directly observable (though operationalisations vary). The dimensions of time and meaning are also deemed important; priming accounts predict that priming effects decrease over time, and work from a grounding perspective tends to consider behaviours to be aligned only when they also involve shared meaning. However, our review of the literature shows that there is as yet limited theoretical and empirical work with respect to the dimensions of sequence and modality—yielding promising avenues for future research.

Discussion

There is an ever-expanding line of research on alignment in interaction, with a broad range of theoretical and empirical approaches. We demonstrated that seemingly related studies have very different approaches to the phenomenon, which are hard to reconcile because

[12] Indeed, an anonymous reviewer brings up the possibility that the distinction has been "amplified by duelling labs".

they refer to qualitatively different types of alignment. In an effort to enable cumulative progress and principled comparison, we unpacked the complex notion of alignment into five constituent dimensions. We distinguished between *priming* and *grounding* as the two most prominent theoretical perspectives, and showed that priming approaches prioritise the dimensions form and time, while grounding approaches mostly focus on sequence, form, and meaning. In this section, we identify a number of open questions in the field, and make suggestions for how the framework can benefit future work.

One opportunity for further research is the relation between forms of alignment at various types and levels of linguistic and communicative behaviour. More work is needed to ascertain whether the current postulated underlying mechanisms (priming vs. grounding) generalise to alignment of any behaviour, or perhaps only apply to a specific subset. For example, repeating another's words to resolve a misunderstanding may seem to point in the direction of grounding, whereas alignment in terms of posture might be better explained through priming. Other kinds of behavioural alignment might fall somewhere in between, with strategic as well as more automatic components being at play simultaneously (cf. Kopp & Bergmann, 2013).

Second, more work is needed on the causal relations between alignment at various channels or (linguistic) levels of behaviour. From a priming perspective, it has been argued that alignment at one level can "percolate" to other levels (Pickering & Garrod, 2004). There is certainly strong evidence for this with respect to syntactic, lexical, and semantic alignment (see Mahowald et al., 2016 for an overview), though we are not aware of published evidence for the "link-between-levels" claim for lower linguistic levels (e.g., phonetics), or across modalities (e.g., lexical choice and co-speech gestures; Oben & Brône, 2016). From the grounding perspective, one might argue that different kinds of behavioural alignment yield different communicative *affordances* (depending on the task at hand), which could have implications for the order in which they occur. For example, when referring to novel objects or concepts, the use and alignment of iconic co-speech gestures can (by virtue of their form-meaning resemblance) constitute a gateway into shared conceptualisations, which might precede any alignment in terms of lexical choice. The qualitative observations from Holler and Wilkin (2011) seem to line up with this reasoning and provide inspiration for follow-up studies.

Some of the opportunities for new research we have identified here result from the challenges involved in comparing findings on various types and levels of (linguistic) behaviour. As we have shown, there is a large space of empirical possibilities for studying alignment, and design choices in this space are often guided by research traditions and theoretical presuppositions. To make such choices more visible, and to increase the commensurability of work across theoretical perspectives, we recommend that studies clearly explain how alignment has been operationalised and which dimensions have been privileged. The theory-agnostic framework proposed here can be a useful resource:

adopting a common terminology for the building blocks of alignment will greatly enhance comparability and theory building in the field.

Our overview of the alignment possibility space has also shown a dearth of theoretical and empirical work with respect to the dimensions of sequence and modality. Regarding sequence, many quantitative analyses tend to ignore the inherent sequential structure of (task-based) interactions altogether. However, this could be an interesting test bed for differentiating between diverging theories. From a grounding perspective, there are good reasons to believe that alignment rates will be higher within certain sequences. In contrast, presuming that an automatic priming mechanism underlies alignment, we could hypothesise that only temporal proximity affects alignment, irrespective of sequential relation.

When it comes to the modality dimension, various theories leave open the possibility of cross-modal alignment, although empirical evidence is still lacking. Cross-modal alignment is presumably not considered to be alignment (nor “repetition”, “mimicry”, or “behaviour matching”), because there is a lack of form-resemblance, which is a key characteristic in both grounding and priming accounts. However, when listeners align to the speaker’s verbal narration in a nonverbal manner, such as wincing or showing a concerned facial expression when someone tells a close-call story (Bavelas et al., 2000), this could be considered a form of meaning alignment. Yet cross-speaker speech-gesture relationships remain understudied (Tabensky, 2001), which is remarkable, given that speech and gesture are semantically co-expressive (McNeill, 1992), and tightly linked in both production and comprehension (Cassell et al., 1998; de Ruiter et al., 2012; Kita & Özyürek, 2003; Kopp & Bergmann, 2013; Mol et al., 2012; for a review, see Özyürek, 2018). Thus, little is known about whether, and if so how, lexical and gestural alignment are interrelated, making it a promising avenue for further research.

In closing, we outline three specific recommendations for work on cross-participant alignment of communicative behaviour:

I. *Theorise alignment phenomena using common conceptual foundations.* Use the dimensions of time, sequence, meaning, form, and modality to delineate alignment, and to formulate theories and testable predictions about its cognitive mechanisms and communicative functions.

II. *Describe operationalisations to enable targeted comparisons.* Explicitly describe *which* instances or streams of behaviour are compared and *how* those are compared. That is, describe how behavioural similarities are measured and how observations are selected, grouped, manipulated, or measured in terms of the five dimensions of time, sequence, meaning, form, and modality.

III. *Combine methods to build a more comprehensive view of alignment.* Combine observational and experimental methods, and qualitative and quantitative approaches, to further unravel the multidimensional nature of alignment—especially in terms of sequence and modality, which remain largely unexplored.

Following these recommendations will contribute to increased interdisciplinary coherence, will enhance the reproducibility and generalisability of results, and will enable more principled comparisons across the fields that study the alignment of communicative behaviour.

Conclusion

A paper with the goal of charting different takes on alignment and related phenomena in human interaction cannot escape the ironic observation that there appears to be, on the surface, a relative lack of alignment on basic terminology in related fields that would benefit from working together. However, as we have argued, even different lexical labels may mask deeper underlying similarities. Here we have sought to bring out the most important of these in terms of five constituent dimensions relevant to any notion of cross-participant alignment in interaction: time, sequence, meaning, form, and modality.

By decomposing the multidimensional nature of alignment in this way, we have brought into view a wealth of theoretical interpretations and empirical operationalisations of alignment. We hold that no account of alignment in interaction can be complete without explicating the phenomenon in terms of these five dimensions, which crosscut levels of analysis and assumed mechanisms. In time, the rise of explicit operationalisations of alignment and kindred notions in terms of these basic dimensions will result in greater commensurability and comparability of empirical and theoretical work. We hope the framework will be of use as a conceptual tool to disclose hidden assumptions, refine theoretical accounts, and so enable cumulative progress in the study of alignment in interaction.

Acknowledgments

This paper came about in the context of a multidisciplinary research project within the Language in Interaction consortium, called Communicative Alignment in Brain and Behaviour (CABB). We wish to make explicit that the work has been shaped by contributions of CABB team members: Mark Blokpoel, Sara Bögels, Lotte Eijk, Judith Holler, Branka Milivojevic, Wim Pouw, Iris van Rooij, Ivan Toni, and Marieke Woensdregt. We are also grateful to the following people for their comments on an earlier version of this paper: Marlijn ter Bekke, Sara Bögels, Lotte Eijk, Judith Holler, Dilay Karadöller, Ezgi Mamus, Francie Manhardt, Branka Milivojevic, Iris van Rooij, Jan Peter de Ruiter and Ivan Toni. The authors remain solely responsible for contents of the paper. This work was supported by the Netherlands Organisation for Scientific Research (NWO) Gravitation Grant 024.001.006 to the Language in Interaction Consortium and Vidi grant 016.vidi.185.205 to MD.

Chapter 3

The primacy of multimodal alignment in converging on shared symbols for novel referents

This chapter is based on:
Rasenberg, M., Özyürek, A., Bögels, S.,
& Dingemanse, M. (2022).
The primacy of multimodal alignment in
converging on shared symbols for novel referents.
Discourse Processes, *59*(3), 209–236.
https://doi.org/10.1080/0163853X.2021.1992235

Data and code are available at:
https://doi.org/10.34973/7kbd-5g86

Abstract

When people interact to establish shared symbols for novel objects or concepts, they often rely on multiple communicative modalities as well as on alignment (i.e., cross-participant repetition of communicative behaviour). Yet these interactional resources have rarely been studied together, so little is known about if and how people combine multiple modalities in alignment to achieve joint reference. To investigate this, we systematically track the emergence of lexical and gestural alignment in a referential communication task with novel objects. Quantitative analyses reveal that people frequently use a combination of lexical and gestural alignment, and that such multimodal alignment tends to emerge earlier compared to unimodal alignment. Qualitative analyses of the interactional contexts in which alignment emerges reveal how people flexibly deploy lexical and gestural alignment in line with modality affordances and communicative needs.

Introduction

Even when sharing a common language, we sometimes talk about things for which we do not have conventional labels, such as abstract ideas, new innovations or unfamiliar objects. How do people create shared symbols to refer to these novel referents? Here we study this question in the context of multimodal interaction, the natural ecology of human language. Our aim is to understand when and how people converge on referential expressions and how they use spoken and gestural resources in this process. We focus on the interplay between two key interactional processes that are known to underlie the emergence of novel symbols: alignment (i.e., cross-participant repetition of communicative behaviour) and the flexible deployment of communicative affordances of the vocal (e.g., speech) and manual (e.g., gesture) modalities.

The importance of alignment for collaborative referring to (novel) objects or concepts has been substantiated in work on alignment. People have been shown to perform better in joint cooperative tasks (such as the Map Task; Brown et al., 1984) when they align their communicative behaviours, such as lexical and syntactic choice (Dideriksen et al., 2020; Fusaroli & Tylén, 2016; Reitter & Moore, 2014). There is also evidence for a causal effect of alignment on the process of creating shared symbols: in a study involving drawings, communicative success (that is, how accurately matchers were able to identify the correct meaning based on a drawing) was higher when participants were allowed to make their drawings alike, compared to when they were forbidden to do so (Fay et al., 2018, Experiment 2).

The different affordances of the vocal and manual modalities for symbol creation have been a key topic in the field of language evolution or emergence (e.g., Goldin-Meadow, 2017; Levinson & Holler, 2014). When people cannot rely on conventionalised symbols to refer to (novel) objects or concepts, gestures are effective because of their iconic potential (Fay et al., 2013, 2014; Macuch Silva et al., 2020; Zlatev et al., 2017), and may therefore help "bootstrap" a communication system (Fay et al., 2013). However, when people are faced with unfamiliar stimuli, there is also evidence for a multimodal advantage of combining gestures and non-linguistic vocalisations (i.e., non-word sounds) compared to using either of those modalities alone (Macuch Silva et al., 2020), which implies that their joint contribution might facilitate shared symbol creation.

So, previous work has revealed that behavioural alignment plays a key role in collaborative referring, and that the manual modality (in combination with the vocal modality) can be used effectively for establishing joint reference to (novel) objects or concepts. Yet we know very little about how people use the communicative affordances of multiple modalities in the process of alignment in emergence contexts. This is because alignment has mostly been studied in terms of just lexical choice or co-speech gesture, without looking at the relation between modalities, and because studies of language emergence have rarely focused on the analysis of cross-modal alignment in interactive contexts. There is a missing link in our understanding of the interplay between alignment and the affordances of communicative modalities: how do people deploy alignment in one or multiple modalities when referring to novel referents?

Here we aim to provide a first step towards answering this question by looking at lexical and gestural alignment in a multimodal corpus of dyads performing a referential communication task with novel objects (similar to the Tangram task (Clark & Wilkes-Gibbs, 1986) but in a face-to-face setting, see Figures 3.1 and 3.2). Our primary focus is on the *emergence* of alignment: we examine the first time speakers repeat each other's lexical choice and/or gesture (i.e., align) when referring to a particular referent in a conversational context. We quantify *how often* and *when* this happens and in which modality or modalities (i.e., lexically, gesturally or in both modalities). If alignment is established in both modalities for a particular referent (i.e., multimodal alignment), we ask next whether it emerged simultaneously (lexical and gestural alignment emerge at the same time) or successively (alignment in one modality preceding alignment in the other modality). To investigate *how* alignment is employed for collaborative referring, we qualitatively inspect its turn-by-turn unfolding and the affordances of the spoken and gestural modalities as they are recruited by participants.

Modality and alignment

A key element of the process of achieving collaborative reference is for participants to establish a shared conceptualisation: a *conceptual pact*. Such conceptual pacts can be encoded in particular verbal expressions (Brennan & Clark, 1996), but also gestures (Holler & Wilkin, 2011a) or drawings (Fay et al., 2018). For example, communicators can align on lexical items such as "ice skater" to refer to a Tangram figure (Clark & Wilkes-Gibbs, 1986), "line" to refer to a particular part of a maze (Garrod & Anderson, 1987) or "loafer" to refer to one out of multiple shoes (Brennan & Clark, 1996). When used repeatedly over time, such conceptual pacts are considered to have become "entrained" (Brennan & Clark, 1996).

Conceptual pacts do not appear out of the blue; they take interactional work: "speakers and addressees work together in the making of a definite reference" (Clark & Wilkes-Gibbs, 1986, p. 1). During this collaborative process (known as *grounding*), repetition of lexical choice can be particularly useful; it can be employed to accept a referring expression (Clark & Brennan, 1991), or to repair or expand it (Clark & Brennan, 1991; Clark & Wilkes-Gibbs, 1986; Dingemanse, Roberts, et al., 2015; Fusaroli et al., 2017). Co-speech gestures can be effective for this process as well (Chui, 2014; Holler & Wilkin, 2011a; Tabensky, 2001). For example, in one study by Holler and Wilkin (2011a), a participant playing the role of director described a Tangram figure with the verbal expression "with two things sticking out" along with a co-speech gesture where her two arms represent the position of the figure's arms sticking out from the back. The matcher replied with "yeah" while repeating the gesture, which signalled in a "definite manner that the entirety of the reference has been understood" (p. 143).

Holler and Wilkin's (2011a) results show how lexical and gestural alignment can be recruited jointly or separately in various ways when forming conceptual pacts. They can go together, for example when speakers both produce the lexical phrase "the ice skater", as well as the same gestural representation of the figure. Or they can part ways, as when a director said "an ostrich" together with a re-enactment of the figure and the matcher replied with "Yeah, okay that, that looks like a woman to me, kicking her leg up behind her, yeah?", while producing the same gesture. Here the matcher repeated the iconic gesture while replacing the verbal expression with an alternative conceptualisation. Other work, too, shows examples where speakers copy each other's gestures in casual conversation, either with or without lexical alignment (Chui, 2014; de Fornel, 1992; Graziano et al., 2011; Kimbara, 2006).

If we were to make predictions about how frequently lexical and gestural alignment co-occur, we could expect prevalence of multimodal alignment based on the interactive alignment model by Pickering and Garrod (2004). Here alignment is considered to be the result of linguistic representations being automatically primed during comprehension, which "percolates" across levels, such that alignment at one linguistic level leads to alignment at other levels as well. This claim has been supported by evidence showing that lexical and semantic alignment "boost" syntactic alignment (Branigan et al., 2000; Cleland & Pickering, 2003; Mahowald et al., 2016). However, so far there is little evidence that this would generalise to alignment across modalities. While one study reported that various verbal and non-verbal channels show reliable covariation (Louwerse et al., 2012), more fine-grained studies of lexical and gestural alignment found no correlation between alignment in the two modalities (Oben & Brône, 2016) and revealed that when gestures do not match a discourse context, they are unlikely to be copied, yielding alignment on the lexical level only (Mol et al., 2012).

In sum, prior qualitative work has demonstrated *how* alignment is employed as a resource for collaborative referring, with quantitative studies providing mixed evidence for *how frequently* lexical and gestural alignment (co-)occur. The two modalities can be recruited flexibly—yielding unimodal or multimodal alignment—which appears to be governed by the interactional needs at hand.

Modality and symbol creation

Talking about novel objects or concepts without conventionalised names brings along specific interactional challenges. If modality and alignment are indeed deployed flexibly to suit communicative demands (as previous work suggests; Chui, 2014; Holler & Wilkin, 2011a; Mol et al., 2012; Oben & Brône, 2016), then the pressures of emergence contexts might invoke a preference for alignment in one particular modality, or they could call for the combination of alignment in both modalities. Though alignment has been studied in interactive tasks involving unfamiliar configurations (e.g., Fusaroli et al., 2012; e.g., Garrod & Anderson, 1987) or novel objects (e.g., Holler & Wilkin, 2011a), we are not aware of any studies quantitatively investigating both lexical *and* gestural alignment in such settings. So, to derive hypotheses on the extent to which lexical and gestural alignment might be jointly or separately recruited when referring to novel referents, we turn to studies on language emergence and language development. Though not specifically targeting the phenomenon of alignment and its role in the process of shared symbol creation, this work is useful for its focus on contexts where conventional symbols are not yet established or acquired.

What the field of language emergence and language development have in common is the wealth of evidence for the importance of the manual modality. Children use gestures to refer to objects before they learn to produce words for those objects (Iverson & Goldin-Meadow, 2005) and have been shown to convey abstract concepts through gesture when they cannot yet do so in speech (Perry et al., 1988). Adults, too, employ the gestural modality as an effective means of communication when verbal labels are missing. In referential tasks, people have been shown to communicate more effectively when they use only gestures compared to only non-linguistic vocalisations (Fay et al., 2013, 2014; Macuch Silva et al., 2020; Zlatev et al., 2017), and more efficiently when they used multimodal symbols compared to either gestures or vocalisations alone (Macuch Silva et al., 2020; but see Fay et al., 2014 where there was no advantage for multimodal over gesture-only communication). Gestural and multimodal symbols probably offer such communicative benefits because of their versatility in establishing transparent form-meaning mappings: gestures can be used to visually depict object attributes, spatial relationships, actions, and motions. Through its iconicity and indexicality, gesture lends itself well for the production of motivated signs (i.e., signs that are linked to meaning by structural resemblance or by natural association; Fay et al., 2013; see also Perniss & Vigliocco, 2014).

The iconic and indexical potential of gesture is one reason that gesture (alongside speech) is thought to play an important role in the initial stages of language evolution (e.g., Levinson & Holler, 2014; Sterelny, 2012; though there are "speech-first" accounts of language evolution too; Cheney & Seyfarth, 2005; MacNeilage, 2008; Mithen, 2005). Fay et al. (2013) argue that gesture is an effective means to bootstrap a communication system: "grounding a basic set of shared meanings in this way, during the very earliest stages of language, could then pave the way for the further expansion of the lexicon" (p. 1365).

In sum, when people align their behaviour, they are likely to do so in one or multiple modalities depending on the communicative demands. The communicative demands of symbol creation settings (i.e., settings where conventionalised referring expressions are not yet established) appear to call for the use of gestural and/or multimodal symbols, though little is known about the use of alignment of those symbols during social interaction. Here we aim to take the next step: we examine the interplay of lexical and gestural alignment in referring to novel referents. Combining quantitative and qualitative analyses, we chart the emergence of alignment in relation to modality and capture the interactional dynamics of how unimodal and multimodal alignment are employed for communicative purposes.

Present study

We aim to investigate how frequently, when and how alignment of co-speech gestures and lexical choice emerge when converging on shared symbols for novel referents. To do so, we use a multimodal corpus of interactions where people negotiate referring expressions for novel objects, which allows for (i) systematic quantitative observations of lexical and gestural alignment for particular referents, and (ii) qualitative inspection of the communicative environment in which alignment naturally unfolds.

We used a referential communication task in which participants used speech and gesture freely as they took turns to describe and find images of novel 3D objects, over six consecutive rounds. We also asked participants to individually name the objects both before and after the interaction. This set-up enables us to investigate:

i) the extent to which participants managed to create shared symbols for the novel objects;
ii) *how frequently* alignment emerges in the lexical modality only, the gestural modality only or in both modalities in the interaction;
iii) *when* alignment emerges in the lexical and gestural modality in the interaction;
iv) *how* the different alignment patterns—independent, simultaneous or successive emergence of lexical and gestural alignment—are functionally deployed to effectively refer to novel referents.

We expect participants to establish referential conventions during the interaction. The pre- and post-interaction naming of the objects serves as a rough proxy for the creation of such shared symbols, and so we hypothesise that participants will use more similar names to label the objects after the interaction, compared to before the interaction (prediction 1). Given that the interaction in our task is multimodal, we expect participants to recruit both lexical and gestural alignment as interactional resources for collaborative referring. Since participants share a spoken language, we expect that participants will work towards alignment on lexical choice, as shared lexical symbols are arguably more robust and efficient compared to relatively unconventionalised co-speech gestures. So, we predict that multimodal alignment and lexical alignment will emerge more frequently than gestural alignment alone (prediction 2).

We do not have a specific hypothesis as to whether alignment in both modalities will be more or less frequent than alignment in lexical choice only. Multimodal alignment might be expected based on psycholinguistic research showing that speech and gesture are produced and comprehended in an integrated way (Kelly et al., 2010; Kita & Özyürek, 2003; McNeill, 1992), yielding benefits of multimodality for message comprehension (Hostetter, 2011); as well as based on work underlining the affordances of the gestural modality for referential communication (especially in language emergence contexts, cf. section "Modality and symbol creation"). However, this might not necessarily result in frequent use of multimodal alignment in this task, because people differ substantially in the amount of gestures they naturally produce, which has consequences for the opportunities for gestural alignment (Özer & Göksun, 2020).

Our third prediction concerns the temporal relation between lexical and gestural alignment in cases where multimodal alignment is deployed. Alignment can emerge in both modalities at the same time (simultaneous emergence), or alignment in one modality could precede alignment in the other modality (successive emergence). Based on prior work in the domains of language emergence and development, we expect that gestural alignment will either emerge together with lexical alignment or precede it, and we expect that least frequently of all, gestural alignment follows lexical alignment (prediction 3).

Quantitative analyses necessarily abstract away from important details of how alignment is interactionally achieved in the turn-by-turn context of conversational sequences. We attend to these details through qualitative, sequential analysis of the communicative environments in which lexical, gestural, and/or multimodal alignment naturally unfold. This ensures empirical grounding for the quantitative analyses and sheds light on how modality in alignment is employed to establish joint reference to novel referents.

Methods

Dataset

The current study is based on data collected within a larger research project aimed at investigating various kinds of cross-speaker alignment. For this project, participants performed a referential communication task, similar to the classic Tangram task (Clark & Wilkes-Gibbs, 1986; Holler & Wilkin, 2011a), with images of 3D objects. Before and after this interactive task, participants individually named the objects: the naming task. For the current study, we draw on a subset of this dataset, by analysing data from half of the dyads and half of the objects.

Participants

We analysed data from 20 Dutch participants (11 women and 9 men, M_{age} = 22.9 years, $Range_{\text{age}}$ = 18–32 years). Prior to the task the unacquainted participants were randomly grouped into dyads, resulting in 7 same-gender dyads (3 male dyads, 4 female dyads) and 3 mixed-gender dyads. The participants were recruited via the Radboud SONA participant pool system. Participants provided informed consent prior to starting the experiment and were paid for participation (12–16 euros, depending on total participation time). The study met the criteria of the blanket ethical approval for standard studies of the Commission for Human Research Arnhem-Nijmegen (DCCN CMO 2014/288).

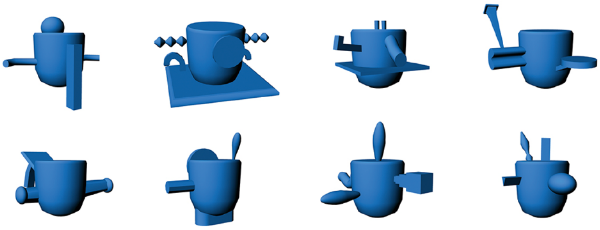


Figure 3.1. “Fribbles” that were used as stimuli; selection of 8 (out of the 16 used in the task) that were selected for the analyses.

Apparatus and materials

We used a set of 16 “Fribbles” (Figure 3.1 displays the 8 used in the analyses of the present study), illustrations of novel three-dimensional objects (based on Barry et al., 2014), designed in such a way as to ensure cross-participant and cross-dyadic variation in elicited names. During both the naming task and the interactive task, all 16 Fribbles were simultaneously presented on a grey background in a size of about 4x4 cm per figure. The Fribbles were randomly distributed over 16 positions (forming rows of 5, 6, and 5 items respectively). In the interactive, but not in the naming task, the Fribbles were labelled with letters for one participant, and numbers for the other (see Procedure). The naming task was conducted in two separate booths, where each participant was seated in front of a computer screen and used a keyboard to name the Fribbles. In the interactive task, participants were standing and faced each other (see Figure 3.2). Each had their own 24” screen (BenQ XL2430T), slightly tilted so participants could easily view the screen and their partner, and positioned at hip height to ensure mutual visibility of upper torso and gesturing area. Each participant had a button box to move to the next trial. Verbal and nonverbal behaviour was recorded using two head-mounted microphones (Samson QV) and three HD cameras (JVC GY-HM100/150).

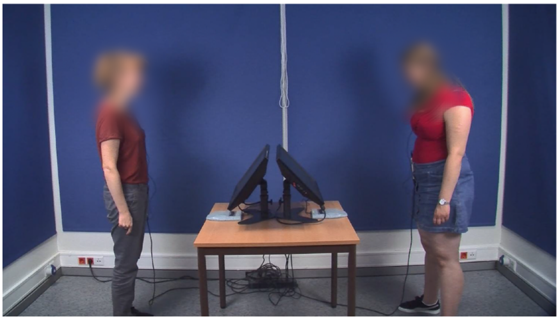


Figure 3.2. Set-up during interactive director-matcher task.

Procedure

In the naming task, participants were asked to give a name or description of 1 up to 3 words for each image (i.e., the Fribbles) in such a way that their partner (the other participant) would be able to find it amongst the other images. Target Fribbles were indicated with a red rectangle, and participants could use "ENTER" to move to the next Fribble (the order was randomised across participants). During this task, participants knew that they would take part in a communicative task afterwards, but they were not informed that this would involve the same images, nor that they would have to do the naming task again afterwards. The naming task before the interaction took 5.41 minutes on average (range = 2.24–8.01 minutes).

The referential communication task consisted of six consecutive rounds, consisting of 16 trials each, with Director and Matcher roles alternating after each trial. In each trial, a single target Fribble was highlighted for the Director by means of a red rectangle. Participants were instructed to work together in order to come to a shared understanding of what the target item is. The order in which the Fribbles were presented on the screen varied across the participants. To avoid confusion about the different orders, the Fribbles were labelled with numbers for one participant and letters for the other. Once the Matcher was confident they identified the item described by the Director, they said the corresponding positional label out loud and pressed a button to go to the next trial. Once all 16 trials had been completed, the Fribbles were shuffled and a next round started. The trial order was such that each participant took on the Director role for a certain Fribble either in rounds 1, 3, and 5 or in rounds 2, 4, and 6. No time constraints were posed and the participants did not receive feedback about accuracy. Participants were told that they were "free to communicate in any way they wanted" (an instruction phrased to be agnostic about communicative modality, i.e., speech and/or gesture), and that their performance would be a joint achievement. The communicative task lasted for 24.92 minutes on average (range = 16.38–34.56 minutes).

After the interaction, participants again individually named the Fribbles, with the same instructions as before (the only change was an additional sentence stating that the name could be the same as before, but did not have to be). This took 1.89 minutes on average (range = 0.87–3.14 minutes).

Analysis

To assess the extent to which dyads had shared symbols for the Fribbles before and after the interaction, we computed the similarity of the names they provided in the naming task. We considered names to be similar when they consisted of the same base words. All words were first spell checked, lemmatised (i.e., inflected verbs changed into infinitives, plural and diminutive forms into singular nouns) and compounds were split if they were not standard Dutch words (verified with the online Van Dale dictionary). Naming similarity was computed by taking the cosine similarity of the participants' names (i.e., vectors of words), resulting in a score ranging from 0 (no similarity) to 1 (perfect similarity; cf. Duran et al., 2019 where the same measure is used for computing lexical alignment). For example, the comparison of "right round disk" and "disk horizontal right" resulted in a similarity score of 0.67.

Since the Fribbles are new to participants, the interactive task primarily involved talking about them in terms of subparts (each Fribble has about four distinctive subparts, while the "base" figure is the same, see Figure 3.1). We took these subparts as the primary target of possible alignment in gesture and/or speech, so they make up the main unit of analysis in this study. To keep the amount of hand-coded data manageable, here we analyse half of the target items (i.e., 8 out of 16 Fribbles, with a total of 34 subparts, see Figure 3.1). We arbitrarily selected which half to use, while ensuring that the dataset remained balanced (i.e., participants start as a director viz. matcher in the first round for four items each).

Transcription and coding of multimodal interaction

Transcription of speech and annotation of gestures was done in ELAN. Speech was segmented into Turn Constructional Units (TCU; Couper-Kuhlen & Selting, 2017; Schegloff, 2007) and orthographically transcribed based on the standard spelling conventions of Dutch. For co-speech gestures, only the stroke phase was annotated (i.e., the meaningful part of the gestural movement; Kendon, 2004; McNeill, 1992), for the left and right hand separately. Gestures were categorised into three types: 1) iconic gestures, which depict physical qualities of concrete referents or movements or actions related to those referents, 2) deictic gestures, or pointing gestures, and 3) other gestures, which were mostly beat gestures and interactive gestures. Only the first category (iconic gestures) was used in the analyses below.

For the iconic gestures, we coded which Fribble subpart(s) the gesture referred to, using a pre-defined coding protocol as illustrated for Fribble 14 in Figure 3.3. Gesture referents were coded based on the kinematics of the gesture together with the co-occurring speech and overall discourse context. Gestures can refer to one subpart (e.g., a curved hand as if holding a ball to depict 14A), or to more than one subpart simultaneously (e.g., using both arms alongside the body to represent 14B+14D). Inter-rater reliability for gesture identification and gesture coding (gesture type and gesture referent) was moderate to high (for details of the inter-rater reliability analyses and results, see Appendix A).

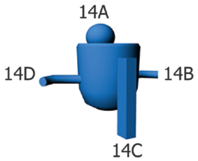


Figure 3.3. Example of Fribble subpart codes as used in coding protocols and transcripts.

Operationalisation of alignment

Any notion of communicative alignment makes relevant an operationalisation with respect to five dimensions: sequence, time, meaning, modality, and form (Rasenberg, Özyürek, et al., 2020). Our research questions primarily concern the sequential and temporal patterning of alignment, so we impose no a priori restrictions on the dimensions of sequence or time (so two instances of similar behaviour may count as aligned whether they occur within the same sequence or round, or at larger time spans across sequences and rounds). We fixate the phenomenon by focusing on the remaining three dimensions. For meaning, our criterion is referential alignment: we consider cross-speaker repetition of words or gestures to be a case of alignment only if they are used to refer to the same referent, and we exclude non-referential speech and gestures. So, if both participants use the word *egg* to refer to Fribble subpart 14A, this would count as lexical alignment, but not if one of them used it to describe another Fribble subpart. For modality, we look at alignment *within* modalities (comparing words with words and gestures with gestures), not across modalities. For form, finally, we use modality-specific criteria designed to yield a maximally commensurate measure of form similarity across modalities, as detailed in Appendix A.

To summarise our criteria, we consider lexical choice to be aligned if there is at least one common word (after lemmatizing) that both participants use to refer to the same referent, and which is informative for distinguishing referents. We consider gestural behaviour to be aligned if both participants use an iconic gesture to refer to the same referent. This is based on an explorative analysis showing that the majority of those gesture pairs overlap in one or more form features, while exact copies are rare (see Appendix B).

Quantitative and qualitative analyses of alignment

Our analyses were performed on the level of Fribble subparts (N = 340; 10 dyads * 34 subparts), where we first disregarded subparts that were never referred to (with speech or gesture) by either one or both members of a dyad, as by definition alignment would be impossible in those cases. For the remaining subparts (n = 276), we investigated whether dyads aligned lexically and/or gesturally in their referring expressions. For each case of alignment, we inspected when the “first element” (i.e., the initial word or gesture) and “second element” (i.e., the first time that word or gesture is used by the other speaker) were produced. We consider alignment to have emerged at the moment the second element is produced. Note that temporal distance between the respective elements can vary greatly (e.g., they might occur in adjacent turns within a trial, but also in different rounds of the interaction). Since we are interested in the emergence of alignment, we only coded the first occurrence of alignment for a given modality. To exemplify: once we found the emergence of lexical alignment, we did not code later re-occurrences of the aligned-upon verbal expression, nor did we check whether the dyad aligned on a different set of words later on.

In sum, for each Fribble subpart that both members of a dyad referred to, we noted in which modality/modalities alignment emerged (NO ALIGNMENT, LEXICAL ONLY, GESTURAL ONLY, or MULTIMODAL), as well as when it emerged (i.e., in which of the six rounds of the interaction the second aligning element was produced). When multimodal alignment emerged for a Fribble subpart, we grouped it into one of three categories: MULTIMODAL EMERGENCE, LEXICAL FIRST, or GESTURAL FIRST. We regarded a case as MULTIMODAL EMERGENCE when the second element of both lexical and gestural alignment was produced in the same TCU. We coded a case as LEXICAL FIRST if lexical alignment had emerged earlier than gestural alignment, that is, when the second element of the lexically aligned pair occurred in an earlier TCU than the second element of the gesturally aligned pair; and vice versa for the category GESTURAL FIRST (see Figure 3.4). The cases thus identified formed the dataset for which quantitative and qualitative analysis were conducted.

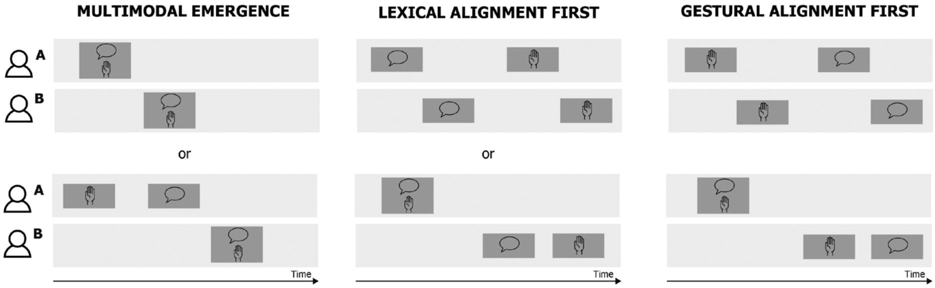


Figure 3.4. Examples of how temporal order of emergence is categorised when alignment is achieved both lexically and gesturally. Speech balloons icons are used for speech, and hand icons for co-speech gestures. Grey rectangles represent TCUs.

To analyse shared symbol creation, we used a paired samples Wilcoxon signed-rank test to assess whether naming similarity was higher after compared to before the interaction (prediction 1). To compare the frequencies of the alignment categories and orders of emergence, we used intercept-only mixed effects models with random intercepts for dyads and subparts (unless otherwise specified). These were binomial models, where specific categories were coded as 0 versus 1 to test the comparisons as specified in the hypotheses (predictions 2 and 3). Finally, we used two sample, two-sided Kolmogorov-Smirnov tests to exploratively compare categories in terms of their distributions of time of emergence (i.e., in which round of the interaction alignment emerged).

For the qualitative analysis, we used observational methods from interactional linguistics and conversation analysis (Clift, 2016; Couper-Kuhlen & Selting, 2017) to make visible the interactional work that participants accomplish with alignment. This allows us to study the sequential and formal properties of multimodal alignment as it emerges in interaction, enriching our understanding of the quantitative patterns.

Results

Shared symbols in the naming task

To find out to what extent dyads created shared symbols for the Fribbles, we compared how similar the names (consisting of 1 to 3 words) were that members of a dyad used to label a Fribble, both before the interaction (*pre*) and after the interaction (*post*); see Figure 3.5, panel A. As expected, we found that the naming similarity scores increased from *pre* (*M* = 0.07, *Median* = 0) to *post* (*M* = 0.46, *Median* = 0.41). A paired samples Wilcoxon signed-rank test indicated that this difference was statistically significant ($Z = 0.70$, $p < .001$).

By itself this leaves unclear whether the increased naming similarity is contingent on the history of the interaction, or simply a result of spending time with the stimuli and repeatedly formulating references over six rounds. To tease these options apart we compared the scores of "real dyads" with those of "non-dyads" (i.e., people who did not interact with each other), see Figure 3.5, panel B. We computed the non-dyad scores with a simple shifting function, where all names from participants B were paired with the names from participant A from the next dyad, while keeping Fribble and Session (*pre/post*) constant. In contrast to the real dyads, for the non-dyads there is no systematic improvement from *pre* (*M* = 0.11, *Median* = 0) to *post* (*M* = 0.05, *Median* = 0; *p* = .06). This allows the inference that symbol creation was indeed contingent on dyadic interaction.

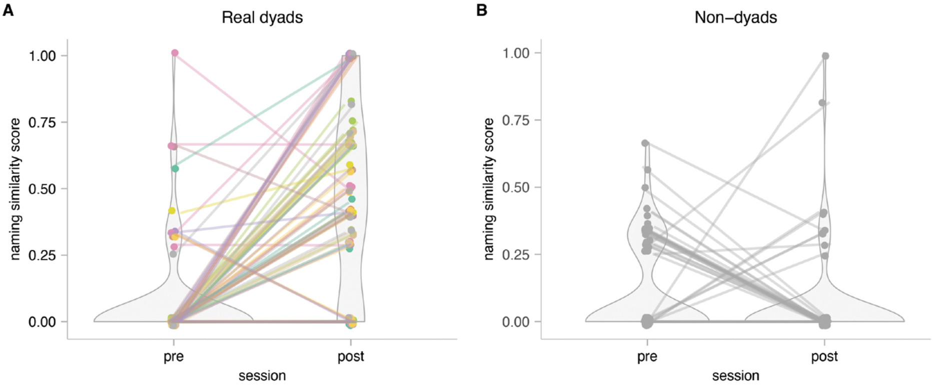


Figure 3.5. Distribution of naming similarity scores (i.e., cosine similarity of overlapping words in the names provided by participant A and B of a dyad for a particular Fribble), before (pre) and after (post) the interaction. Results from real dyads (panel A) are contrasted with those from non-dyads (i.e., pairs which did not interact with each other; panel B). Dots represent individual datapoints (*N* = 80); colours represent dyads (*N* = 10).

Remarkably, even for real dyads there are quite some name pairs with zero similarity *post* interaction (*n* = 21). Further investigation revealed that these were often cases where the two members of a dyad labelled different subparts of a Fribble. For example, participant

A's name referred to the orientation with respect to one subpart ("stands on rectangle"), while participant B's name referred to another subpart ("spoon top right"). Conversely, names with naming similarity scores of 1 (n = 12) were usually labels for one specific subpart (e.g., "chimney") or a more holistic name for the whole Fribble (e.g., "rabbit").

Though the *post* naming similarity scores might be expected to follow from the degree of alignment in the interaction, an explorative investigation yielded no evident relationship between the two (see Appendix B). This is unsurprising given the fact that the naming task elicits short written forms at the level of whole Fribbles, whereas for our measure of alignment we focused on the emergence of alignment in both speech and gesture, and at the level of the subparts, creating many opportunities for differences in selection and construal. We will get back to this in the Discussion.

Prevalence of alignment in the interactive task

Task performance was high, with Matchers selecting the correct target Fribble in 99.8% of the trials. Alignment was highly frequent in the task: on average across dyads, alignment emerged in at least one modality at some point in the interaction for 92% of Fribble subparts that had been (lexically and/or gesturally) referred to by both members of a dyad. For the subparts where alignment emerged (n = 255), 56% involved multimodal alignment, 38% lexical alignment only and 6% gestural alignment only (see Figure 3.6). As predicted, gestural alignment only occurred less frequently than multimodal alignment (β = 2.53, SE = 0.49, z = 5.53, p < .001)[13] and lexical alignment only (β = 2.08, SE = 1.05, z = 1.97, p = .048).[14]

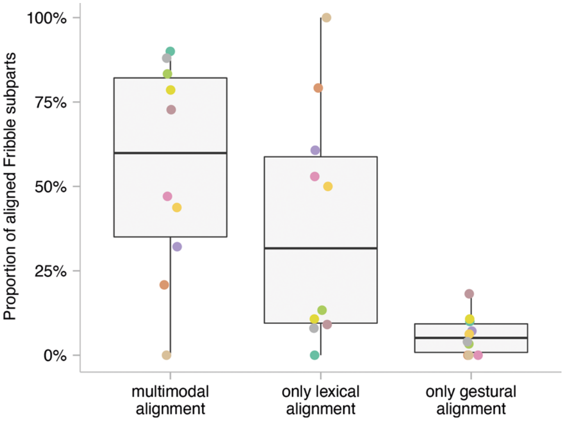


Figure 3.6. Average proportion of Fribble subparts (that have been referred to by both participants of a dyad) for which alignment occurred, by modality. Coloured dots represent dyads (N = 10).

[13] For the model comparing gestural alignment only to multimodal alignment, we only included a random intercept for subparts (not for dyads), due to convergence issues.

[14] Though we did not have a hypothesis about the difference in frequency of multimodal alignment and lexical alignment only, we compared them to provide a complete picture and found no statistical difference (β = 0.70, SE = 0.74, z = 0.94, p = .348).

Temporal distribution of unimodal and multimodal alignment

In answering *when* lexical and gestural alignment are deployed to refer to novel referents, we first compare unimodal with multimodal alignment. Alignment tended to emerge early in the interaction: for 80% of the subparts for which alignment was achieved, it emerged in the first or second round (Figure 3.7). Note that emergence in the second round is more common than in the first. This is to be expected because Director/Matcher roles switched over trials; Directors usually (lexically and/or gesturally) described the Fribbles extensively in the first round (while the contributions from Matchers varied), which was then "aligned to" in the second round by the other participant when taking up the role of Director for that Fribble.

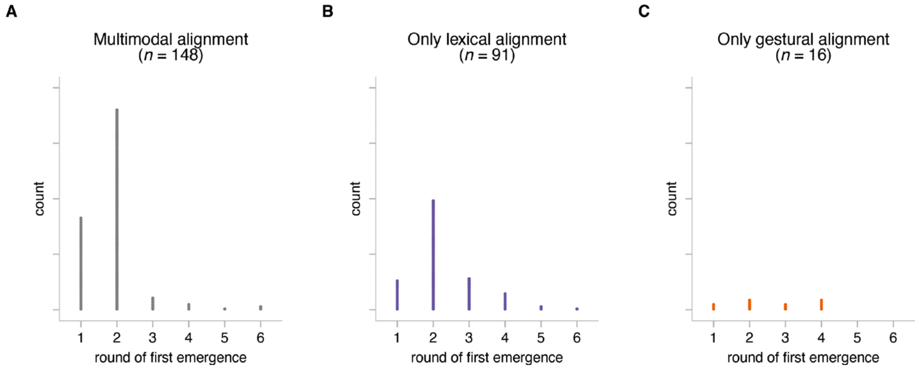


Figure 3.7. Distribution of rounds of the interaction in which alignment first emerged. For multimodal alignment this represents the time point of the first instance of alignment in either modality (see Figure 3.8 for details).

Early emergence was especially prevalent for multimodal alignment. The first instance of alignment emerged in the first or second round in 92% of the multimodally aligned subparts (Figure 3.7, panel A). Emergence in round 1 or 2 occurred less frequently for unimodal alignment, with 71% for lexical only and 50% for gestural only (Figure 3.7, panels B and C). Kolmogorov-Smirnov tests revealed that the distribution of time of emergence was different for the category multimodal alignment when compared to lexical alignment only (p = .018) and gestural alignment only (p = .013); the distributions of the latter two categories did not differ significantly (p = .560).[15]

Order of emergence in multimodal alignment

For cases of multimodal alignment, we investigated whether lexical and gestural alignment emerged simultaneously, or whether alignment in one modality preceded alignment in the

[15] Note that the category gestural alignment only is rather small (n = 16); however, when comparing multimodal alignment to unimodal alignment (thus collapsing lexical alignment only and gestural alignment only), the distributions were significantly different as well (p = .002).

other modality. For the subparts where alignment emerged in both modalities (n = 48), we found that emergence was simultaneous in 51% of cases; gestural preceded lexical alignment in 28% of cases; and lexical preceded gestural alignment in 21% of cases (Figure 3.8, panel A). As predicted, simultaneous multimodal emergence occurred more frequently than lexical alignment first (β = 0.94, SE = 0.34, z = 2.74, p = .006). But contrary to our hypothesis, we found no evidence for a difference between the frequency of lexical alignment first and gestural alignment first (β = 0.33, SE = 0.30, z = 1.10, p = .274).[16,17]

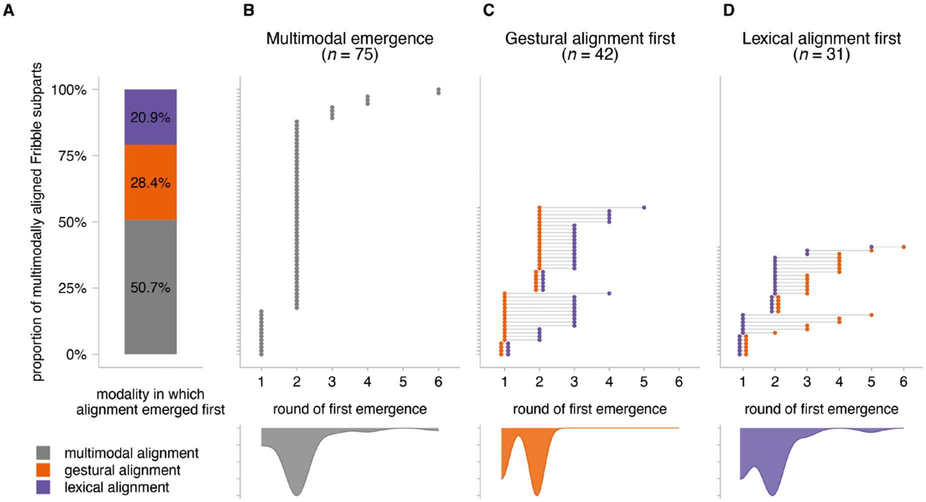


Figure 3.8. Temporal order of the emergence of lexical and gestural alignment. Panel A shows that simultaneous, multimodal emergence is most frequent, followed by gestural alignment preceding lexical alignment and lexical alignment preceding gestural alignment. The dumbbell plots in panels B-D display in which rounds of the interaction lexical and gestural components of multimodal alignment emerged (ticks on y-axis represent individual datapoints; i.e., Fribble subparts). For example, in panel C, the very top row shows that for one particular Fribble subpart, gestural alignment emerged in round 2, followed by lexical alignment in round 5. The bottom plots are density plots corresponding to the (first) dots of the dumbbell plots above.

To explore the relation between order of emergence and time of emergence, we compared the temporal distributions of the first instance of alignment for the three categories (see the density plots in Figure 3.8, panels B-D). The Kolmogorov-Smirnov tests revealed no differences between the three categories (all p > .05).

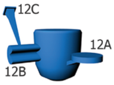
[16] For the models comparing lexical alignment first to gestural alignment first and to simultaneous emergence, we only included a random intercept for dyads (not for subparts), due to convergence issues.

[17] Though we did not have a hypothesis about the difference in frequency of simultaneous emergence and gestural alignment first, we compared them to provide a complete picture and found that simultaneous emergence was more frequent (β = 0.59, SE = 0.25, z = 2.33, p = .02).

Multimodal alignment: qualitative analyses

With the quantitative evidence in hand we are in a position to consider qualitative evidence for *how* lexical and gestural alignment are recruited as interactional resources. Multimodal emergence of alignment (i.e., simultaneous emergence of lexical and gestural alignment) most often consisted of cases where lexical and gestural alignment went "hand in hand", where a particular composite utterance (e.g., "ball" + ball gesture) was repeated as a whole by the other speaker (n = 67). Transcript 3.1 shows a representative case of how alignment emerged multimodally in the interaction. In all transcripts "A" and "B" refer to participants A (standing on the left side) and B (on the right), and the underlined speech temporally overlaps with the gesture strokes depicted in the video still with the corresponding subscript (cf. Mondada, 2018).

Transcript 3.1. Simultaneous emergence of gestural and lexical alignment when "expanding" a referential expression

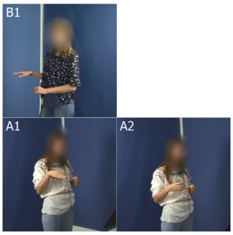


round 1 B (matcher): ja en met zo'n **plateautje** aan de rechterkant $_{B1}$?

yes and with a ***plateau*** *(like this) on the right side* $_{B1}$*?*

A: ja een soort uh cirkelachtig $_{A1}$ **plateautje** $_{A2}$ inderdaad

yes a sort of uh circular $_{A1}$ ***plateau*** $_{A2}$ *indeed*



$_{B1:}$ right-handed gesture depicting the horizontal orientation and relative position of 12A to the base shape; the flat palm-down hand makes small lateral movements

$_{A1:}$ right-handed gesture depicting 12A (similar to $_{B1}$), with a circular motion depicting the shape

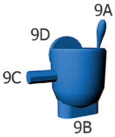
$_{A2:}$ right-handed gesture depicting 12A, with a curved handshape depicting the shape

Here, the director (A) confirms the matcher's question about subpart 12A through verbal and gestural repetition (i.e., repetition of the noun "plateau" and the accompanying gesture), with meaningful variation to provide further information. She adds the adjective "circular", which is also expressed gesturally by adding a circular motion to gesture A1 (which otherwise looks similar to the matcher's gesture B1). So the director refashions the presented referential expression through what has been called "expansion", though rather than a mere verbal process (as in the original account by Clark & Wilkes-Gibbs, 1986), here it is done in both speech and gesture.

Besides cases where lexical and gestural alignment emerge "hand in hand", there was a less frequent pattern of multimodal emergence (n = 8), where the first word and gesture were *not* produced in a single speech turn, while the repeated word and gesture were (see visualisation of this distinction in Figure 3.4). For example, one speaker introduced the lexical choice "zeppelin" in an initial TCU, which was followed by a gestural depiction with different co-expressive speech in the next TCU. Yet later on they were produced together as one composite utterance by the other speaker ("zeppelin" + gesture), yielding simultaneous multimodal alignment.

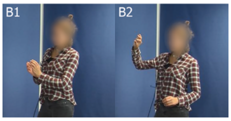
When multimodal alignment is not simultaneous, there appear to be two types of temporal patterns of successive occurrence (Figure 3.8, panels C and D). First, lexical and gestural alignment can closely succeed each other, where both emerge within the same round. We find this pattern in both directions: sometimes lexical alignment emerged first, followed by gestural alignment; and vice versa, gestural alignment first, shortly followed by lexical alignment. Such close successions of lexical/gestural alignment only occurred in the first or second round of the interaction. Alternatively, alignment could emerge at larger sequential and temporal distances (e.g., gestural alignment in round 1, followed by lexical alignment in round 3), which sometimes involves very late emergence for one modality (even as late as round 6). These two types of patterns appear to be qualitatively different from each other, and will be discussed in turn.

Transcript 3.2. Gestural alignment preceding lexical alignment in search of lexical convergence



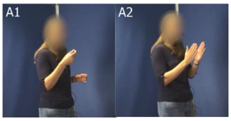
round 1 B (director): en dan boven steekt dus laat maar zeggen [B1] zo'n op- ja **op de kop** zo'n **kegel** uit [B2]

and then on top sticks so let's say [B1] a up- yes ***upside down*** *(this kind of) a* ***cone*** *out [B2]*



round 2 A (director): en rechts bovenop de ronde, op, bovenop de hoofdvorm heb je een soort van [A1] (.) ja [A2]

and right on top of the round, on, on top of the main shape you have a sort of [A1] (.) yes [A2]



B: kegel?

cone?

A: uh ja uh hoe noem je zoiets? een uh

uh yes how do you call something like that? a uh

B: kegel op de kop

cone upside down

A: ja een **kegel op de kop** inderdaad

yes a ***cone upside down*** *indeed*

B1: two-handed gesture depicting the shape of 9A; static gesture with the wrists held together and the curved palms slightly apart

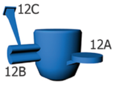
B2: right-handed gesture depicting 9A; the index finger and thumb are held slightly apart (illustrating the width of the subpart), while making a single upward (slightly diagonal) movement, depicting the orientation of the subpart.

A1: right-handed gesture depicting 9A (similar to B2); the index finger and thumb are held slightly apart, while making an up-and-down movement

A2: two-handed gesture depicting 9A (similar to B1): the hands start out put against each other, then move upward with the palms slightly apart, and end with the fingertips touching each other

Transcript 3.2 provides an example of the pattern in which the emergence of gestural alignment is followed by the emergence of lexical alignment in relatively close succession. In round 1, participant B referred to subpart 9A with the words “cone” and “upside down”, along with depictive gestures representing the same subpart. In round 2, A runs into trouble verbally describing the subpart, and produces a disfluent utterance supported by two depictive gestures that resemble both of B’s earlier gestures. The emerging gestural alignment appears to be used here in search of lexical convergence. Participant B gazes at A’s gestures and then suggests a lexical completion for A’s utterance (“cone?”), which A accepts while seeking and receiving further clarification of the fuller lexical formulation (“a cone upside down”), establishing lexical alignment. The interactional work done by the gestures appears to support a word search and is likely aided by their visible similarity.

Transcript 3.3. Lexical alignment followed by gestural alignment for calibrating a conceptual pact



round 1 A (director): uh deze heeft aan de rechterkant een platte ronde **schijf** en aan de linkerkant heb je een uitsteeksel met daar bovenop nog zo’n heel langwerpige [zo’n toeterding

uh this one has on right side a flat round ***disk*** *and on the left side you have a projection with on top of that another like very elongated [such a horn thing*

B: [ja G

[yes G

round 2 B (director): dit is die uh de beker waarvan er uh een horizontale **schijf** B1 rechts zit en dan links zit nog een uitsteeksel met zo’n hele lange ja kegel

this is that uh the cup of which uh one horizontal ***disk*** *B1 is on the right and then on the left there is another projection with such a very long uh cone*

A: [oh ja

[oh yes

B: [zo’n spijl erbovenuit

[such a bar above

A: en aan de rechterkant zo’n ronde schijf A1 toch?

and on the right side such a round disk A1 right?

B: ja gewoon die plat staat ja

yes just which is flat yes

A: ja 15

yes 15

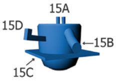
B1: left-handed gesture where the hand models 12A, with a sharp lateral movement marking the horizontal orientation

A1: left-handed gesture with a curved handshape depicting the shape of 12A

We see the reverse, with lexical alignment coming first, in Transcript 3.3. Here both participants use “disk” to refer to 12A: in the first round produced by A as the director without a gesture, and in the second round by B as the director with a gesture. However, the combination of B's noun phrase ("horizontal disk") and gesture (representing the horizontal orientation of the disk with a sharp lateral movement) is treated as inconclusive by A, who seeks to clarify the *shape* of the subpart. This is done, much like in Transcript 3.1, by presenting a modified version of both the noun phrase ("round disk" instead of "horizontal disk") and the gesture (as if moulding a disk, with a curved handshape), establishing gestural alignment in the process. Although a partial form of lexical alignment was established at the start of round 2 (where participants align on the noun ("disk"), but not on the adjective ("round" versus "horizontal"), the subsequent lexical and gestural refinements serve to further disambiguate and calibrate the emerging multimodal conceptual pact.

Transcripts 3.2 and 3.3 demonstrate how alignment in the two modalities emerge in close succession early on in the interaction, working together to establish mutual understanding. We now turn to the patterns of more distant emergence, starting with the category "gestural alignment first". Participants frequently use gestures to establish joint reference early on the interaction (with gestural alignment emerging in round 1 or 2, while the lexical references are not yet aligned, or rather underspecified, e.g. "protrusion"), which is later on followed by lexical alignment (e.g., "horn" in round 4). With respect to the category "lexical alignment first", participants at times appear to resort to gestures later on in the interaction to deal with interactional trouble, such as to further calibrate a (somewhat underspecified or partial) lexical pact (much like in Transcript 3.3) or when they appear to have trouble retrieving a lexical item, as shown in Transcript 3.4:

Transcript 3.4. Gestural alignment in an environment of lexical disfluency



round 6 A (director): uh dit is de glijbaan, de pijp en de en het en het vlak A1 dat er doorheen zit
uh this is the slide, the pipe and the and the and the plane A1 which is through it

B: yes dat is L
yes that is L

A1: right-handed gesture depicting 15C; flat hand palm-up, making a lateral movement depicting the horizontal orientation

Though lexical alignment emerged in round 5, in round 6, participant A's description of this Fribble that runs into disfluency ("and the and the and the"), foreshadowing trouble in retrieving a lexical item. He finally produces a lexical item ("plane") that is different from the one they aligned on before, but does so together with a gestural depiction of 15C, using gestures that are similar to those produced much earlier by B (in rounds 1 and 3). So, a similarity in gestural representation is used to restore collaborative reference. The use of gesture in an environment of disfluent speech is similar to what we saw in Transcript

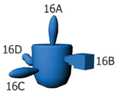
3.2, and underlines the flexible way in which language users shift the division of labour across modalities. Two non-exclusive ways to interpret the use of gesture here are that gesture helps lexical retrieval and/or that gesture is used as compensation for the “broken” lexical pact.

Unimodal alignment: qualitative analyses

While multimodal alignment was prevalent and emerged early in the interactive task, both lexical and gestural alignment separately also warrant our analytical attention, starting with lexical alignment only (the most common case after multimodal alignment).

Lexical alignment was most likely to emerge in round 2. It is useful to look more closely at the interactional work alignment is doing in such cases. We found that often a director produced a particular noun phrase in round 1, which was reused by their partner when taking on the role of director in round 2. But this reuse was rarely straightforward repetition and typically involved some modification or expansion. Consider Transcript 3.5:

Transcript 3.5. Lexical alignment for calibrating a conceptual pact

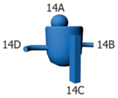


round 1	A (matcher):	((inbreath)) o:h ja maar hij heeft ook één zo’n uh vierkante neus *o:h yes but he also has one like uh square nose*
	B:	ja *yes*
	A:	en nog een soort *and also a sort of*
	B:	**slurfje**, staartje ***trunk**, tail*
	A:	ja *yes*
round 2	A (director):	met een vierkante schroef als neus en een **slurfje** aan de achterkant *with a square screw as nose and a **trunk** on the back*

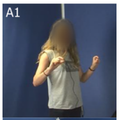
In round 1, the matcher appears to have found the target Fribble (as suggested by an inbreath and a stretched change of state token “o:h”), and subsequently describes several other subparts of the Fribble to verify her selection. After describing subpart 16B as a “square nose”, she goes on to describe 16D, but is interrupted by the director who completes her sentence with “slurfje, staartje” [(elephant’s) trunk, tail], which A confirms by saying “yes”. This double-barrelled candidate description (casting part 16D as a small trunk or tail) provides source material for a conceptual pact, but does not yet commit to a single conceptualisation; indeed, the two candidate nouns imply opposite animal parts. In the next round, participant A (now director) reuses B’s word “trunk” in her description. The immediate result of this case of lexical alignment is to commit to one particular conceptualisation, which is taken up without further problems by B. Though this example came from a dyad where both participants gestured regularly, it shows that sometimes lexical alignment can be sufficient for the task at hand.

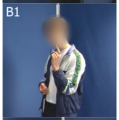
Turning to the category of "gestural alignment only", even if this is relatively rare, two salient patterns emerged in the data. The first one is where gestural alignment emerges early on for subparts that may be hard to capture in speech, as shown in Transcript 3.6. Here, A and B refer to 14D, both verbally and gesturally. Both start speaking in overlap, with A resolving the overlap by withholding completion of the spoken turn while launching into a depictive gesture. B's turn is completed in the clear with an alternative gestural depiction occupying the slot of the noun (Clark, 2016). This composite utterance is treated as sufficient by A, as seen by her spoken confirmation and another gesture produced with her left hand (which she still had in the air, i.e., in a post-stroke hold). However, she somewhat changes the gesture's handshape and motion (now showing more resemblance to B's gesture), as if to say "what you just gestured is the same as what I was gesturing about". With the spoken utterances conveying only limited information, the dyad appears to rely heavily on coordinating their gestures to achieve collaborative reference.

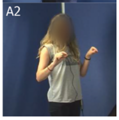
Transcript 3.6. Gestural alignment as a substitute for speech



round 1	A (director):	[en aan de linkerkant A1 *[and on the left side A1*
	B:	[en links zit nog een soort zo gebogen B1 *[and on the left there is also a sort of bended B1*
	A:	ja A2 *yes A2*
	B:	ja *yes*
	A:	ja *yes*


A1


B1


A2

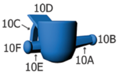
A1: left-handed gesture depicting subpart 14D; the index finger and thumb are held slightly apart (illustrating the width of the subpart), while making small sideward movements. The right-handed gesture is a post-stroke hold (depicting 14B, which is irrelevant for current purposes).

B1: left-handed gesture depicting subpart 14D; a curve is traced with the extended index-finger.

A2: left-handed gesture depicting subpart 14D; the index finger is slightly extended in a single sideward movement.

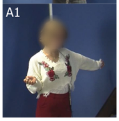
The second pattern of gestural only alignment is where speakers resort to gestures for a particular referent throughout (most) of the interaction in a way that compensates for the lack of lexical alignment on that referent. Consider Transcript 3.7: throughout the interaction, the two speakers of a dyad used different nouns ("lumps" versus "spheres") to refer to subparts 10B+10F. While A has produced accompanying gestures in rounds 1 and 3, participant B produces a similar gesture for the first time as late as round 4. The sequential environment in which this happens is telling. After B's initial verbal description in round 4, A produces a soft verbal repetition of part of the formulation ("arms with...") while visibly scanning the array of Fribbles on her screen. This display of trouble is followed by an upgraded formulation on the part of B, who now produces a multimodal utterance that is both more lexically specific ("two arms with a sphere attached") and features a two-handed gestural depiction of the spheres time-aligned with "sphere attached". So where a mere lexical formulation proved insufficient for A, the dyad resorted to the gestural modality to establish collaborative reference, and continued to rely on the gestural depiction (in the absence of lexical alignment) in rounds 5 and 6 as well.

Transcript 3.7. Gestural alignment for re-establishing collaborative reference under uncertainty



round 4 B (director): dit is die met uh de armen met bolletjes eraan
this is the one with uh the arms with spheres attached
en achter een uh ding nog
and behind a uh another thing
A: ^{0}armen met^0 ((visibly searches on screen))
^{0}*arms with*^{0}
B: twee armen met een [bol $_{B1}$ eraan en dan
two arms with a [sphere $_{B1}$ attached and then

A: [oh die zo $_{A1}$
[oh the one (like this) $_{A1}$



B: ja
yes

B1: two-handed gesture where curved handshapes depict the round shapes of subparts 10B+10F, somewhat away from the body thereby depicting the subparts' positions relative to the base shape
A1: two-handed gesture where clenched fists model subparts 10B+10F, right extended arm models 10A and left tucked-in arm models 10E.

What unites both of these patterns of gestural-only alignment is that they rely on the visuo-spatial affordances of the gestural modality to achieve joint reference by iconically depicting aspects of a referent, either because it is hard to capture in speech, or because a spoken formulation turned out hard to interpret.

Discussion

Quantitative findings

With the present study we aimed to reveal how frequently, when and how lexical and gestural alignment emerge when creating shared symbols for novel referents. First of all, our results confirm that symbol creation took place over the course of the interaction, as we found that the names that participants used to label the novel objects were more similar to each other after compared to before the interaction (prediction 1 supported). As for the interactions, we found that alignment was very frequent overall: for 92% of the novel referent subparts that dyads referred to, some form of alignment occurred at some point in the interaction, with multimodal and lexical alignment being more frequent than gestural alignment only (prediction 2 supported). We found a distinctive pattern for multimodal alignment: it was both more frequent than gestural alignment only and tended to emerge earlier in the interaction compared to both lexical and gestural alignment only. For those cases of multimodal alignment, we found mixed support for prediction 3: emergence of alignment in both modalities simultaneously was more frequent than successive emergence (i.e., lexical alignment preceding gestural alignment or vice versa), but contrary to our expectations, the two types of successive emergence (gestural alignment preceding alignment, and lexical alignment preceding gestural alignment) were equally frequent.

The prevalence of alignment in our study corroborates the notion that alignment plays an important role in collaborative referring (Brennan & Clark, 1996; Fay et al., 2014, 2018; Holler & Wilkin, 2011a; Reitter & Moore, 2014).[18] We found that lexical and gestural alignment can be deployed flexibly: they can occur in tandem as well as independently, which is in line with earlier work showing no systematic relation between the two (Oben & Brône, 2016), and qualitative reports on various combinations of lexical and gestural alignment (Chui, 2014; Holler & Wilkin, 2011a). Yet multimodal alignment was clearly favoured. This finding relates to psycholinguistic work on multimodal communication in two ways. First, given that speech and gesture are integrated during both production and comprehension (Kelly et al., 2010; Kita & Özyürek, 2003; McNeill, 1992), multimodal alignment may be the result of cross-participant repetition of the composite utterance as a whole. Second, since receivers have been shown to benefit from multimodality in message comprehension (Hostetter, 2011), participants could have relied mostly on multimodal, rather than unimodal alignment, to ensure more robust communication in this task.

The prevalence of multimodal alignment also ties in with the previously reported efficiency advantage for multimodal signals in the field of experimental semiotics (Macuch

[18] Note that in our study we were not able to relate patterns of alignment to task performance, as all dyads scored at or near ceiling in the referential task.

Silva et al., 2020), and with accounts of multimodal origins of language (Levinson & Holler, 2014; Perlman, 2017; Zlatev et al., 2017). Our study complements this prior work by showing that when people cannot rely on conventionalised referring expressions, multimodality is not only a useful property of communicative signals, it is also a resource for *aligning* to the signals of other participants. Furthermore, we found that *early* alignment tends to be multimodal rather than unimodal. This may be because most referents were hard to describe, putting pressure on people to use both multimodal utterances and alignment as resources to establish joint reference early on in the interaction (which then yields early emergence of alignment in at least one modality). Conversely, for easier referents, both the need for alignment and multimodal communication could be lower (yielding later emergence of unimodal alignment).

Turning to the order of emergence for multimodal alignment, we found that simultaneous emergence of alignment in both modalities was most frequent, again underscoring the need to consider multimodal origins of language. However, we also found ample cases where alignment emerged in one modality first and later in the other, but contrary to our expectations, the two orders were equally frequent. We hypothesised to find ample "gestural alignment first", as this would resemble patterns in contexts of language development and language emergence where gestures (paired with vocalisations) can "pave the way" for the emergence of conventionalised lexical items (Fay et al., 2013; Iverson & Goldin-Meadow, 2005; Perry et al., 1988). While the quantitative finding that "lexical alignment first" was not rare was surprising, our qualitative analyses revealed that this occurred to deal with particular communicative challenges, as we will argue later in the discussion.

Qualitative findings

Our qualitative findings demonstrate *how* (multi)modality and alignment interact in collaborative referring. The results corroborate earlier work showing that alignment can be employed to accept or further negotiate a referring expression, which can be done through lexical alignment (Clark & Brennan, 1991; Clark & Wilkes-Gibbs, 1986), but also gestural alignment (Chui, 2014; Holler & Wilkin, 2011a), or—as we showed here—by aligning in both modalities simultaneously. But our results bring to light another function as well: when various candidate expressions have been used for a referent, alignment can be used to commit to one of those conceptualisations.

A second insight from the qualitative analyses is that people employ both similarity *and* variation in gesture form for communicative purposes. We find evidence for what appears to be "strategic" alignment of communicatively "significant" form features (Bergmann & Kopp, 2012), where the sequential context governs which features (e.g., handshape, motion) are relevant at that moment. But our results also bring to light an alternative strategy: speakers can communicatively employ mis-alignment or deviation in salient form features to negotiate referring expressions (cf. Chui, 2014; Tabensky, 2001;

and see also Fusaroli, Rączaszek-Leonardi, et al., 2014; Healey et al., 2014 on this notion of complementarity in interaction). And finally, people might communicatively employ alignment of less significant form features as well, as a way to mark the common ground before adding new information. Transcript 3.1 provides an example of how these latter two strategies are combined: a participant repeated their partner's gesture with the same (non-salient) handedness, position, orientation and handshape (constituting the link to their partners gesture), but changed the movement into a salient, circular motion (to further specify the shape of the "plateau").

The analyses revealed that people employ modality-specific features when aligning. Whereas the discrete combinatorial format of speech allows for extending or modifying parts of noun phrases, the iconic and dynamic nature of gestures allows for copying or modifying form features to bring certain aspects of the referent in focus. These different affordances also enable people to balance the communicative load between the lexical and gestural modalities depending on the interactional needs at hand. Though overall the emergence of lexical alignment was more frequent, we also showed cases of how gestural alignment is used for achieving mutual understanding in the absence of a lexical pact (Transcript 3.7), or even in the absence of content words all together (Transcript 3.6). Gestural alignment also emerged when people experienced problems producing a verbal reference (where gestural alignment preceded lexical alignment; e.g., Transcript 3.2) or recalling an already established lexical pact (in which case gestural alignment follows lexical alignment; e.g., Transcript 3.4).

In summary, the spoken and gestural modalities offer their own affordances for alignment to establish joint reference, and these modalities are usually employed in combination. Our qualitative analyses help to make sense of the nuanced patterns that emerge from the quantitative findings. While the primacy of multimodal alignment emerges clearly throughout the study, the relative order of its building blocks, lexical and gestural alignment, appears to be governed by an interaction between the moment-by-moment communicative demands and the affordances offered by each modality.

Future research

Coming back to the initial question of how alignment and communicative modality are employed for establishing shared symbols, three challenges remain to be further explored: 1) how to operationalise alignment, 2) how to generalise the results and 3) how to account for variation in shared symbol creation.

In order to systematically track both lexical and gestural alignment, we formulated maximally commensurate measures of what constitutes alignment, regarding behaviour as aligned when it was produced in the same modality and for the same referent, and with modality-specific criteria for the required overlap in form. Our quantitative results should be interpreted and compared to prior work with this specific operationalisation kept in mind. Specifically, while most studies on gestural alignment emphasise overlap in gesture form (Rasenberg, Özyürek, et al., 2020), here we considered form overlap loosely. By pairing this with both a quantitative (see Appendix B) and qualitative investigation of gesture form overlap, we revealed how overlap *and* deviation in gesture form can be employed for communicative purposes. Future work could broaden the definition of alignment even further by also investigating cases where people verbally re-encode the information that their partner provided through gesture, or vice versa—that is, investigate alignment *across* modalities (de Fornel, 1992; Rasenberg, Özyürek, et al., 2020; Tabensky, 2001).

As to the issue of generalisability, our dataset appears to be representative of this kind of task-based setting, as we find the same phenomena as described in earlier work using similar tasks (e.g., emergence of conceptual pacts, shorter references over time, vast amounts of iconic gestures; e.g., Brennan & Clark, 1996; Clark & Wilkes-Gibbs, 1986; Holler & Wilkin, 2011a). While the interactions are clearly different from everyday conversations, they do fulfil all basic characteristics of face-to-face conversation (Clark, 1996; see also the discussion in Holler & Wilkin, 2011a) and show resemblances with common communicative situations, such as singling out a familiar referent from a set of similar referents (e.g., asking for a specific cup from a set of cups in a cupboard), or talking about novel objects or concepts (e.g., when working on an art project). Furthermore, since the Fribbles lack conventionalised labels, our data enabled us to shed some light on the potential interplay between alignment and modality in emergence contexts.

Lastly, we found quite some variation in the degree of shared symbol emergence, that is, the similarity of the names after the interaction. This variation could not be explained with the patterns of alignment in our data. This may be due to our focus on the *emergence* of alignment (i.e., the first occurrence), as opposed to repeated use (*entrainment*) later on in the interaction (see also Appendix B). Variation in systematicity and efficiency of novel symbols has previously been linked to the presence viz. absence of interactive feedback (Fay et al., 2018; Krauss & Weinheimer, 1966; Motamedi et al., 2019). Given that participants were allowed to interact as much as they wanted in our task, why did this not always give rise to simple, shared symbols as measured post-interaction? Future studies could explore this question further by investigating the kind of interactional work that is needed to go from the first occurrence of alignment to entrainment and simplification of shared symbols.

Conclusion

By systematically tracking lexical and gestural alignment in a referential communication task in a clearly operationalised way, we uncovered the primacy and prevalence of multimodal alignment when referring to novel objects. Moreover, by closely inspecting the interactional dynamics of independent, simultaneous, and successive emergence of lexical and gestural alignment, we found that the multimodal system can be flexibly adjusted to communicative pressures and constraints to yield referring expressions that contribute towards the ultimate goal of achieving joint reference. We believe a combination of qualitative and quantitative analyses akin to those in the present study have the potential to provide more insights into the joint contribution of different modalities (speech and gesture) in alignment of communicative behaviour when creating novel symbols.

Acknowledgments

This paper came about in the context of a multidisciplinary research project within the Language in Interaction consortium, called Communicative Alignment in Brain and Behaviour (CABB). We wish to make explicit that the work has been shaped by contributions of CABB team members (in alphabetical order): Mark Blokpoel, Lotte Eijk, Judith Holler, Branka Milivojevic, Wim Pouw, Iris van Rooij, Ivan Toni and Marieke Woensdregt. We would additionally like to thank Wim Pouw and Marieke Woensdregt for their comments on an earlier version of this paper, Wim Pouw, Susanne Brouwer and the M3 members of the Radboud University for help with statistics, and Emma Berensen, Junfei Hu, Vicky Ioannidou and Clarissa de Vries for their help with annotating the data. The authors remain solely responsible for the contents of the paper. This work was supported by the Netherlands Organisation for Scientific Research (NWO) Gravitation Grant 024.001.006 to the Language in Interaction Consortium and Vidi grant 016.vidi.185.205 to M.D.

Chapter

4

Handling trouble together: Multimodal strategies for repair initiations and solutions in conversation

Abstract

To engage in joint meaning-making people deploy a wide range of semiotic resources to achieve particular communicative goals. Here we study multimodal strategies for targeting and resolving interactional trouble in other-initiated repair sequences. In particular, we investigate the use of co-speech gestures in face-to-face task-based interactions in which people co-create new labels for novel objects. We find variation in how often and how gestures are used across the initiating and responding positions of the different repair types, and discuss how this distribution follows from modality-specific affordances and constraints. Our findings contribute towards a comprehensive understanding of other-initiated repair as an optimally organised, flexibly deployed system in multimodal interaction.

Introduction

Our lives are filled with social interactions. We talk about a movie, buy bread at a bakery, or give a colleague instructions. Such interactions often appear to flow naturally and effortlessly, but upon closer inspection they involve intricate processes of multimodal joint meaning-making. People do not just transmit meaning through singular turns, but incrementally build up understanding through coordinative efforts, which involve not only speech but also embodied resources such as gestures. Here we aim to advance our understanding of talk-in-interaction as a collaborative and multimodal undertaking, by asking when, why and how people select and deploy different semiotic resources.

One place to study this issue systematically is the domain of other-initiated repair (Schegloff, 2000; Schegloff et al., 1977). This is when an addressee halts the ongoing conversation to attend to some interactional trouble (in a "side sequence"; Jefferson, 1972). In these other-initiated repair sequences, the addressee signals a problem with perceiving or understanding a prior turn (in a *repair initiation*), inviting the producer of the trouble-source turn to repair the trouble (in a *repair solution*). As a simplified example, consider a sequence like "A: 'you mean this one?' ((pointing gesture)) | B: 'yeah ((nods))'". For any such sequence, we can ask how semiotic resources like words and gestures are recruited across the sequential positions of repair initiation and repair solution, and how the communicative affordances of these resources shapes and constrains the work they do in these positions.

In order to systematically study other-initiated repair from this multimodal angle, we start from the notion that addressees can initiate repair in different ways, which form a crosslinguistically general typology of three repair types (Dingemanse & Enfield, 2015):

i) *open request*; signals trouble but leaves open where or what the problem is, requests repetition or clarification (e.g., "Huh?")
ii) *restricted request*; localises trouble by signalling out a particular element of the trouble-source turn as problematic, requests repetition or clarification (e.g., "Who?")
iii) *restricted offer*; offers a candidate understanding, asks for confirmation or correction (e.g., "You mean X?")

In principle, this typology is independent of modality. For example, an open request can be realised with the interjection "Huh?" (Drew, 1997), with a gesture cupping the hand behind the ear (Mortensen, 2016) or with a manual sign for "wait" or "what" in sign language (Manrique, 2016; Skedsmo, 2020). Nonetheless, attempts to characterise other-initiated repair as a unified system have so far focused mainly on speech, showing systematicity in the use of linguistic resources for the different types of repair (Dingemanse & Enfield, 2015; Schegloff et al., 1977). Work on the role of embodied resources is much more fragmented, as studies have usually focussed on the role of certain behaviours for one particular type of repair (e.g., how bodily or facial signals are used to signal trouble in open requests; Oloff, 2018) or on one particular sequential position (repair initiations or repair solutions; e.g. Hoetjes, Krahmer, et al., 2015; Holler & Wilkin, 2011b). As a result of this, we know little about how speech and other modalities are combined to constitute multimodal strategies for the different types of repair initiations and solutions, nor do we have a good understanding of how people coordinate their use of semiotic resources across repair turns.

To address these issues, we start by surveying prior work on multimodal resources used in repair sequences, taking the two elements of the sequence in order (repair initiation, then repair solution). Rather than presenting an exhaustive review, the main aim of the overview is to demonstrate that different semiotic resources can be used for particular purposes, while simultaneously highlighting the gaps in our knowledge about repair as a multimodal practice. From this it follows that we would benefit from a holistic look at the use of manual co-speech gestures together with speech in the repair system as a whole. To do so, we carry out a quantitative and qualitative exploration of how co-speech gestures are used—together with speech—in initiating and responding positions and across repair types.

Repair initiations

How are the different types of repair initiations (open request, restricted request and restricted offer) realised in face-to-face interaction? To answer this question, we need to consider the interactional work that is accomplished in these different repair initiation types, and how they relate to the prior turn that they target as being in need of repair (i.e., the *trouble-source*). In open requests, people *signal* trouble with a prior turn without specifying where or what the trouble is. Such trouble-signalling can be achieved in semiotically-diverse ways; people can use spoken or signed interjections, question words and apology-based formats (Dingemanse et al., 2014), as well as embodied displays, such as leaning forward (Li, 2014), turning, tilting or "poking" the head (Andrews, 2014; Seo & Koshik, 2010), furrowing the eyebrows (Hömke, 2019) or continuing to gaze at one's interlocutor ("freeze displays"; Levinson, 2015; Manrique, 2016; Oloff, 2018; Skedsmo, 2020). In terms of hand gestures however, apart from the cupped-hand behind the ear to signal acoustic trouble (Mortensen, 2016), little is known about if and how representational co-speech gestures (i.e., iconic or pointing gestures which convey semantic meaning) are used together with speech to signal trouble in open requests.

In restricted requests and restricted offers, people localise trouble; they single out a particular aspect of a trouble-source turn as problematic. The localisation of trouble mostly has been investigated in the lexicosyntactic domain (e.g., question words such as "Who?"; Dingemanse et al., 2014), but prior work has also uncovered a general strategy for localising trouble: repetition. People have been found to repeat words or signs from a trouble-source turn to mark it as being in need of repair (e.g., Dingemanse et al., 2014; Dively, 1998; Skedsmo, 2020). Repetition of gestures might be used for this purpose as well, as it has been found that gestural repetition is used to negotiate meaning more generally (Chui, 2014; Graziano et al., 2011; Holler & Wilkin, 2011a; Kimbara, 2006; Rasenberg et al., 2022). To our knowledge, so far only one example of a restricted request with gestural repetition to target trouble has been described in the literature. In that example, A requested B to put something on the table by tapping on the table, which B had trouble interpreting. In this sequence consisting of multiple repair initiations, B initiates repair the second time through a restricted request, which consisted of a verbal question word and a repetition of the tapping gesture, as to say "what {do you mean by tapping}?" (Baranova, 2015, p. 568). However, we are in need of more evidence on how often and in which different ways gestures are used for localising trouble in restricted requests and restricted offers.

Finally, restricted offers involve an additional component; they offer a candidate understanding as a solution to the problem, for which both the spoken and gestural modality offer useful resources. A straightforward way in which referent ambiguities can be resolved is to suggest a candidate referent by pointing to an object, location or person (with an extended index finger or otherwise; e.g., Floyd, 2020; Kraut et al., 2003; Levinson, 2015; Dingemanse, 2015). Iconic co-speech gestures can also be used to convey a wide range of meanings through form-meaning resemblance. In one reported case, an iconic gesture was used together with speech to offer a multimodal candidate understanding of a trouble-source: the pronoun "it" (Sikveland & Ogden, 2012). In another case, someone responded to a multimodal trouble-source turn by repeating the trouble-source gesture but rephrasing the speech, thus presenting a novel multimodal understanding (Bertrand et al., 2013; see also Tabensky, 2001). The relationship between speech and gesture is highly flexible and potentially allows for more ways to express candidate understandings. This highlights the need to consider how composite utterances (i.e., utterances in which different semiotic resources are combined) are used in restricted offers. Finally, gestures can also be used to convey a candidate understanding in the absence of accompanying speech. In such cases, the meaning can follow in part from its relation to the material environment, for example when making a circling gesture close to a pie to negotiate the understanding of how a topping needs to be poured on top of it (Jokipohja & Lilja, 2022).

Repair solutions

How are repair solutions realised in co-present, multimodal interactions? A first thing to note is that repair solutions are contingent on the type of repair initiation. Usually, they involve *repetition* or *clarification* of the trouble-source turn in response to open and restricted requests, and *confirmation* or *correction* in response to restricted offers. While there is some research on vocal strategies for repeating and clarifying (Curl, 2005; Schegloff, 2004), there is very little known about gestural and multimodal strategies in the context of other-initiated repair as a cooperative system in which people work together to resolve trouble.

However, we can derive insights from studies which have investigated how peoples' gesture use changes from turns before compared to after "addressee feedback" (from a confederate in a task-based interaction; Hoetjes, Krahmer, et al., 2015; Holler & Wilkin, 2011a) and "trouble spots" (in a mathematics lesson; Alibali et al., 2013). This work has found that people add gestures (Alibali et al., 2013), or use more salient gestures; i.e., gestures that are more precise and larger compared to the gestures in their prior turn, and to which attention was drawn through deixis in speech and gaze (Hoetjes, Krahmer, et al., 2015; Holler & Wilkin, 2011b). Most of these sequences resemble or qualify as other-initiated repair, so the findings suggest that gestures are likely to be frequent and prominent in repair solutions. However, it is not clear how the addition or modulation of gestures is used together with speech for *repeating* or *clarifying* a trouble-source, or for *confirming* or *correcting* an offer, and how the design of the solution relates to the repair initiation. Let's unpack this further.

First, the prominence of gestures in repair solutions can be apprehended by looking closely at the work they do to repair particular problems. If the trouble is considered to be a problem with hearing, people can repeat speech and/or gestures (without modification) from the trouble-source turn in the repair solution (Kendon, 2004, pp. 128–134). If the trouble is considered to be one of understanding, people have been shown to clarify the trouble by (partially) repeating the trouble-source speech, while adding gestural or embodied components (Olsher, 2008) or by changing the gesture (Kendon, 2004, pp. 128–134). By systematically analysing responses to a variety of spontaneously produced repair initiations, we can see if these examples generalise and if there are perhaps more multimodal strategies for the differential tasks of repair solutions.

The use of gestures in repair solutions in response to restricted offers seems to stand out, as it was found that participants' affirmative responses to restricted offers have lower gesture rates compared to the trouble-source turn which is repaired (Holler & Wilkin, 2011b). This makes sense since confirmatory responses are often short, for example merely consisting of a verbal response token (e.g., "yes") and/or nodding. But manual gestures can still play a role in confirming (Jokipohja & Lilja, 2022). For example, a case is reported where a restricted offer involved a pointing gesture (to someone's house to resolve a referent ambiguity), which was responded to by repeating the person's name and an arm movement in the same direction, thus providing a multimodal confirmation (Dingemanse, 2015). Thus, while quantitative work shows that people use fewer gestures in repair solutions that confirm offers (Holler & Wilkin, 2011b), gestures can still play an important role in such turns. However, so far, we have a limited understanding of the types of gestures and multimodal strategies that are likely to be used in repair solutions, and importantly, how these relate to (multimodal) repair initiations.

Present study

The aim of the present study is to advance our understanding of how people collaboratively and multimodally resolve interactional trouble throughout other-initiated repair sequences. We focus on manual co-speech gestures, because prior work suggests that gestures—in contrast to other embodied resources in repair, such as facial and bodily signals—offer affordances for quite a broad range of functions in repair sequences. However, this conjecture is based on scattered findings and singular examples in the literature, which we will subject to systematic inquiry in this study, thereby working towards a more holistic understanding of repair as a multimodal system.

To this end, we first quantify the use of co-speech gestures across different types of repair (open requests, restricted requests and restricted offers), at two sequential positions (repair initiations and repair solutions). As we have seen, the interactional work carried out by people differs across these repair turns; for example, people *signal* trouble in open requests, and are likely to *confirm* (or correct) a candidate understanding in response to restricted offers. The distribution of gestures across repair initiations and repair solutions will provide us with a general impression of *when* gestures are most (un)likely to be used, and how this relates to the different repair types. Next, we qualitatively investigate *why* and *how* co-speech gestures are used together with speech in these repair turns, thereby outlining the range of the multimodal strategies that people use for localising trouble, offering solutions, clarifying trouble and confirming offers. We pay special attention to how people design their composite utterances in relation to their partners turn.

We carry out these quantitative and qualitative analyses on a dataset of task-based interactions., where participants take turns as “director” and “matcher” to describe and find novel 3D shapes (see Figure 4.1). Participants are standing face-to-face and are instructed to “communicate in any way they want”. As such, the task-based setting yields interactions in which participants recruit multimodal utterances in relatively free-form in order to negotiate mutual understanding to jointly solve the task. Referring to novel referents which lack conventionalised labels is communicatively challenging, which likely incites participants to resort to other-initiated repair to establish mutual understanding. Furthermore, the concrete, visual shapes lend themselves well for iconic depiction (Masson-Carro et al., 2016), which enables us to quantitatively study the distribution of gestures in the repair system, and investigate the role of iconicity in the gestural modality.

Methods

Data

The data for this study comes from a larger research project, in which participants performed a referential communication task with images of novel objects in a face-to-face setting (see Figure 4.1). Before and after this interactive task, participants individually labelled the objects. For the present study, we only draw on the audio-visual recordings of the communication task, as described below. The full procedure, materials and apparatus for the study are detailed in Rasenberg et al. (2022).

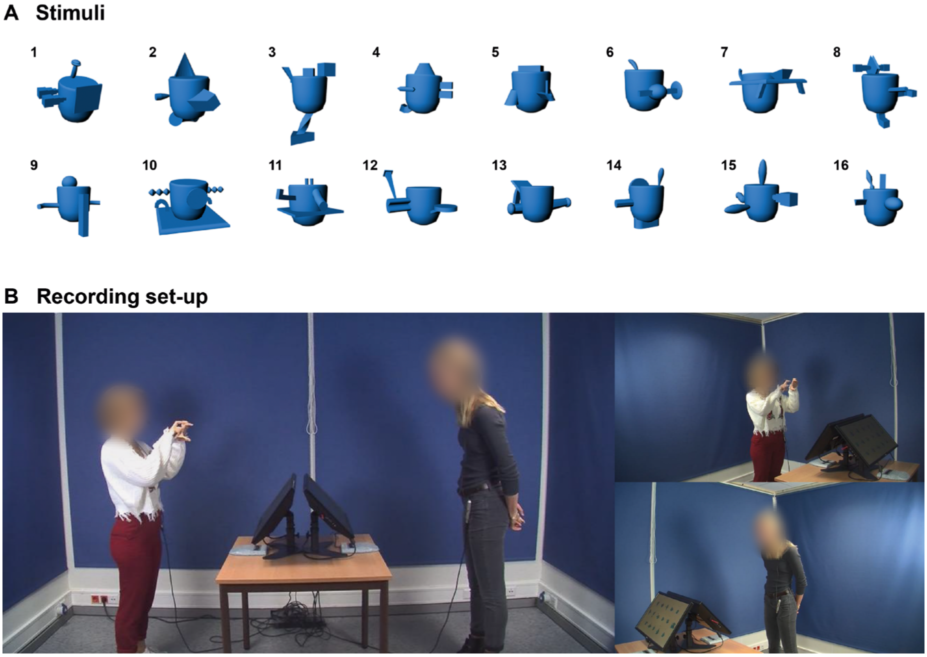


Figure 4.1. Panel (A) shows the “Fribbles” that were used as stimuli. Panel (B) shows the set-up by means of screenshots from the three cameras.

Participants

40 native speakers of Dutch (M_{age} = 22.3 years, $Range_{\text{age}}$ = 18–32 years) took part in the study. The unacquainted participants were randomly grouped into 20 dyads (10 mixed-gender, 6 female-only and 4 male-only dyads). Participants were recruited via the Radboud SONA participant pool system. They provided informed consent and were paid for participation. The study met the criteria of the blanket ethical approval for standard studies of the Commission for Human Research Arnhem-Nijmegen (DCCN CMO 2014/288).

Procedure

Participants took turns as director and matcher to describe and find 16 images of 3D objects called “Fribbles” (adopted from Barry et al., 2014). The Fribbles were displayed in a random order on two screens; the screens were slightly tilted and positioned at hip height to ensure mutual visibility of upper torso and gesturing area (see Figure 4.1, panel B). In each trial, a single target Fribble was highlighted on the director’s screen by means of a red square.

The participants were instructed to communicate “in any way they wanted” in order for the matcher to find the target item on their screen. The matcher indicated their selection by saying the corresponding label (letter or number) out loud. They then pressed a button to go to the next trial, where the participants switched director/matcher roles. Once all 16 Fribbles were matched, the first round was concluded. In total six rounds were completed; that is, all Fribbles were described and matched six times, yielding a total of 96 trials. On average, dyads spent 24.4 minutes on the task (range = 14.2–34.6 minutes). Recordings were made using two head-mounted microphones and three HD cameras.

Coding

ELAN was used for transcription and coding. Speech was segmented and transcribed at the level of Turn Constructional Unit (TCUs; i.e., potentially complete, meaningful utterances; Clayman, 2013; Couper-Kuhlen & Selting, 2017; Schegloff, 2007). For other-initiated repair we created a coding scheme on the basis of Dingemanse et al. (2016), see Appendix A. To facilitate the annotation process and allow for quantitative analyses of speech (in another study), we made sure that the boundaries of the repair annotations correspond to those of the speech annotations. As such, a single repair annotation corresponds to a single TCU or spans multiple TCUs. For minimal sequences (where one repair initiation followed by a repair solution was sufficient to resolve the trouble), we annotated the trouble-source turn, repair initiation and repair solution. As an example, consider Example 1, where each line corresponds to a TCU and the question mark signifies question prosody:

Example (1)

1	TROUBLE SOURCE	A (director):	dit is de hoofdvorm waarbij um er rechts uh een cirkeltje is gewoon zo plat *this is the main shape where um on the right uh a circle is just flat (like this)*
2		A:	links heb je die vorm die half uitgesneden is met een soort spitse punt erin *on the left you have that shape that is half cut out with a sort of pointed point in it*
3	REPAIR INITIATION	B	rechts is een ? *on the right is a ?*
4	REPAIR SOLUTION	A:	ja een cirkeltje die eraan vast is geplakt waar je iets aa- op kan zetten *yes a circle that is pasted on it where you can put something a- on*

For non-minimal sequences (where multiple repair initiations were produced to attend to the same trouble), the trouble-source turn was only annotated for the first repair initiation. Repair initiations were categorised into three types: *open request, restricted request* and *restricted offer*. Annotation and coding decisions were based on the entire multimodal context, taking into account both speech and co-speech gestures.

For co-speech gestures, we annotated the stroke phase (i.e., the meaningful part of the gestural movement; Kita et al., 1998; McNeill, 1992), for the left and right hand separately. Gestures were categorised into three types: *iconic, deictic* and *other*. Iconic gestures are gestures which depict physical qualities of concrete referents or movements or actions related to those referents, deictic gestures are pointing gestures (with extended finger or hand) and other gestures are a heterogeneous group of gestures which do not fit the prior two categories (e.g., beat or interactive gestures). Inter-rater reliability for repair and gesture identification, segmentation and coding was moderate to high (all yielded minimally 75% agreement; for details and additional reliability measures, see Appendix A).

Importantly, gestures were annotated independently from the repair environments in ELAN. Gesture annotation and coding was set up first (for the study reported in chapter 3), which was done in such a way that gestures were annotated on tiers independent from the speech tiers. This made it possible to also annotate gestures which do not (or only partially) overlap with spoken utterances. Furthermore, this approach enabled us to identify and code gestures independently from the repair annotations (which were linked to the speech annotations).

In order to quantitatively analyse the use of gestures in repair sequences, gestures had to be linked to the corresponding repair annotations (trouble-source turn, repair initiation or repair solution). This was done in a semi-automatic fashion: gesture strokes which overlapped completely with a repair annotation from the same speaker were automatically linked to it, while gesture strokes with partial or no overlap were subjected to fine-grained rules and manual inspection (see Appendix A, section "Linking gestures to repair annotations").

Results

In what follows, we will first explore gesture use quantitatively, by presenting the distribution of gestures across the initiating and responding positions of repair sequences of the three initiation types (open request, restricted request and restricted offer)[19]. We will then complement this with qualitative insights derived from illustrative examples, to demonstrate how gestures are used as part of multimodal strategies in these positions.

Quantitative findings

Overall, 378 repair initiations were found in the dataset, divided into 24 open requests, 39 restricted requests and 315 restricted offers. We found a mean of 18.9 repair initiations

[19] Since we have no specific hypotheses about these patterns, we present descriptive findings but refrain from using statistical tests.

per dyad (*SD* = 9.92, range = 6–45), which amounts to a repair initiation occurring once every 1.5 minutes on average. Most of these repair initiations (69.8%) occurred in a minimal sequence; i.e., most of the time a single initiation and solution sufficed to resolve the trouble. However, sometimes more repair initiations were used to address one particular problem. Such non-minimal sequences consisted of a maximum of six repair initiation-solution pairs (*M* = 2.52, *Mode* = 2).[20] Dyads used more repair initiations at the start of the task; in the first round, on average 4.5% of all TCUs where repair initiations, while in the last round this was 1.6% of TCUs (see Figure B6 in Appendix B).

In the set of 378 repair sequences we found a total of 479 co-speech gestures. The majority of the gestures were iconic (89.8%); the remaining gestures were deictic (3.1%) or classified as "other" (7.1%). Note again that participants were not instructed to gesture; all gestures were produced spontaneously. 63% of the repair sequences contain at least one gesture; the average number of gestures produced in a repair turn (repair initiation or solution) is 0.63 (*SD* = 1.23, range = 0–12), with ample variation in means across dyads (M_{range}= 0.06–1.63). None of these turns consisted of manual gestures only; gestures were always combined with speech to form repair initiations and solutions.

The number of gestures produced in the repair initiation and repair solution varies across repair types, as shown in Figure 4.2. Gestures are rarely used in repair initiations when those are open requests (only 1 out of 24 contained a gesture) or restricted request (5 out of 39 contained a gesture). On the contrary, gestures were used in about half (48.9%) of the restricted offers (154 out of 315 contained minimally one gesture). When inspecting repair solutions, we find the reverse pattern; gestures were produced in 79.2%, 66.7% and 24.1% of the responses to open requests, restricted requests and restricted offers respectively. For the distribution of gesture types across repair turns and types, see Table B1 in Appendix B.

In summary, in terms of raw number of gestures per turn (Figure 4.2), the number of gestures per repair turn goes up in repair initiations and down in repair solutions as the repair initiation type of the sequence becomes more specific (open request < restricted request < restricted offer).

[20] The finding that lengthy non-minimal sequences are rare is in line with earlier work (Dingemanse, 2015; Schegloff, 1979; Skedsmo, 2020). In this dataset, there were two sequences which consisted of six repair initiations. In one of these, the Matcher justified the multiple initiations by saying "ik moet het wel goed hebben hoor" [I should get it right] before launching the fourth repair initiation. In the other case, the dyad did not seem to have reached mutual understanding even after six repair initiations; it appears as though the Matcher finally just gave up, as she provided an incorrect answer (on average the task was performed at ceiling with 99.3% accuracy). The quantitative and qualitative evidence from cases like this provides support for the claim that interactants prefer shorter repair sequences, and rarely use more than three attempts to resolve interactional trouble. This has been put forward as a tentative "rule of three", which might apply because after more than three initiations the preference for intersubjectivity no longer outweighs the disruption of progressivity (Dingemanse, 2020).

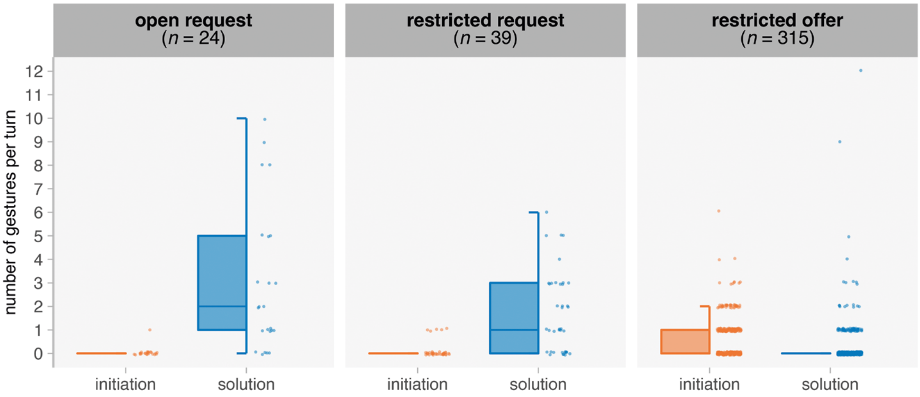


Figure 4.2. The number of gestures in repair initiations (orange) and repair solutions (blue), in sequences of repair types of increasing specificity (open request < restricted request < restricted offer). Some boxplots are just a horizontal line, indicating that the median is 0. Dots are individual datapoints; every dot represents a repair initiation or solution. The number of gestures per turn goes up in repair initiations and down in repair solutions as repair formats become more specific.

Qualitative findings

To further uncover the role of the gestural modality in the other-initiated repair system, we will present examples of how gestures are used together with speech in initiating and responding positions of these three repair types. Note that we have not quantified different usages, and hence we do not make claims about how common the presented multimodal strategies are. Yet the examples are representative of the full range of multimodal strategies found in this dataset, and the usage of gestures in *localising* trouble, *offering* solutions, *clarify*ing trouble and *confirming* offers—as exemplified below—all occur more than once. We discuss repair initiations and repair solutions in turn.

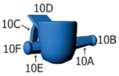
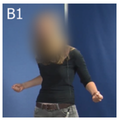
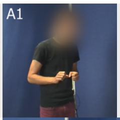
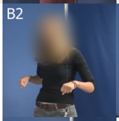
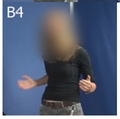
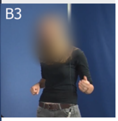
Repair initiations

Open requests. We start with the observation that gestures are rarely used as part of repair initiations of the type open request and restricted request. In the set of 24 open requests, we found only one instance of a gesture (of the type “other”). The matcher raised his index finger while saying “uh wacht even, nog een keer” [uh wait a second, come again?]. This is a conventional (or “pragmatic”) gesture used more commonly to signify “hold on” in spoken dialogue (Uskokovic & Talehgani-Nikazm, 2022). Here, it is used to signal trouble, and to halt the conversation to attend to the problem. It resembles request formats in sign languages, where the manual sign “wait” or “wait-a-minute” (paired with eyebrow movements or another sign with the other hand) can be used to initiate repair (Dively, 1998; Manrique, 2016).

Restricted requests. In restricted requests, people localise trouble; they make clear which part of a prior turn is in need of repair. A key strategy for localising trouble is repetition.

While prior work has already shown that this can involve repetition of speech (Dingemanse et al., 2014; Schegloff et al., 1977) or gesture (Baranova, 2015) here we find that people can also employ multimodal repetition (i.e., repetition of both speech and gesture) to invite repair of a multimodal trouble-source, as shown in Transcript 4.1. In all transcripts, "A" and "B" refer to participants A (standing on the left side) and B (on the right), and the underlined speech temporally overlaps with the gesture strokes depicted in the video still with the corresponding subscript. Whenever multiple video stills are presented on one line, they represent separate gestures. Fribbles along with labels for the distinct subparts (10A, etc.) are added, to indicate more precisely what participants are referring to (with speech and/or gesture). Lines with a grey background are the focal point in each transcript.

Transcript 4.1. Repetition of speech and gesture to localise trouble in a restricted request

	1		B (director):	de gieter *the watering can*	 B1	
	2	TROUBLE SOURCE	B:	((breath/laugh)) die twee balletjes rechts $_{B1}$ *those two balls on the right $_{B1}$*		
	3	REPAIR INITIATION	A:	twee balletjes rechts? $_{A1}$ *two balls on the right?$_{A1}$*	 A1	
	4	REPAIR SOLUTION	B:	ja twee van die uitsteeksels $_{B2}$, links en $_{B3}$ rechts $_{B4}$, uh zit een balletje aan vast *yes two of those protrusions $_{B2}$, left and $_{B3}$ right $_{B4}$, uh is a ball attached to it*	 B2  B4	 B3

$_{B1:}$ two-handed gesture where the fists model subparts 10B+10F (referred to with "balls").

$_{A1:}$ two-handed gesture; similar to $_{B1}$ but the fists are held closer together in front of the torso, the movement of tucking in the fingers is repeated twice.

$_{B2:}$ two-handed gesture depicting 10A+10E; the hands start close together in front of the torso and they then move sideways (depicting the orientation and relative location), the fingers tucked in (depicting the shape).

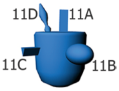
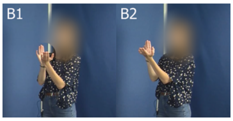
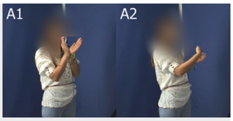
$_{B3:}$ left-handed gesture signifying the relative location of subpart 10F (right hand is post-stroke hold of $_{B2}$).

$_{B4:}$ right-handed gesture signifying the relative location of subpart 10B (left hand is a post-stroke hold of $_{B3}$).

Line 2 is treated as a trouble-source turn, likely due to a gesture-speech mismatch. In that turn, B says "two balls on the right" to refer to subparts 10B+10F, but she iconically depicts the balls by holding her clenched hands on the left and right side of her body. We could also characterise this as a "Fribble-speech" mismatch, as there are no Fribbles with two "balls" (or round shapes) on the right side. In line 3, A initiates repair by repeating the composite utterance: he says “two balls on the right” with a questioning intonation, and also produces a gesture depicting the balls by clenching his fingers, somewhat closer together in front of his body. B responds by clearing up the mismatches in line 4; she now uses two separate gestures to depict the location of the “balls” (one to her left, the other to her right), which are time-aligned to the co-expressive speech (“left” and “right” respectively). Though multimodal restricted requests are rare (only 5 out of the 39 restricted requests contained a gesture), this example thus shows that people can invite clarification of a prior composite utterance through repetition in both speech and gesture.

Restricted offers. Within repair initiations, we find that gestures are used most commonly in restricted offers. These typically serve one of two functions: to localise trouble or to convey a candidate understanding. Transcript 4.2 shows a repair initiation in which gestures are used for both functions.

Transcript 4.2. Repetition of speech and gesture to localise (distant) trouble in a restricted offer

	1	TROUBLE SOURCE	B (director):	een beetje alsof je een boek hebt opengeslagen zo $_{B1}$ en dan er tegenaan $_{B2}$ gezet *a bit as if you have opened a book (like this) $_{B1}$ and then put against $_{B2}$ it*	
				[intervening talk; 43 sec.]	
	2	REPAIR INITIATION	A:	oh je bedoelt met dat opengeslagen $_{A1}$ boek zegmaar dat dat zegmaar dat handvaatje $_{A2}$ van de linkerkan[t (is) *oh you mean with that opened $_{A1}$ book that that’s like the handle $_{A2}$ of the left si[de*	
	3	REPAIR SOLUTION	B:	[ja *[yes*	

$_{B1:}$ two-handed gesture depicting the shape of 11C; the wrists touch, the hands are slightly apart, the fingers extended.

$_{B2:}$ two-handed gesture depicting how 11C is attached to the base shape; similar to $_{B1,}$ but now with a movement from the participants’ very left to the right.

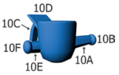
$_{A1:}$ two-handed gesture depicting the shape of 11C; similar to $_{B1.}$

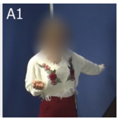
$_{A2:}$ right-handed gesture depicting the relative location and orientation of 11C; the hand is rotated sideward, the fingers extended.

In line 2, A first orients to the trouble by repeating lexical and gestural material from a trouble-source turn (“opened book” + gesture in line 1, referring to 11C). Whereas repair initiations most commonly directly follow the trouble-source turn, here the localisation spans many intervening turns (in which the participants attended to the same problem, but also referred to other Fribble subparts). As such, multimodal repetition serves as a powerful technique, used to tie back to a multimodal trouble source in a distant trouble source turn (Schegloff, 2000). Subsequently, A offers a candidate understanding of that trouble-source; she produces an iconic gesture (gesture A2) together with the lexical phrase "handle of the left side" (referring to the same subpart). In this composite utterance, speech and gesture were co-expressive, with the gesture providing a more detailed representation of "left side" by depicting the location of the "handle" subpart relative to the base shape (represented with her body).

Besides such synchronous speech-gesture combinations, speech and gesture can also be combined to offer a candidate understanding in such a way that only the gesture conveys specific information about the referent, which is indexed by speech. Consider Transcript 4.3.

Transcript 4.3. Providing a gestural candidate understanding in a restricted offer



1	TROUBLE SOURCE	B (director):	dit is de twee armen met bolletjes eraan B1 *this is the two arms with bulbs attached* B1	
2			en uh een ding naar ach[teren *and uh a thing towards the [back*	
3	REPAIR INITIATION	A:	[die zo? A1 *[the one (like this)?* A1	
4	REPAIR SOLUTION	B:	ja *yes*	

B1: two-handed gesture where curved handshapes depict the round shapes of 10B+10F, somewhat away from the body thereby depicting the subparts' positions relative to the base shape.

A1: two-handed gesture where clenched fists model 10B+10F, right extended arm models 10A and left tucked-in arm models 10E.

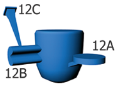
In line 1, B (the director) refers to subparts 10B+10F of the Fribble by means of a multimodal expression ("de twee armen met bolletjes eraan" [the two arms with bulbs attached] ((gesture))). In line 3 (which partly overlaps with B's continued description in line 2), A seeks confirmation by offering a multimodal candidate understanding. Here, it is the gesture that is offered as a candidate understanding, to which attention is drawn through deictic speech ("zo" ['like this']). So, unlike in Transcript 4.2 where the speech and gesture were co-expressive and jointly constituted a candidate understanding, in Transcript 4.3 the bulk of the communicative work is done through the gestural modality only, conveying semantic information beyond what is expressed in speech. The gesture in the repair initiation (line 3) can be taken as a "gestural rephrasing" (Tabensky, 2001) of the gesture in the trouble-source turn (line 1): in both cases, the hands represent the bulbs (10B+10F), but in the second gesture the "arms" that they are attached to are depicted more explicitly; the left extended arm depicts subpart 10A, while the right tucked-in arm depicts the shorter subpart 10E. As such, some of the characteristics of the gesture are kept intact (two-handed gesture; both arms to the side of the body; hands representing bulbs), while other features are changed for the purpose of negotiating meaning (a process described in Jokipohja & Lilja, 2022; Rasenberg et al., 2022; Tabensky, 2001).

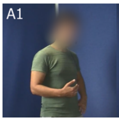
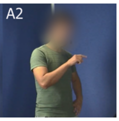
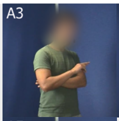
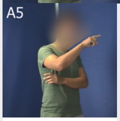
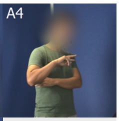
Together, Transcripts 4.2 and 4.3 show how iconic gestures can be used for different purposes (even within one repair initiation), through various speech-gesture combinations. This versatility helps to characterise the quantitative finding that almost half of the restricted offers contained one or more gestures.

Repair solutions

Responding to open and restricted requests. Turning to repair solutions, we start with responses to open and restricted requests, which invite repetition and/or further clarification of the trouble. Consider Transcript 4.4.

Transcript 4.4. Multimodal clarification in response to a restricted request



1	TROUBLE SOURCE	A (director):	uh van links lijkt het op een hondenkoekje $_{A1}$ en dan daar $_{A2}$ bovenop staat een trompet *uh from the left it looks like a dog biscuit $_{A1}$ and then there $_{A2}$ on top stands a trumpet*		
2	REPAIR INITIATION	B:	((laughs)) wat? een hondenkoekje met een trompet? *what? a dog biscuit with a trumpet?*		
3	REPAIR SOLUTION	A:	ja *yes*		
4	REPAIR SOLUTION	A:	links heeft hij zo zo'n twee $_{A3}$ twee ronde uh uitsteeksels $_{A4}$ en daarboven zit gewoon zo'n $_{A5}$ lange toeter trompet *on the left it has (like this) such two $_{A3}$ two round uh protrusions $_{A4}$ and above that is just such a $_{A5}$ long horn trumpet*	 	

$_{A1:}$ right-handed gesture depicting the shape and orientation of 12B; the fingers are loosely extended, and the hand makes a single sideward movement.

$_{A2:}$ right-handed gesture depicting the shape and relative location of 12C; the extended index finger moves up and down once.

$_{A3:}$ right-handed gesture depicting 12B; the index and middle finger are extended, representing the upper and lower part of 12B.

$_{A4:}$ right-handed gesture depicting 12B; similar to $_{A3}$, but the hand makes a sideward movement (similar to $_{A1}$).

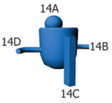
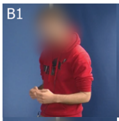
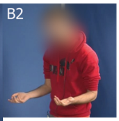
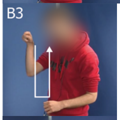
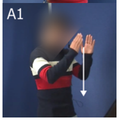
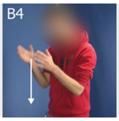
$_{A5:}$ right-handed gesture depicting 12C; similar to $_{A2}$, but the upward movement is repeated once more.

In line 1, participant A provides a multimodal description of subparts 12B and 12C. In line 2, participant B initiates repair by combining an open request ("what?") with a restricted request (“a dog biscuit with a trumpet?”), known as a “double” repair initiation (Kim, 1999). Participant A’s response consists of a clarification, where the earlier holistic referring expression is unpacked using more concrete, geometrical terms: “two round protrusions” instead of “dog biscuit” (for 12B) and adding “long horn” to “trumpet” (for 12C). Notably, the gestures that are produced as part of the repair solution are very similar to those in the trouble-source turn, yet they are slightly changed and put into a new relationship to speech. Whereas the first gesture was produced along with the phrase “honden(koekje)” [dog biscuit] in the trouble-source turn (line 1), in the repair solution two similar gestures are produced while saying “two” and “protrusions” (line 4). The gestures look different in that they have two extended index fingers (with the rest of the fingers curled in), thereby more explicitly depicting the two subcomponents of what looks like a “dog biscuit”, which is verbally expressed as well. The second gesture from the trouble-source turn (line 1) is repeated in the repair solution (line 4) while adding deixis in speech (“zo’n” [‘like this’]), thereby drawing more attention to it—a phenomenon that has previously been found in responses to “negative addressee feedback” as well (Holler & Wilkin, 2011b). Together, the speech and gesture modifications in the repair solution can be regarded as a *multimodal* clarification, a strategy we also saw in Transcript 4.1. Here, the director relies on iconicity in the gestural modality to connect the holistic construal of the referent (“dog biscuit”, “trumpet”) to a more concrete, compositional one (“two round protrusions”, “long horn”).

Responding to restricted offers. The primary work for repair solutions in response to restricted offers is to confirm or correct a proposed candidate understanding. Transcript 4.5 provides an example of how gestures can be used to confirm an offer through repetition. The repair initiation on line 2 is a restricted offer, which (similarly to Transcripts 4.2 and 4.3), offers a *multimodal* candidate understanding (of what 14C looks like) for confirmation. The multimodal offer is subsequently accepted in a multimodal way; by combining a verbal confirmation (“yes exactly”) with a repetition of the offered gesture. As such, it is not just the spoken part of the restricted offer that is accepted, but the whole multimodal offer.

Gestural repetition as shown in Transcript 4.5 is one way in which iconic gestures can be used in repair solutions in response to restricted offers. Another way is when the candidate understanding was in fact incorrect, thus requiring more clarifying work in the repair solution (recruiting iconic gestures in the way as described for responses to requests for clarification). But we also find gestures of the category “other” (i.e., gestures which were not iconic nor deictic) in repair solutions. This category of gestures and the way they are used in repair solutions in response to restricted offers is very diverse, comprising for example beat gestures, hedging gestures (tilting palm-down hand while saying “sort of”) and metaphorical gestures (pointing to one’s head while saying “yes ((laughs)) you call it that, I am going to remember it”).

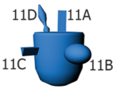
Transcript 4.5. Repetition of gesture to confirm a restricted offer

	1	TROUBLE SOURCE	B (director):	oké en dan dan heb je dus aan de voorkant (een) B1 daar zie ik er B2 ook niet heel veel van een zo'n groot blok B3 *okay and then then you thus have on the front (a) B1, I don't see B2 so many of those, a such a big block B3*	  
	2	REPAIR INITIATION	A:	gewoon die zo recht naar beneden A1 gaat? *just that goes straight down A1 (like this)?*	
	3		B:	zie je re- *do you see ri-*	
	4	REPAIR SOLUTION	B:	ja precies B4 *yes exactly B4*	

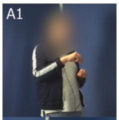
B1: two-handed gesture depicting the relative location of 14C; the hands are loosely held in front of his torso, and then moved towards himself.
B2: two-handed "palm-up open-hand" gesture (cf. Cooperrider et al., 2018; Müller, 2017).
B3: right-handed gesture drawing the shape of 14C with the index finger.
A1: two-handed gesture depicting the shape and orientation of 14C; the hands are slightly apart and make a single downward motion.
B4: two-handed gesture depicting 14C; almost identical to A1.

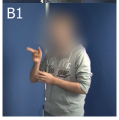
One type of "other" gesture that we find in the dataset has a more distinctive function for the repair system, as shown in Transcript 4.6. Here, the matcher offers a multimodal candidate understanding of 11D in line 1, which is confirmed by the director in line 2. This repair solution contains a verbal confirmation ("oh yes that is a good one"), paired with a gesture which is a quick flick of the hand (lasting about 60 msec). This specific gesture is only produced by this participant in the dataset (always for the same purpose; i.e., to confirm a restricted offer), and might therefore be regarded as idiosyncratic. Yet importantly it is a gesture that has been recognised as a conventionalised, interactional gesture of the type "citing gestures", and more specifically, of the subcategory "acknowledgements" which "indicate that the speaker saw or heard that the addressee understood the speaker" (Bavelas et al., 1995, p. 397).

Transcript 4.6. Gestural acknowledgment to confirm a restricted offer



1	REPAIR INITIATION	A (matcher):	[ja het is een soort pipet uh dingetje [toch? A1 *[yes it is a sort of pipet uh thingy [right? A1*
2	REPAIR SOLUTION	B:	[oh ja B1 [dat is wel een goede *[oh yes B1 [that is a good one*





A1: right-handed gesture depicting 11D; the fingers are pressed together multiple times as if squeezing the bulb of a pipet.

B1: right-handed gesture which is a very quick flick of the hand with an extended index finger.

Discussion

Summary and discussion of main findings

To further our understanding of how people collaboratively and multimodally co-construct meaning, this study set out to investigate multimodal sequences of other-initiated repair, in which people jointly solve problems with perceiving or understanding. We focused on how co-speech gestures are used together with speech in repair sequences in task-based interactions with novel referents where people are likely to encounter issues with understanding. Our results are largely line with isolated findings spread across earlier studies; but by quantitatively and qualitatively investigating how and when gestures are used as part of multimodal strategies in the repair system as a whole, we were able to put those puzzle pieces together.

Other-initiated repair occurred frequently in this dataset (once every 1.5 minutes on average), and we found ample use of co-speech gestures in repair sequences, especially iconic gestures. Yet there is large variation in how often and how gestures are used across the initiating and responding positions of the different repair formats. In what follows we will make sense of the distribution by connecting the quantitative findings to the qualitative insights, in order to answer the question we started out with: when, why and how people use co-speech gestures together with speech to resolve interactional trouble.

Repair initiations

Why do people use co-speech gestures in repair initiations? Our qualitative findings show that gestures are rarely used to signal trouble, but instead are used together with speech to localise trouble and offer candidate understandings, thereby restricting the problem space. This helps to explain their distribution across repair initiation types; they are predominantly used in the most specific repair initiation type (restricted offers). As such, our findings go beyond a quite extensive body of work on embodied resources in repair initiations, which has focussed on bodily and facial signals whose main purpose is to *signal* trouble, without specifying where or what the trouble is[21] (Andrews, 2014; Hömke, 2019; Li, 2014; Manrique, 2016; Manrique & Enfield, 2015; Oloff, 2018; Rasmussen, 2014; Seo & Koshik, 2010; Skedsmo, 2020).

How do gestures make repair initiations more specific? In terms of localisation trouble, a common strategy is to repeat the trouble, which can in principle be done with any semiotic resource that affords repetition (Baranova, 2015; Dingemanse & Enfield, 2015; Manrique, 2016; Skedsmo, 2020). In this study we provide integrative evidence that this indeed holds for both speech and gesture, and we show that they can be combined to target multimodal trouble. Repetition of gestures differs from speech in that the gesture can visually present the repairable, which might be a more effective way to tie back to a distant trouble-source turn, as in Transcript 4.2 (cf. Baranova, 2015 on the use of gestural repetition to "upgrade" a repair initiation). Such usage sets manual co-speech gestures apart from facial and bodily signals which can also be used to signal trouble, but which have "little power" to specify where or what that trouble is (Albert & de Ruiter, 2018). Consequently, facial and bodily signals are only used in close proximity to the trouble-source turn, whereas here we found that manual co-speech gestures can still be used to tie back to the trouble-source after several intervening turns.

Besides gestural repetition to locate trouble, we found that gestures can be used to convey a candidate understanding as part of restricted offers. Some examples were reported in the literature, showing that deictic gestures are used to resolve referent ambiguities (e.g., Dingemanse, 2015) and that people provide iconic depictions of candidate understanding, where the gesture is co-expressive with speech (Bertrand et al., 2013; Sikveland & Ogden, 2012). In this dataset we also find co-expressive gestures in restricted offers, but with Transcript 4.3 we have additionally shown that iconic gestures can be the main resource for conveying semantic meaning (with speech merely serving a deictic function; "like this?"). Furthermore, we revealed that partial gestural repetition can be used as a strategy in restricted offers; people change some form features of a trouble-source turn gesture to negotiate mutual understanding (cf. Rasenberg et al., 2022; Tabensky, 2001). As such, we find that different types of repetition (with full or partial form overlap) are used for differential purposes (localising trouble versus depicting a candidate understanding), highlighting the need to consider cross-participant repetition—or "alignment"—carefully (Rasenberg, Özyürek, et al., 2020).

[21] Though recipients interpret these embodied displays as signals of trouble with either hearing or understanding, and design their repair solution accordingly (Hömke, 2019; Oloff, 2018).

Repair solutions

What is the purpose of co-speech gestures in repair solutions? Our quantitative and qualitative findings show that gestures are used differently in responses to open and restricted requests on the one hand, and responses to restricted offers on the other hand. Starting with responses to open and restricted requests, we showed that iconic gestures can be used to clarify a trouble-source turn. How do they do that? Prior work focusing on gestures has revealed that gestures in repair solutions are more salient than gestures in the trouble-source turn; they are more precise and larger, and attention is drawn to it through deixis in speech and gaze (Hoetjes, Krahmer, et al., 2015; Holler & Wilkin, 2011b; see also Jokipohja & Lilja, 2022; Kendon, 2004; Olsher, 2008). Here we considered how both speech and gesture differ between the trouble-source turn and the repair solution, and in Transcript 4.4 we showed that besides changing the gestures, there were also prominent changes in speech. Specifically, abstract referring expressions ("dog biscuit", "horn"), were replaced by more concrete, geometrical terms (cf. Schegloff, 2004 on the replacement of technical or uncommon reference terms). These new (better recipient designed) descriptions were produced together with the (slightly modified) gestures, providing a new speech-gesture ensemble. This corroborates insights from research on naturalistic interactions, which showed that people use more specific words and gestures to clarify trouble (Jokipohja & Lilja, 2022). Furthermore, the potential of these multimodal clarification strategies is underlined by our quantitative finding that gestures are commonplace in responses to open and restricted requests (79.2% and 66.7% of these turns contain one or more gestures), in line with quantitative results from studies on gestures in responses labelled as "addressee feedback" or "trouble spots" of various kinds in prior work (Alibali et al., 2013; Hoetjes, Krahmer, et al., 2015; Holler & Wilkin, 2011b).

Finally, in repair solutions in response to restricted offers gestures are less prevalent (these turns usually contain no gestures, or only one). This also matches earlier work, which revealed that compared to the trouble-source turn, gesture rates decrease in responses to confirmation requests which included a correct candidate understanding (Holler & Wilkin, 2011b). Here we have qualitatively inspected such turns, which revealed two gestural strategies to confirm offers; i) repetition of the iconic gesture which was part of the restricted offer, as shown in Transcript 4.5, and ii) the use of "acknowledgement" gestures, as shown in Transcript 4.6 (Bavelas et al., 1995). As such we showed that even though people use fewer gestures in response to restricted offers, gestures appear to be very effective in confirming offers, especially when the offer was already multimodal. As such, these findings compliment prior qualitative work which has described the use of (repetition of) deictic gestures in repair solutions (Dingemanse, 2015; Sidnell, 2007).

Limitations

An important question is whether the gestures as found in the current dataset of task-based interactions is an accurate reflection of gesture use in natural interactions. One domain where they clearly differ, is with respect to the use of deictic gestures. We found very few, probably because people did not have shared visual access to the referents under discussion, while in natural interactions pointing is a straightforward way to resolve referent ambiguities (Dingemanse, 2015; Floyd, 2020; Levinson, 2015). Conversely, the nature of the 3D stimuli in our task might have yielded an inflated amount of iconic gestures. Despite these differences, we reported converging evidence of patterns and strategies reported in prior work. We also take the observed distribution of the frequency of co-speech gesture across initiating and responding positions of the three repair types to be informative for our understanding of other-initiated repair as a multimodal system, where future work can investigate which types of gestures are used across different conversational settings.

A second concern is the generalisability of the findings which were obtained in a specific cultural-linguistic context. Just like different languages have different verbal resources for the repair system (such as noun-class specific interrogatives in Murrinh-Patha; Dingemanse et al., 2014; or fingerspelling in sign languages; Skedsmo, 2020), so different languages might show variation in gestural resources. Though there is evidently cultural variation in relative frequency, gesture size or position and recruitment of emblematic gestures (see e.g., Cooperrider, 2019; Kita, 2009) which can shape other-initiated repair practices, the findings of this study suggest that there might also be general principles that can be found across languages and cultures. Likely candidates are gradient forms of gestural repetition to localise trouble or negotiate understanding (of which we found various examples in this dataset, reported here and in Rasenberg et al., 2022), and the use of iconic or indexical (manual or non-manual) gestures to offer or confirm candidate understandings (cf. work on Russian, French, Siwu, Yélî Dnye and Norwegian; Baranova, 2015; Bertrand et al., 2013; Dingemanse, 2015; Levinson, 2015; Sikveland & Ogden, 2012).

Conclusion

This study advances our understanding of other-initiated repair as a semiotically-diverse, optimally organised system for resolving interactional trouble. Our results are in line with isolated findings spread across earlier studies. However, by systematically investigating how and when speech and gestures are used to form multimodal strategies in the repair system as a whole, we were able to put those puzzle pieces together for the first time.

We have shown how people combine speech with the dynamic and iconic properties of the gestural modality to localise trouble and offer candidate understandings (in repair initiations) and to clarify trouble or confirm offers (in repair solutions), where a key strategy is to repeat or modify the multimodal contributions of the conversational partner. These findings underscore the notion that multimodal turn design is an interactive process. People orient to the unfolding turns of their partner, and treat both speech and gesture as meaningful contributions which can be interrogated (in repair initiations) or affirmed (in repair solutions). Together the findings provide novel insights into how people incrementally build up understanding through coordinative, multimodal efforts.

Chapter 5

The multimodal nature of communicative efficiency in social interaction

This chapter is based on:
Rasenberg, M., Pouw, W., Özyürek, A.,
& Dingemanse, M. (2022).
The multimodal nature of communicative
efficiency in social interaction. *Scientific Reports*, *12*, 19111.
https://doi.org/10.1038/s41598-022-22883-w

Data and code are available at:
https://doi.org/10.34973/12dp-9q56

Abstract

How does communicative efficiency shape language use? We approach this question by studying it at the level of the dyad, and in terms of multimodal utterances. We investigate whether and how people minimise their joint speech and gesture efforts in face-to-face interactions, using linguistic and kinematic analyses. We zoom in on other-initiated repair—a conversational microcosm where people coordinate their utterances to solve problems with perceiving or understanding. We find that efforts in the spoken and gestural modalities are wielded in parallel across repair turns of different types, and that people repair conversational problems in the most cost-efficient way possible, minimizing the joint multimodal effort for the dyad as a whole. These results are in line with the principle of least collaborative effort in speech and with the reduction of joint costs in non-linguistic joint actions. The results extend our understanding of those coefficiency principles by revealing that they pertain to multimodal utterance design.

Introduction

In joint actions, people coordinate their behaviours in order to achieve joint goals (Clark, 1996; Sebanz et al., 2006). Whether people are moving a couch or having a chat, joint action appears to be organised according to a principle of efficiency or effort minimisation (Engelbrecht, 2001; Gibson et al., 2019; Levshina & Moran, 2021; Ray & Welsh, 2011; Zipf, 1935). Empirical work on joint action shows that this effort minimisation appears to target overall *joint effort* (or *coefficiency*) rather than individual effort (Santamaria & Rosenbaum, 2011; Török et al., 2019, 2021). Work on spoken language likewise suggests that people work together to minimise the cost for the dyad as a social unit—known as *the principle of least collaborative effort* (Clark & Brennan, 1991; Clark & Schaefer, 1987; Clark & Wilkes-Gibbs, 1986). One consequence for the study of efficiency in language is that language use is not about idealised speakers producing optimal one-off utterances; instead, we need to consider the work that interacting participants jointly undertake to actively construe possible meanings.

The notion of coefficiency is in principle agnostic to the type of behaviour involved. That is, joint action is recognised to involve a complex interplay of efforts exerted through various types and levels of behaviour. However, when it comes to language use, efficiency is usually studied in *unimodal* ways (by focusing on written representations of speech), where communicative acts are considered to be *linear* (one word after the other). Complementary or parallel contributions across modalities are overlooked in accounts of efficiency in human languages (Gibson et al., 2019; Levshina & Moran, 2021), despite the communicative capacities of composite utterances as revealed by research on multimodal interaction (Enfield, 2009; C. Goodwin, 1979, 1981; Kendon, 2004; Stivers & Sidnell, 2005). So, work on efficiency in coordinated spoken language use has yet to take into account how simultaneous articulators are concurrently employed to convey information (for work on sign language, see Slonimska et al., 2020). Here we take on the challenge to study how people efficiently coordinate multiple types of communicative behaviour, by investigating if and how people minimise joint speech and gesture efforts in a task-based conversational setting.

We focus on stretches of conversation where people explicitly coordinate their utterances with the goal of jointly solving problems of perceiving or understanding—known as other-initiated repair (Schegloff, 2000; Schegloff et al., 1977). In a typical sequence of other-initiated repair, one participant temporarily halts the conversation in order to ask for clarification with a repair initiation like "huh?" (OPEN REQUEST), "who?" (RESTRICTED REQUEST), or "like this ((gesture))?" (RESTRICTED OFFER) (Dingemanse & Enfield, 2015; Drew, 1997; Hayashi et al., 2013; Kitzinger, 2013; Robinson & Kevoe-Feldman, 2010), to which their conversational partner responds with a repair solution. After having jointly resolved the trouble, the participants end the repair sequence and the main conversation continues

(Jefferson, 1972; Schegloff, 1992). Repair initiations and solutions are defined strictly in terms of sequential positions in conversation, where the turns themselves can recruit any combination of communicative modalities (Andrews, 2014; Floyd et al., 2016; Hoetjes, Krahmer, et al., 2015; Holler & Wilkin, 2011b; Hömke, 2019; Kendrick, 2015; Levinson, 2015; Li, 2014; Manrique, 2016; Manrique & Enfield, 2015; Mortensen, 2016; Oloff, 2018; Rasmussen, 2014; Seo & Koshik, 2010; Sikveland & Ogden, 2012; Skedsmo, 2020; Svensson, 2020).

Sequences of other-initiated repair have played a key role in the development of the notion of least collaborative effort for English task-based and telephone interactions (Clark & Schaefer, 1987; Clark & Wilkes-Gibbs, 1986) and in its generalisation to co-present conversational interaction across diverse languages (Dingemanse, Roberts, et al., 2015). This work revealed that people collaboratively resolve trouble while minimizing their joint efforts in two ways. First, recipients who signal trouble prefer to use restricted formats (e.g., "Which one?" or "You mean X?") over open formats (e.g., "Huh?"), meaning that they initiate repair in the most specific way possible (the *specificity principle*). Second, the more specific the repair initiation, the longer the repair initiation (involving more speech effort), thereby minimizing the efforts needed for the sender to resolve the trouble in the repair solution (the *division of labour principle*). However, this prior work focused exclusively on unimodal utterances, either by classifying the referential formats of noun phrases, or by computing the orthographic length of turns. Gesture efforts in spoken language have been overlooked, even though prior research has shown that manual co-speech gestures can play an important role in repair initiations (Mortensen, 2016; Sikveland & Ogden, 2012) and repair solutions (Alibali et al., 2013; Hoetjes, Krahmer, et al., 2015; Holler & Wilkin, 2011b; Olsher, 2008; Sidnell, 2007). Since gestural efforts to convey meaning have not been incorporated in the division of labour equation, we cannot be sure that the speech-centred findings hold water for interactions in their true multimodal form.

Adopting a multimodal perspective is also warranted in light of recent studies showing that various interactional strategies (initially discovered based on speech-centred research) extend to the design of multimodal utterances. For example, people modulate both speech and gesture when trying to get a message across in noisy environments (multimodal Lombard effect; Trujillo et al., 2021); adapt both speech and gestures to the degree of knowledge that is shared with a recipient (multimodal audience design; Holler & Bavelas, 2017); and are likely to use cross-speaker repetition of both speech and gesture for establishing joint reference to novel objects (multimodal alignment; Rasenberg et al., 2022). These findings reinforce the notion that speech and co-speech gestures operate as part of an integrated system (Kendon, 2014; Kita & Özyürek, 2003; McNeill, 1992), with people flexibly deploying and coordinating their use of both modalities to engage in joint meaning-making (Chui, 2014; de Fornel, 1992; C. Goodwin, 2000; M. H. Goodwin & Goodwin, 1986; Holler & Wilkin, 2011a; Mondada, 2011; Rasenberg et al., 2022; Sikveland & Ogden, 2012).

In work on co-speech gesture, the notion of division of labour is sometimes used in reference to how effort is divided between the modalities of speech and gesture in one speaker's utterances (de Ruiter et al., 2012). Here instead we focus on the dyad as a social unit, where our primary interest is how effort is distributed across contributions of different speakers, taking into account both speech and gesture.

To investigate the distributions of multimodal effort at the dyad level we focus on sequences of other-initiated repair in task-based interaction. Other-initiated repair has several properties that make it an ideal testing ground for studying efficiency in social interaction. First, it is a miniature coordination problem solved in real-time by two participants, making it a relevant domain for understanding joint actions more broadly (Albert & de Ruiter, 2018). Second, its sequential structure (an insert sequence composed of an initiation and proposed solution) is cross-linguistically well-attested and highly frequent (Dingemanse, Roberts, et al., 2015; Fox et al., 1996). Third, it comes in a small number of formats that we can compare in terms of multimodal effort and frequency of use. By tracking participants' speech and gesture efforts in these conversational enclosures, we test whether *multimodal* contributions are optimised for least collaborative effort.

We study social interactions between participants as they carry out a director/matcher task designed to present them with coordination problems to be solved on the fly using multimodal communication. In the task, participants take turns referring to novel 3D shapes (Figure 5.1, panel A). Standing face-to-face and instructed to communicate in any way they want, participants recruit multimodal utterances in relatively free-flowing interactions in order to negotiate mutual understanding and jointly solve the task. Speech and gesture behaviours were recorded with head-mounted microphones, cameras and markerless motion tracking devices.

For the spoken modality, we operationalised effort as the number of orthographic characters per repair turn, as this allows us to compare our findings to those of Dingemanse et al. (2015). Though not a direct measure of talk-in-interaction, orthographic length may be a reasonable proxy because (i) it is not affected by speech rate (while turn duration is) and (ii) it normalises length across different speakers. In our dataset orthographic length strongly correlates with the duration of the repair turn ($r = 0.93$, $p < .001$), in line with earlier work (Dingemanse, Roberts, et al., 2015). For the gesture modality, we use the number of submovements of manual co-speech gestures. This has been used as a kinematic measure of complexity and effort before (Trujillo et al., 2021; Pouw, Dingemanse, et al., 2021; Vesper et al., 2021) and measures akin to it have been shown to correlate with the number of information units in gestures as interpreted by human coders (Pouw, Dingemanse, et al., 2021). While perfect equivalence of measures across modalities is impossible, those proposed here are comparable in the sense that a) both speech and co-speech gesture are used to negotiate meaning in other-initiated repair sequences, and

b) orthographic characters and submovements can both be used as a quantitative proxy for the amount of information that is (verbally or visually) conveyed by a repair turn.

We first investigate speech and gesture efforts separately, where we explore how these efforts are distributed across sequential positions (repair initiation and solution) and repair types (open request, restricted request and restricted offer). To investigate the division of multimodal effort between people, we compute a measure of multimodal effort by summing (standardised) speech and gesture efforts. We hypothesise that the type of repair initiation predicts how the joint amount of multimodal effort is divided between people, similarly to what has been found for the division of speech efforts (Dingemanse, Roberts, et al., 2015). That is, we hypothesise that the more specific the repair initiation (open request < restricted request < restricted offer), the higher the proportion of the multimodal cost paid in the repair initiation relative to the total cost paid in the initiation and solution together. Finally, in line with the principle of least collaborative effort (Clark & Brennan, 1991; Clark & Schaefer, 1987; Clark & Wilkes-Gibbs, 1986), we predict that people design their utterances so as to minimise the total amount of multimodal effort for the dyad as a whole. Specifically, we hypothesise that the repair type which yields the smallest amount of joint multimodal effort will be used most frequently.

Methods

Participants

Twenty dyads took part in the study (10 mixed-gender, 6 female-only and 4 male-only dyads, M_{age} = 22.3 years, $Range_{\text{age}}$ = 18–32 years). The unacquainted participants were recruited via the Radboud SONA participant pool system. Participants provided informed consent and were paid for participation. The participants who are visible in the figures provided informed consent to publish the images in an online open access publication. The study met the criteria of the blanket ethical approval for standard studies of the Commission for Human Research Arnhem-Nijmegen (DCCN CMO 2014/288), and was conducted in accordance with relevant guidelines and regulations.

Apparatus and materials

The stimuli were 16 images of novel 3D objects, called “Fribbles” (adapted from Barry et al., 2014), see Figure 5.1, panel A. During the interaction, participants were standing face-to-face, where each had their own button box and screen (24” BenQ XL2430T), slightly tilted, and positioned at hip height to ensure mutual visibility of upper torso and gesturing area (see Figure 5.1, panel B). The Fribbles were presented on these screens on a grey background in rows of 5, 6, and 5 figures respectively, in a size of about 4x4 cm per figure, with a corresponding label next to it. The order of the Fribbles was random and varied for the participants (but was constant across dyads). Audio was recorded with head-mounted microphones (Samson QV) and videos were made with three HD cameras (JVC GY-HM100/150). 3D motion tracking data was collected using two Microsoft Kinects V2 (for 25 joints, sampled at 30 Hz).

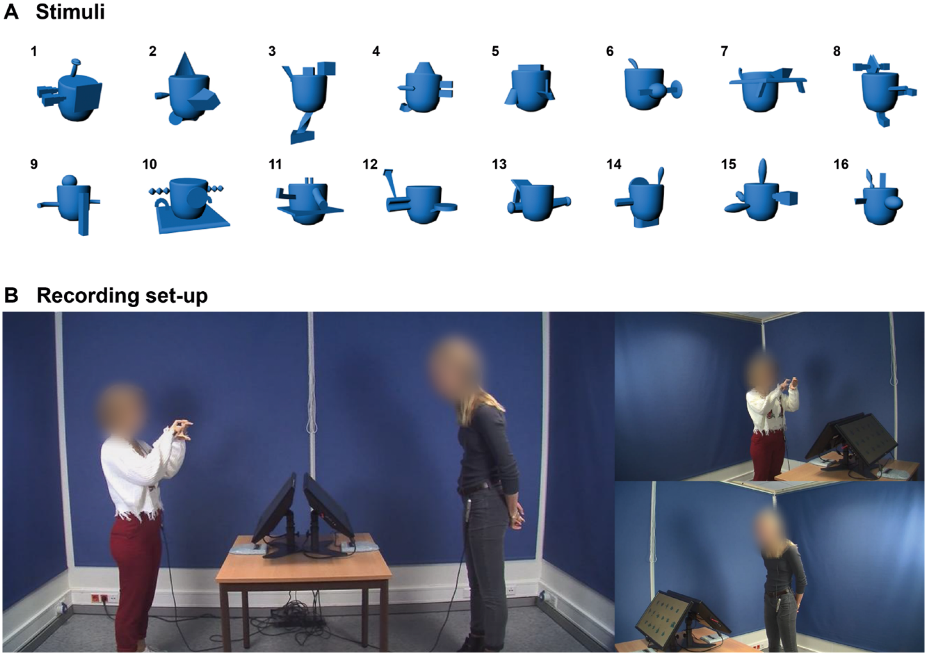


Figure 5.1. Panel (A) shows the "Fribbles" that were used as stimuli. Panel (B) shows the set-up by means of screenshots from the three cameras.

Procedure

Participants were assigned director/matcher roles. In each trial, a red triangle highlighted a single target Fribble on the director's screen. The participants were instructed to communicate in order for the matcher to find the target item on their screen. To indicate their selection, the matcher said the corresponding label out loud and pressed a button to go to the next trial, where the participants switched director/matcher roles. After matching all 16 Fribbles, a new round would start; in total six rounds were completed, yielding a total of 96 trials. No time constraints were posed and the participants did not receive feedback about accuracy. Participants were told that they were "free to communicate in any way they want" (an instruction phrased to be agnostic about communicative modality, i.e., speech and/or gesture), and that their performance would be a joint achievement. Dyads spent 24.4 minutes on the task on average (*range* = 14.2–34.6 minutes).

Analysis

The audio-video data were annotated in ELAN (version 5.8); data processing and statistical analyses were performed with the R statistical programme (version 4.0.2).

Speech was segmented into Turn Constructional Units (TCUs; i.e., potentially complete, meaningful utterances; Clayman, 2013; Couper-Kuhlen & Selting, 2017; Schegloff, 2007) and orthographically transcribed based on the standard spelling conventions of Dutch. Other-initiated repair was coded based on a modified version of the coding scheme by Dingemanse et al. (2016). We annotated the trouble source, repair initiation and repair solution, where the boundaries of those annotations corresponded to the speech annotations (where a single repair annotation could correspond to a single TCU or span multiple TCUs). Subsequently, repair initiations were categorised into three types: *open request, restricted request* and *restricted offer*. Details on the coding procedure including examples can be found in Appendix A. Inter-rater reliability for repair identification, segmentation and coding was moderate to high (all yielded minimally 75% agreement; for details and additional reliability measures, see Appendix A).

For co-speech gestures, the stroke phase was annotated for gesture units (i.e., the meaningful part of the gestural movement (Kendon, 2004; McNeill, 1992)), for the left and right hand separately. Inter-rater reliability was substantial for gesture identification, segmentation and coding (minimally 75% agreement; for details and additional measures, see Appendix A). We considered gestures to be part of a repair turn when the gesture stroke completely overlapped with the repair annotation (which was the case for 91.5% of the gestures), but used fine-grained rules and manual inspection in case of partial or no overlap (see the section "Linking gestures to repair annotations" in Appendix A). All types of manual co-speech gestures were included in the analysis, but the majority of the gestures in the dataset are iconic (89.8%).

Submovements were computed for each gesture stroke, of which the onset and offsets were determined by the manual annotations. The calculation was based on the position of the left- and right-hand tips in 3D space. To ignore noise-related jitter, we smoothed the position traces, and their derivatives (3D speed) with a third order Kolmogorov-Zurbenko (KZ) filter with a span of 2. 3D gesture speed was used to determine submovements, which was based on the speed of the left or right hand for one-handed gestures, or the summed speed of both hands in the case of two-handed gestures. The number of submovements was then computed by identifying local maxima peaks in the 3D speed time series (Pouw, Dingemanse, et al., 2021; Trujillo et al., 2019). To this end, we used R-package *pracma* and considered a peak to be a submovement when it exceeded at least 10 cm/s and only if they had at least 100ms distance between adjacent peaks. The minimum amount of submovements per gesture stroke is 1 (i.e., static strokes are considered to consist of 1 submovement). Two examples of gestures along with their submovement profile are presented in Figure 5.2, panel A (for more examples, see the supplementary materials in Appendix B).

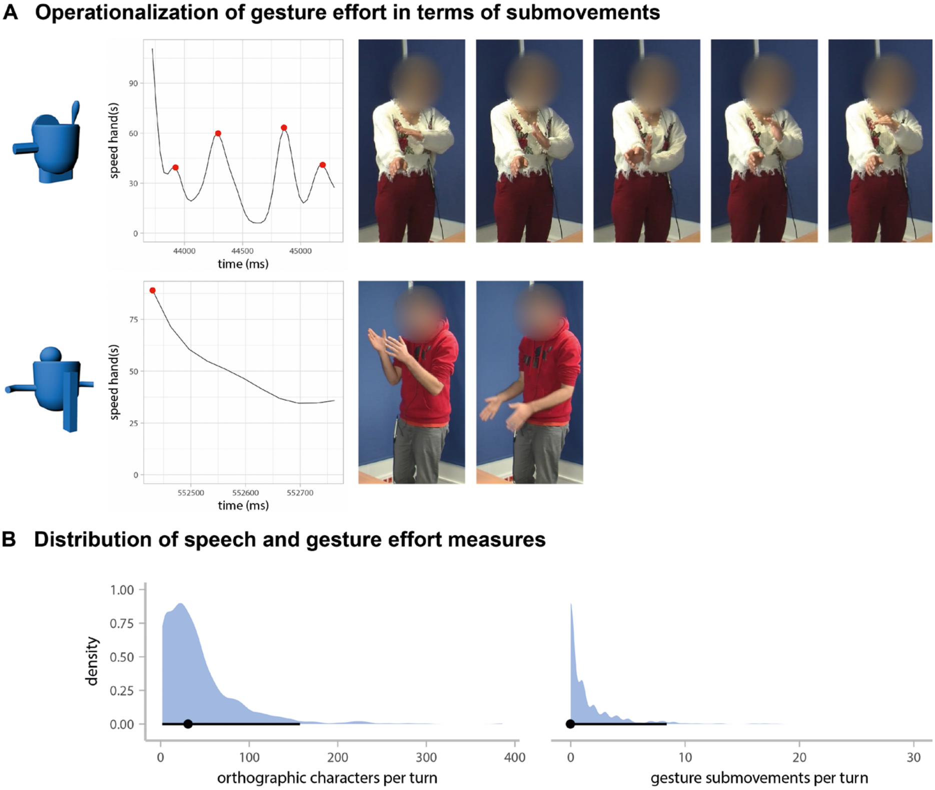


Figure 5.2. In panel (A), the top row shows a gesture which depicts the subpart on the left side of the Fribble. The gesture is produced by a matcher as part of a restricted offer with the following speech: “ah en is zijn arm uh rond maar ook een beetje met hoeken?” [literal translation: ah and is his arm round but also a bit with corners?]. The right arm is extended to model the “arm”, while the left hand is moved around it to depict the angular shape (number of submovements: 4). The bottom row shows a gesture which was produced by a director in a repair solution in response to a restricted offer. The gesture depicts the rectangular subpart on the front side of the Fribble, while saying “ja precies” [yes exactly]. The multimodal utterance thereby confirms the preceding restricted offer (which contained an identical gesture). The hands are kept somewhat apart (to depict the width of the rectangle), and moved straight downwards (number of submovements: 1). The density plots in panel (B) show the distributions for the speech and gesture effort measures (the dots are the median and the lines the 95% quantile interval).

In order to analyse the division of multimodal effort, we combined the speech and gesture efforts to yield a multimodal effort variable. We first standardised the individual speech and gesture measures (as their distributions differed greatly, with gesture submovements being zero-inflated, see Figure 5.2, panel B), and then summed them. We then calculated the proportion of multimodal effort in the repair initiation as compared to the total multimodal effort in the repair initiation and repair solution. To subsequently inspect how the total amount of joint effort varies across repair types, we summed the multimodal effort in the repair initiation and repair solution for each repair sequence. The resulting measures (i.e., the proportion of effort in the repair initiation and total joint effort) were used as dependent variables in mixed effects models with random intercepts and slopes for dyads and target items (unless reported otherwise, when a maximal model was not possible due to convergence issues) and repair initiator type as predictor. We used backward difference contrast coding to compare restricted requests to open requests, and restricted offers to restricted requests.

Results

Overall, 378 repair initiations were found in our dataset of task-based interactions from 20 dyads (comprising about 8 hours of audio, video and motion tracking recordings in total). We found a mean of 18.9 repair initiations per dyad (*SD* = 9.92, *range* = 6–45), which amounts to a repair initiation occurring once every 1.5 minutes on average. There were 24 open requests, 39 restricted requests and 315 restricted offers.

Speech and gesture effort

We start by reporting speech and gesture effort separately to allow for comparisons with prior unimodal work on the division of speech efforts in other-initiated repair (Dingemanse, Roberts, et al., 2015)[22]. The effort that people invest through the spoken modality to collaboratively resolve interactional trouble is shown in Figure 5.3, panel A. In repair initiations, speech efforts slightly increase as the type of initiation becomes more specific (open requests: *M* = 20.13, *SD* = 12.21; restricted requests: *M* = 27.05, *SD* = 16.63; restricted offers *M* = 34.38, *SD* = 20.18). In repair solutions, the opposite pattern emerges: when responding to more specific initiations, speech turns tend to become shorter (open requests: *M* = 116.63, *SD* = 87.18; restricted requests: *M* = 63.08, *SD* = 50.75; restricted offers *M* = 16.70, *SD* = 27.29). These findings are in line with the unimodal analyses in prior work (Dingemanse, Roberts, et al., 2015); see further notes on the division of verbal effort in the supplementary materials (Appendix B).

[22] Since we have no specific hypotheses for the distributions of unimodal effort, we refrain from using statistical tests (apart from a supplementary analysis to see whether the speech patterns replicate those of Dingemanse, Roberts et al. (2015), presented in Appendix B).

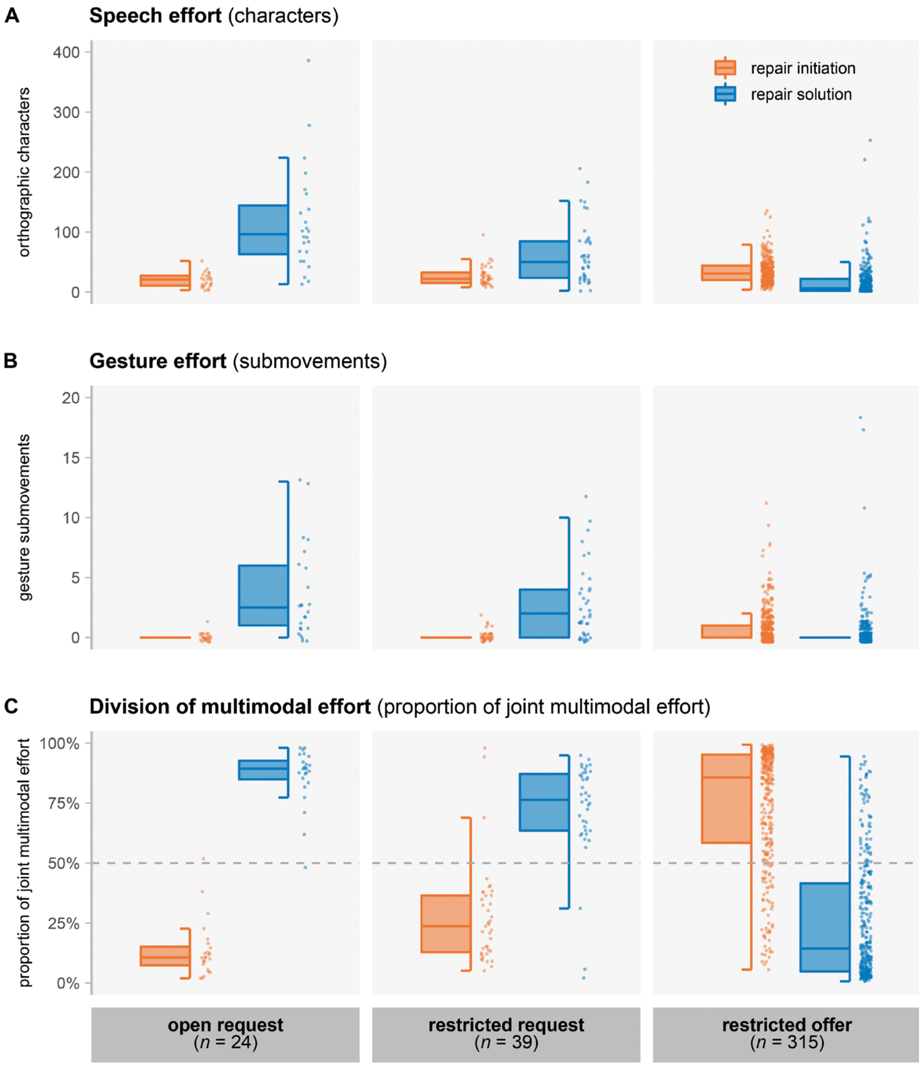


Figure 5.3. Effort invested in repair initiation (orange) and repair solution (blue), for repair formats of increasing specificity (open request < restricted request < restricted offer). Every dot represents a repair initiation or solution. Absolute speech effort (A) and absolute gesture effort (B) both go up in repair initiations and down in repair solutions as repair formats become more specific. Proportional multimodal effort (C) shifts from repair initiation to repair solution as we move towards more specific repair formats. The dashed line represents equal division of effort across participants.

For the gestural modality, there is considerable individual variation, with some people gesturing very rarely or not at all. In total, 479 co-speech gestures were produced across all repair initiations and solutions, with 37.2% of the turns containing at least one gesture. But the likelihood of encountering a gesture in a turn differs greatly across repair types and sequential positions; ranging from 4.2% in repair initiations of the type open request, to 79.2%

in repair solutions in response to open requests. When quantifying gesture effort in terms of submovements, we find a similar pattern as for speech effort (Figure 5.3, panel B). As repair initiations become more specific, more gesture submovements are used in the initiation (open requests: *M* = 0.04, *SD* = 0.20; restricted requests: *M* = 0.15, *SD* = 0.43; restricted offers *M* = 1.02, *SD* = 1.57), and fewer are used in the solution (open requests: *M* = 3.63, *SD* = 3.84; restricted requests: *M* = 2.67, *SD* = 3.15; restricted offers *M* = 0.53, *SD* = 1.74).

Division of multimodal effort

In accordance with the inherently multimodal nature of our dataset, we analyse how the total amount of multimodal effort is divided across participants (Figure 5.3, panel C). Multimodal effort is the sum of the effort invested through the two available modalities, so it can be speech-only or speech-plus-gesture (we found no cases of gesture-only). Adopting a narrow notion of multimodality (focusing on visual information in manual co-speech gestures only), we can consider 63% of the repair sequences as being multimodal in nature, containing one or more gestures by at least one of the participants.

We find that the proportional cost paid by the person initiating repair versus the person resolving the trouble varies as a function of the repair type, and deviates from a case of equal division (where each person would invest 50% of the total effort). The proportion of multimodal effort invested in the repair initiation was higher for restricted requests compared to open requests (β = 0.14, *SE* = 0.06, *t* = 2.35, *p* = .02), and higher for restricted offers compared to restricted requests (β = 0.47, *SE* = 0.04, *t* = 11.42, *p* < .001), as revealed by mixed effects models (with random intercepts for dyads). Overall, when considering multimodal effort counts (rather than proportions), we find a trade-off between the efforts invested by the two members of the dyad; the more multimodal effort is invested by the person initiating repair, the less multimodal effort is used to respond to it (*r* = -0.14, *p* = .007).

Minimisation of joint multimodal effort

The total amount of multimodal effort that was invested by the dyad to resolve interactional trouble (in the repair initiation and repair solution combined) is shown for each repair type in Figure 5.4. On average, we find that the joint multimodal effort is smallest when the repair initiation was a restricted offer. A mixed effect model (with random intercepts and slopes for dyads, and random intercepts for target items) revealed that joint effort is less in sequences involving restricted offers (*M* = 1.37, *SD* = 1.21) compared to restricted requests (*M* = 2.45, *SD* = 1.74; β = -1.15, *SE* = 0.30, *t* = -3.80, *p* = .006), but that restricted requests do not differ significantly from open requests (*M* = 3.51, *SD* = 2.58; β = -0.86, *SE* = 0.55, *t* = -1.58, *p* = .137). Paired with the finding that restricted offers are by far the most frequently used, people appear to do repair in the most cost-efficient way possible, minimizing the joint multimodal effort for the dyad as a whole.

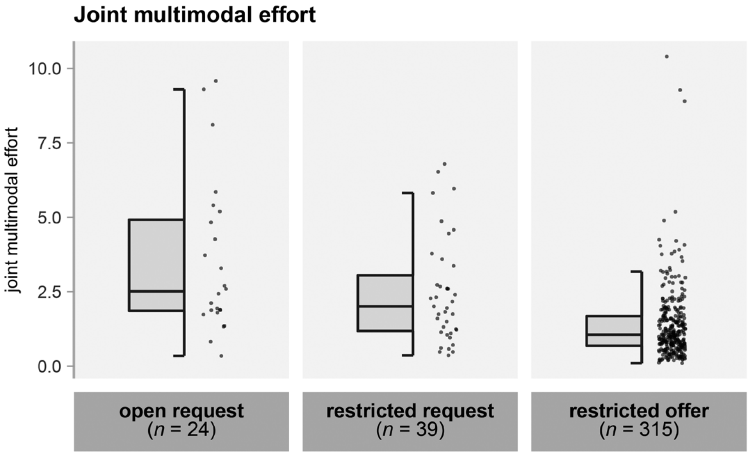


Figure 5.4. Joint amount of multimodal effort invested by both participants to resolve the interactional trouble. Every dot represents a repair sequence, i.e., repair initiation and repair solution together. As the specificity of repair formats goes up, joint multimodal effort invested goes down.

Discussion

The present study investigated how language use is shaped by communicative efficiency from a multimodal and interactional perspective. We focused on short time windows in turn-by-turn interaction where people work together to achieve a particular joint goal: repairing a problem with perceiving or understanding talk. We analysed how speech and co-speech gesture efforts are distributed across repair types (open requests, restricted requests and restricted offers) and sequential positions (the repair initiation and the repair solution), with the aim to test whether the division of multimodal effort is optimised for least collaborative effort.

There are three main findings. First, we find that speech and gesture efforts rise and fall together across repair types and sequential positions. This corroborates the view that speech and gesture are integral parts of a single multimodal communicative system (Kendon, 2014; Kita & Özyürek, 2003; McNeill, 1992; Perniss, 2018), and matches speech-gesture parallelism as reported for other interactional phenomena; for example, people increase both speech and gesture efforts in noisy environments (Trujillo et al., 2021), and people use fewer words and fewer gestures as common ground increases (for a review, see Holler & Bavelas, 2017).

Second, we show in detail how people orchestrate efforts in speech and gesture to achieve rapid coordination. In particular, the type of repair initiation used predicts how people divide their multimodal efforts: the more specific the repair initiation, the more multimodal effort is invested by the person initiating repair, leaving less work for the sender of the original message to resolve the trouble. This replicates prior unimodal work showing systematicity in how verbal effort is distributed across repair initiation and solution (Dingemanse, Roberts, et al., 2015), and shows that the pattern is robust enough to hold in both naturalistic as well as task-based conversations. Our results extend this division of labour principle to composite utterances, providing an unprecedented view of multimodal contributions to the coordination of joint action.

Third, we find that people overwhelmingly converge on repair formats (i.e., restricted offer) that minimise multimodal effort for the dyad as a social unit. This is a novel, direct attestation of the principle of least collaborative effort (Clark & Brennan, 1991; Clark & Schaefer, 1987; Clark & Wilkes-Gibbs, 1986) that is made possible by combining quantifications of speech effort with new, reproducible methods to measure gestural effort in terms of kinematics.

Taken together, these findings provide a novel unifying perspective on studies of language use (Clark & Brennan, 1991; Clark & Schaefer, 1987; Clark & Wilkes-Gibbs, 1986) and non-linguistic joint action (Santamaria & Rosenbaum, 2011; Török et al., 2019, 2021). The coordination of joint action minimally involves dynamically updated task representations, monitoring processes, and adjustable behaviours (Vesper et al., 2010). Although linguistic coordination has sometimes been cast as a qualitatively different form of coordination (Vesper et al., 2010), here we have shown that the micro-environment of interactive repair—which occurs at frequencies and timescales more commensurate with joint action—provides a unique window onto the real-time negotiation of distributed agency. In repair, people provide public evidence of representations and monitoring processes, allowing them to rapidly hone in on optimal coordinative solutions (Healey et al., 2007). By zooming in on these miniature coordination problems, we reveal systematicity in how people adjust multiple types of communicative behaviour to minimise joint efforts, thereby unravelling the multimodal nature of coefficiency in conversational joint action.

Beyond the empirical findings, our study also makes conceptual and methodological contributions. One is to extend and enrich standard notions of efficiency in language use. Prior work has usually considered efficiency in terms of unimodal message length (Gibson et al., 2019). One limitation of such operationalisations is that they easily lose sight of the fact that in conversation, the interactional work of achieving mutual understanding is often distributed across turns and participants (Colman & Healey, 2011; van Arkel et al., 2020). We argue that communicative effort and efficiency are best studied at the level of the dyad as a social unit, and we show how interactive repair provides a microcosm that allows us to study the public negotiation of mutual understanding over multiple turns.

Another challenge of the most common unimodal operationalisations is that, when applied to co-present conversational settings, they are incomplete and reductive, focusing on language as a unimodal discrete symbol system while overlooking multimodal, continuous and dynamic properties of language use (Bolinger, 1968; Pouw, Dingemanse, et al., 2021). Our contribution towards solving this challenge consists of using methods and insights from studies of joint action and behavioural dynamics (Meulenbroek et al., 2007; Sacheli et al., 2013; Vesper & Richardson, 2014). Our measures capture how people use both categorical and gradient semiotic resources in multiple modalities to make meaning together, where we operationalised effort with a) a linguistically informed quantification of speech in terms of orthographic characters (Dingemanse, Roberts, et al., 2015; Piantadosi et al., 2011) and b) a kinematic measure of submovements derived from continuous manual movement (Namboodiripad et al., 2016; Pouw, Dingemanse, et al., 2021; Trujillo et al., 2019; Vesper et al., 2021). These measures do not fully capture the multi-semiotic dimension of social interaction, as we disregard dynamic properties in the spoken modality (see e.g., research on the phonetic characteristics of repair solutions (Curl, 2005)) as well as non-manual embodied resources (further discussed below). However, we believe the combination of measures for the spoken and gestural modality used in the present study are a step in the direction of a truly multimodal linguistics that takes semiotic diversity seriously (Kendon, 2014).

Limitations

The nature of our task may have invited more representational gestures than some other conversational settings, as the 3D objects lack conventional names and lend themselves well to iconic depiction (Masson-Carro et al., 2016). The lab setting may also have influenced the relative amount of open requests (the least frequently attested repair format in our as well as others' task-based data (Dideriksen et al., 2019; Fusaroli et al., 2017)), as this is associated with noise interference and other low-level perceptual problems (Dingemanse, Roberts, et al., 2015). Despite these differences, the observed distribution of multimodal effort across repair types is likely to be robust enough to generalise to everyday language use. That is, while we might expect to find fewer manual gestures and more open requests, we would still predict people's multimodal productions to be more effortful in repair initiations of the type restricted offer compared to open and restricted requests (and vice versa for solutions).

We have investigated the cost-efficient use of words and manual co-speech gestures. By foregrounding efficiency in a task-based setting and focusing on speech and gestures, we have of course captured only a partial view of what it means for people to coordinate their multimodal utterances to resolve conversational problems. Future studies could broaden this view, for example by incorporating eye gaze, eyebrow movements, head movements and forward leans, which are known to play a role in signalling trouble (Andrews, 2014; Egbert, 1996; Floyd et al., 2016; Hömke, 2019; Li, 2014; Rasmussen, 2014; Seo & Koshik, 2010). This could be complemented by a consideration of other social or expressive factors which influence communicative behaviours, as people can for example opt to use an open request rather than a restricted offer for face-saving purposes (Kim, 1999). More empirical and theoretical work is needed to understand how pressures and constraints in human sociality interact with principles of efficiency in joint action. We take this to be an important avenue for future research, where moving our attention from efforts of the individual to those of the dyad as a cooperative social unit is an important first step.

Conclusion

In summary, in this study we investigated communicative efficiency from a multimodal and interactional perspective by zooming in on other-initiated repair sequences. As a conversational environment in which there is a clear goal, a limited set of turns to reach that goal and a limited inventory of communicative resources to use in those turns, other-initiated repair is a natural laboratory for the systematic study of coefficiency in language use. Our findings reveal that people divide the total amount of multimodal effort between them in such a way as to minimise the overall amount of speech and gesture efforts for the dyad as a whole. By investigating how communicative efficiency is realised in multimodal language use at the level of the dyad as social unit, we have shown how minimizing effort in language use is an interactional achievement.

Acknowledgements

This study came about in the context of a multidisciplinary research project within the Language in Interaction consortium, called Communicative Alignment in Brain and Behaviour (CABB). The work has been shaped by contributions of CABB team members, with special thanks to Sara Bögels, Lotte Eijk, Judith Holler and Ivan Toni. We would additionally like to thank Emma Berensen and Junfei Hu for their help with annotating the data, and James Trujillo and Maarten van den Heuvel for their help with the collection and synchronisation of the motion tracking data. The authors remain solely responsible for the contents of the paper. This work was supported by the Netherlands Organisation for Scientific Research (NWO) Gravitation Grant [024.001.006] to the Language in Interaction consortium and Vidi Grant [016.vidi.185.205] to M.D.

Chapter 6

General discussion

General discussion

How do people negotiate mutual understanding in co-present social interaction? In this thesis, I studied this question from a multimodal and interactional perspective, in the context of collaborative referring to novel referents. Specifically, I focused on the use of speech and co-speech gesture in alignment and other-initiated repair. With this overarching question and research foci in mind, I set three objectives: (i) situate alignment in multimodal, sequential talk-in-interaction—both conceptually and empirically, (ii) investigate the other-initiated repair system from a multimodal perspective, and (iii) examine the division of multimodal labour between participants in social interaction from a joint action perspective.

Objective (i) called for the development of an integrative understanding of alignment, as there appeared to be a lack of consensus in the multidisciplinary research area of social interaction. Therefore, in chapter 2, I set out to clarify how alignment is theorised to support mutual understanding (e.g., through priming or grounding), where researchers look when they study alignment (in the minds or behaviours of people), and how they study alignment (with various operationalisations). I aspired to develop conceptual tools that can help researchers working on alignment (and mimicry, repetition, behaviour matching, etc.) to conceptualise and empirically study cross-participant behavioural alignment in its natural but complex environment: multimodal talk-in-interaction.

In chapter 3, I empirically investigated the multimodal and sequential realisation of alignment in relatively free-flowing task-based interactions about novel referents, focusing on lexical choice and co-speech gesture. This interactional approach complements the paradigms with which alignment is commonly studied in psycholinguistics, such as experiments with highly-constrained labelling tasks (e.g., Branigan et al., 2000, 2007; Cai et al., 2021) and corpus studies focussing on the temporal distance and form overlap of aligned words or gestures, without taking their sequential relation into account (e.g., Bergmann & Kopp, 2012; Dideriksen et al., 2019; Oben & Brône, 2016; Reitter & Moore, 2014). Studying lexical and gestural alignment in the context of collaborative reference to novel referents also brings into view a connection to research on symbol creation, where alignment and modality have been singled out as key factors (see e.g., Fay et al., 2018; Lister & Fay, 2017; Macuch Silva et al., 2020), but where the role of modality in alignment has not yet been studied.

To meet **objective (ii)**, I investigated how people use co-speech gestures together with speech to target and resolve interactional trouble in chapter 4. I built on the foundational speech-centred work on other-initiated repair in conversation analysis (Schegloff, 2000; Schegloff et al., 1977), and more recent investigations on the role of embodied signals, such as the use of facial expressions to signal trouble in repair initiations (e.g., Hömke, 2019; Seo & Koshik, 2010) and the use of co-speech gestures to clarify trouble in repair solutions (e.g., Alibali et al., 2013; Hoetjes, Krahmer, et al., 2015; Holler & Wilkin, 2011b). In this thesis I took the next step: I aimed to work towards a more holistic understanding of other-initiated repair as a multimodal system. I did so by quantitatively and qualitatively investigating how people carry out interactional work with speech and co-speech gesture in both repair initiations and solutions.

Finally, to attain **objective (iii),** I set out to characterise the collaborative, multimodal nature of social interaction in chapter 5. I used other-initiated repair as a testing ground where communicative efforts can be quantified in a systematic manner. In particular, I investigated how people divide their speech and gesture efforts across repair initiations and repair solutions, and tested whether the division of multimodal efforts adheres to the principle of *least collaborative effort* or *coefficiency*, which has previously been attested in spoken dialogue (Clark & Brennan, 1991; Clark & Schaefer, 1987; Clark & Wilkes-Gibbs, 1986) and non-linguistic joint actions (Santamaria & Rosenbaum, 2011; Török et al., 2019, 2021).

This chapter summarises the main findings of chapters 2-5, followed by a discussion of their broader implications—both theoretical and methodological. I then suggest directions for future research, and end with concluding remarks.

Summary of main findings

In **chapter 2**, I started with a quest to synthesise the literature on alignment, in an effort to enable principled comparison of the diverse range of theoretical and empirical approaches. At the heart of this chapter lies the proposal that the relation between any two instances of communicative behaviour can be characterised in terms of five key dimensions: *time, sequence, meaning, form,* and *modality*. Equipped with this integrative framework, I reviewed empirical studies on lexical and gestural alignment, demonstrating how methodological grouping criteria and measurement variables can be captured in terms of the five dimensions. The framework also helped to uncover how empirical approaches pattern together with theoretical presuppositions, revealing blind spots in the research on and our understanding of alignment. Specifically, the form and time dimensions have been prioritised in experimental studies, while the dimensions of modality and sequence remain largely unexplored; studying these latter dimensions can advance our understanding of how behavioural alignment is realised and used in co-present talk-in-interaction.

It is this final conclusion of chapter 2, i.e., the importance of further investigating the modality and sequence dimensions of alignment, that fed into chapter 3. Specifically, the relation between lexical and gestural alignment was highlighted as a promising and unexplored avenue for future research. I expected that by virtue of the semiotic properties of words and gestures, lexical and gestural alignment can offer different affordances for the joint negotiation of meaning, calling for a careful operationalisation and investigation of the dimensions modality and sequence. In the context of referring to novel objects or concepts (chapter 3), this could have implications for the order in which alignment comes about. Alignment of iconic co-speech gestures might initially be used to establish mutual understanding (making use of form-meaning mappings), which could precede alignment in terms of lexical choice when referring to novel referents. To test these hypotheses the five-dimensional framework proposed in chapter 2 was used in chapter 3 to operationalise alignment in multimodal talk-in-interaction in a systematic and transparent manner.

The specific aim of **chapter 3** was to investigate how frequently, when and how lexical and gestural alignment emerges in the process of converging on shared symbols for novel referents. To this end I used a referential communication task with novel objects called "Fribbles" (the main dataset of this thesis, of which I analysed a subset in this chapter), along with an individual naming task in which participants labelled the Fribbles. I expected that people would establish shared labels for the novel referents as indicated by the naming task (prediction 1), and that people would be more likely to align lexically or multimodally in the interaction, compared to only aligning in gesture (prediction 2). In case of multimodal alignment, I expected that alignment would emerge in both modalities simultaneously, or in the gestural modality first (later followed by the lexical modality; prediction 3). Furthermore, I qualitatively investigated the sequential environments in which alignment emerges to understand how lexical and gestural alignment are (independently, simultaneously or successively) used as interactional resources to effectively refer to novel referents.

The results of chapter 3 revealed that, first of all, symbol creation was largely successful: people used more similar names for the Fribbles after compared to before the interaction (prediction 1 supported). Quantitative analysis of the interaction revealed that multimodal and lexical alignment were more frequent than gestural alignment (prediction 2 supported). I found a distinctive pattern for multimodal alignment: it was more frequent than gestural alignment and tended to emerge earlier in the interaction compared to both lexical and gestural alignment. Emergence of alignment in both modalities simultaneously was more frequent than successive emergence over multiple turns or rounds of the interaction (i.e., lexical alignment preceding gestural alignment or vice versa). Contrary to our expectations however, the two types of successive emergence (gestural alignment preceding alignment, and lexical alignment preceding gestural alignment) were equally frequent, providing mixed support for prediction 3. The qualitative analyses further showed how people flexibly deploy lexical and gestural resources in line with modality affordances and communicative needs, in the service of establishing and calibrating joint reference to novel objects. Most notably, people can rely on gestural alignment to compensate for problems with (the absence of) lexical pacts, and use partial alignment in gesture form to negotiate mutual understanding of a referent.

Having investigated how people use speech and gesture in alignment to establish joint reference, I turn to the role of modality in other-initiated repair in **chapter 4**. The aim of this study was to contribute towards a comprehensive understanding of other-initiated repair as a multimodal system, by investigating how manual co-speech gestures can be used together with speech for targeting and resolving interactional trouble. To this end, I investigated multimodal strategies in initiating and responding positions of different types of repair, this time using the complete dataset of task-based interactions. I found a distinctive distribution of co-speech gestures across repair turns; the number of gestures per turn go up in repair initiations and down in repair solutions as repair initiations become more specific (open request < restricted request < restricted offer). In that respect, speech and gesture function similarly; people use both modalities to make repair initiations and solutions more specific, and they can be combined to form effective multimodal strategies. I presented illustrative examples showing that, in repair initiations, gestures can be used together with speech to localise trouble (through repetition of the multimodal trouble-source) and/or to *offer* solutions (through iconic depiction). In repair solutions multimodal strategies are mostly used to *clarify* trouble (e.g., by mapping gestures onto better recipient-designed verbal expressions), but gestures can also be used to effectively *confirm* multimodal offers (through acknowledgment gestures or gestural repetition). These results are in line with various isolated findings on co-speech gestures in the other-initiated repair literature. Here I put the puzzle pieces together for the first time, and showed that manual co-speech gestures can be wielded together with speech to realise a variety of repair formats. The findings paint a picture of other-initiated repair as a flexible, semiotically-diverse system for resolving problems with perceiving or understanding talk-in-interaction, highlighting the importance of studying mutual understanding in multimodal social interaction.

Building on the insights on multimodal repair strategies from chapter 4, **chapter 5** uses other-initiated repair as a testing ground to study efficiency in collaborative, multimodal interaction as joint action. Chapter 4 already showed that co-speech gestures play a prominent role in other-initiated repair, where the usage and number of gestures in a particular turn is contingent on the partner's turn, showing resemblances with earlier work on speech. The goal of chapter 5 is to investigate how people distribute multimodal effort across repair initiations and repair solutions. For this study, I used the same dataset and annotations (for speech, gesture and repair) as in chapter 4, but completed it with the motion tracking data. I quantified speech effort as the number of orthographic characters per turn (cf. Dingemanse, Roberts, et al., 2015) and gesture effort as the number of gesture submovements per turn (cf. Trujillo et al., 2018, 2021). The results are in line with patterns previously reported for unimodal interactions: the more specific the repair initiation, the more multimodal effort is invested by the person initiating repair (relative to the person resolving the trouble). The findings also provide a novel, direct attestation of the principle of *least collaborative effort* in multimodal interaction, in line with findings of earlier speech-centred studies (Clark & Brennan, 1991; Clark & Schaefer, 1987; Clark & Wilkes-Gibbs, 1986; Dingemanse, Roberts, et al., 2015) and studies on non-linguistic joint actions (Santamaria & Rosenbaum, 2011; Török et al., 2019, 2021). That is, rather than minimizing individual efforts, people structure their multimodal utterances such as to yield least collaborative effort for the dyad as a whole.

Theoretical implications

The starting point of this thesis on mutual understanding was the premise that social interaction is collaborative and multimodal. This led me to focus on the joint, multimodal work that people do in talk-in-interaction. What have we learned about the process of establishing mutual understanding by adopting this view?

To put the answer to this question in perspective, it is useful to recognise that this thesis is written at a time where an increased research focus on social interactions starts to transform our understanding of cognition and language use. This shift has initially been recognised in the field of social cognition and dubbed "the interactive turn" (De Jaegher et al., 2010). Later the phrase "the interactive turn" has also been used (e.g., by Hömke, 2019; Kendrick, 2017) to describe the recognition of the importance of social interaction for psycholinguistic theories (e.g., Levinson, 2016; Pickering & Garrod, 2004). Yet the interactive turn was originally coined by De Jaegher et al. (2010) along with the proposal to study interaction dynamics as interactive explanations in their own right, which can complement or even replace individual mechanisms. However, much psycholinguistic work still focuses on the individual cognitive mechanisms involved in lexical, syntactic and semantic processing (of speech or multimodal utterances), where dialogue is construed as the sum of two such systems (i.e., Aggregate approaches to dialogue; Healey et al., 2018). Conversely, when adopting an interactional approach, our attention shifts to the sequential organisation and joint coordination principles involved in dialogue, raising new questions about underlying cognitive processes and representations (Sebanz et al., 2006; Vesper et al., 2010), and how those might be distributed over multiple minds (Hutchins, 1995).

In this thesis, I set out to study mutual understanding by focusing on the joint work that people do in talk-in-interaction as it unfolds in real time. By adopting this interactional approach to study social interaction, the findings of the current thesis offer new insights, which can be categorised as "perspective shifts" in three domains:

(a) alignment as interactional resource;
(b) gesture as coordination device;
(c) social interaction as efficient joint action.

These domains tie in with the objectives set out at the start of this thesis:

(a) → objective (i): to situate alignment in multimodal, sequential talk-in-interaction—both conceptually and empirically
(b) → objective (i) and (ii): to investigate the other-initiated repair system from a multimodal perspective;
(c) → objective (iii): to examine the division of multimodal labour between participants in social interaction from a joint action perspective.

Alignment as interactional resource

Alignment has been argued to support mutual understanding, which is reinforced by studies reporting a positive relation between alignment and performance measures in cooperative tasks (Dideriksen et al., 2020; Fay et al., 2018; Fusaroli & Tylén, 2016; Reitter & Moore, 2014). But how does alignment come about, and how does it lead to mutual understanding? As discussed in chapter 2, theories in psycholinguistics and neuroscience hold that cross-participant alignment of behaviour comes about through priming or direct mapping mechanisms in the individuals' minds or brains (Brass & Heyes, 2005; Dijksterhuis & Bargh, 2001; Heyes, 2011; Pickering & Garrod, 2004; Rizzolatti et al., 2001). These internal processes have been argued to be the driving force for establishing alignment at higher levels of representation (Pickering & Garrod, 2004), which we can consider to be mutual understanding.

But what if instead of focusing on individual cognitive mechanisms (an aggregate approach to the study of social interaction), we look at how alignment is embedded in talk-in-interaction (an interactional approach)?[23] In chapter 2 I contrasted *priming* with *grounding* approaches to alignment, and showed that the notion of *grounding* (a process that is relevant in interactional approaches) pertains to the joint efforts involved in "assuring that what is said has been heard and understood before the conversation goes on" (Clark & Schaefer, 1987, p. 19). In the domain of referential communication, this warrants a close look at the sequential relation between contributions, whether people are referring to the same or different referents, and what the referential expressions look like. As such, an interactional approach to alignment puts into focus how instances of behaviour are related in terms of sequence, meaning, and form, where more is to be gained by looking at the dimension of modality. By studying how alignment is embedded in sequential, multimodal talk-in-interaction, we can investigate what people do with alignment, and how it can be used as an interactional resource to establish mutual understanding.

[23] Note that I do not consider these two research foci to be mutually exclusive explanations for alignment. See the section "Theoretical approaches to alignment" in chapter 2, and more recent work by Pickering and Garrod linking alignment of situation model representations to interactional processes (Gandolfi et al., in press; Pickering & Garrod, 2021).

By adopting this approach in chapters 3 and 4, I contended that not all alignment is the same from an interactional point of view. I showed how alignment can be used for various interactional purposes, which yields variation in how instances of behaviour are related in terms of sequence (adjacent turns or spanning multiple turns), form (slight variation of what is repeated or exact copy), and modality (alignment in speech, gesture or both). In chapter 3, I showed that people use lexical and/or gestural alignment in various ways to establish joint reference, for example by negotiating a referential expression through multimodal expansion, adding adjectives in speech and changing form features in gesture. In chapter 4 on other-initiated repair, I showed examples of gestural alignment with minimal form deviance, where the "copying" of gestures was used for the purpose of localising trouble (in a turn which was produced much earlier) or confirming the offer of a candidate understanding. Together, chapters 3 and 4 are concrete cases of how the integrative framework developed in chapter 2 helps guide, clarify and specify findings that otherwise would have been easily obscured by overly broad definitions of alignment.

The insights from chapter 2-4 about alignment as an interactional, multimodal resource are of broad relevance for research themes in linguistics and beyond. Consider research on language emergence,[24] which involves the study of mechanisms involved in co-creating novel labels. Prior work has already stressed the advantages of the gestural modality for language emergence (e.g., Fay et al., 2013, 2014; Levinson & Holler, 2014; Sterelny, 2012), and has provided evidence for a causal link between alignment (of drawings) and effective symbol creation (Fay et al., 2018). This thesis has contributed to our understanding of the interactional processes that are likely to be involved in language emergence, by showing how simultaneous and successive alignment in multiple modalities shapes the symbol creation process. For example, chapter 3 showed that multimodal alignment can occur in adjacent turns to establish mutual understanding of the referent, while lexical alignment occurring across larger spans can be used to commit to one of multiple conceptualisations that had been coined earlier (establishing a "conceptual pact"; Brennan & Clark, 1996). Chapter 4 shows that people can use lexical and/or gestural alignment to address and resolve problems with hearing or understanding—problems that inevitably come up in language emergence settings, but which has not received much scholarly attention (but see Macuch Silva & Roberts, 2016; Safar, 2021).

Another research domain that could benefit from the new insights on alignment is language development. Alignment has been studied in child-adult conversations (e.g., Dale & Spivey, 2006; Garrod & Clark, 1993; Laalo & Argus, 2020; Misiek et al., 2020), and different forms of alignment (lexical, syntactic or semantic) have been shown to positively

[24] The link to language emergence was apparent to the participants as well. In the debriefing questionnaire, one of them remarked: "Het was leuk om te zien dat door samenwerking en herhaling je samen tot een eigen taaltje kan komen." [It was fun to see that through collaboration and repetition you can jointly arrive at your own (little) language]

relate to subsequent measures of the child's language development (e.g., Denby & Yurovsky, 2019; Foushee et al., 2021; Fusaroli et al., 2021). A key research question here is how different types of alignment can scaffold different aspects of language learning, pertaining to the lexicon or grammar. The framework developed in chapter 2 can help carefully operationalise different types of alignment in terms of time, sequence, meaning, form, and modality, and can enable commensurability of findings across different studies. But another way in which the work reported here comes in, is through the emphasis on alignment as embedded in talk-in-interaction—sequences of talk that are aimed at establishing mutual understanding and doing things together. Understanding how alignment shapes language development involves understanding how it is used to realise interaction work. As Yorovsky et al. (2016) put it: "They [caregivers, red.] need not be teachers; they need only be communicators. If parents want to communicate with their children, and their children need significant linguistic support, they will have no choice but to align" (p. 2097). It is here that we benefit from an interactional approach, paired with an integrative understanding of the building blocks of alignment.

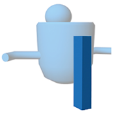
There are more research domains for which the framework could be of relevance, but I have singled out language emergence and development, as there has been a spike of interest in alignment in these fields. I argue that these fields could benefit from a closer look at alignment from an interactional and multimodal point of view akin to the one adopted in this thesis. Take experimental approaches to the study of language emergence for example, where comparisons of signals emerging in "interactive conditions" and "transmission conditions" led to the conclusion that interactive processes are a driving force behind the gradual emergence of efficiency in communication systems (e.g., Garrod et al., 2010; Kirby et al., 2015; Motamedi et al., 2019). To further unravel the exact interactional mechanisms that are responsible for these effects, interactions have been experimentally controlled, for example allowing or restraining participants from copying or modifying each other's drawings in a referential task (Fay et al., 2018; Healey et al., 2007). Yet, as we have seen in this thesis on referential communication in a natural spoken language, not all alignment is alike, and (partial) alignment in different modalities can be used in different ways to establish joint reference. So, I believe we can make progress by complementing experimental approaches with more fine-grained analyses of the interactive work carried out with alignment in spontaneous co-present interactions—involving adults when they do not have access to conventionalised labels, or children who have not yet acquired them. It is there that the insights from this thesis and the integrative framework on alignment can help us understand the interactive processes involved.

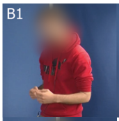
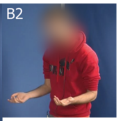
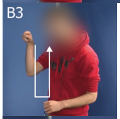
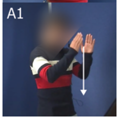
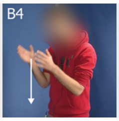
Gesture as coordination device

Let's take a closer look at the kind of gestures that were analysed in chapters 3 and 4: gestures which convey semantic meaning through iconicity, and pragmatic meaning through their relation to a (multimodal) utterance of another speaker (Holler, 2022). There are as of yet no comprehensive accounts on how such gestures are produced, because most of the theoretical work on gesture production (especially representational gestures) aims to characterise the origin of gestures and how gesture and speech production processes are related *within* individuals (e.g., de Ruiter, 2007; Kita & Özyürek, 2003; McNeill, 1992) while foregoing investigations of the coordination of speech and gesture *between* people. The emphasis on intra-speaker coordination applies to unimodal speech production models as well (Levelt, 1989), though some linguistic theories, such as Dynamic Syntax, do account for (intra-individual) turn contingencies (see e.g., Cann et al., 2005; Ginzburg & Cooper, 2004; Purver et al., 2011). However, in what follows I will argue why certain types of "collaborative gestures" (Furuyama, 2002) or "return gestures" (de Fornel, 1992) present special challenges for language production models focused on individual language producers.

Consider again Transcript 4.5 from chapter 4, where a multimodal candidate understanding (line 2), was confirmed through gestural repetition (line 4).

Transcript 4.5 (reproduced from chapter 4)



1	TROUBLE SOURCE	B (director):	oké en dan dan heb je dus aan de voorkant (een) B1 daar zie ik er B2 ook niet heel veel van een zo'n groot blok B3 *okay and then then you thus have on the front (a) B1, I don't see B2 so many of those, a such a big block B3*	  
2	REPAIR INITIATION	A:	gewoon die zo recht naar beneden A1 gaat? *just that goes straight down A1 (like this)?*	
3		B:	zie je re- *do you see ri-*	
4	REPAIR SOLUTION	B:	ja precies B4 *yes exactly B4*	

B1: two-handed gesture depicting the relative location of the highlighted Fribble subpart; the hands are loosely held in front of his torso, and then moved towards himself.

B2: two-handed "palm-up open-hand" gesture (cf. Cooperrider et al., 2018; Müller, 2017).

B3: right-handed gesture drawing the shape of the highlighted Fribble subpart with the index finger.

A1: two-handed gesture depicting the shape and orientation of the highlighted Fribble subpart; the hands are slightly apart and make a single downward motion.

B4: two-handed gesture depicting the highlighted Fribble subpart; almost identical to A1.

Of main concern is line 4. Here, gesture B4 expresses information about the shape and orientation of the referent (the highlighted Fribble subpart), while this information is not conveyed through the co-occurring speech ("yes exactly"). The only way we can make sense of this utterance, is by virtue of its connection to the prior turn. As de Fornel (1992) puts it: "it is possible to observe in the gestural shape some visual aspects of content that are not linked to the verbal component of the utterance of the speaker but to the verbal and gestural component of the utterance of another speaker" (p. 175). Furuyama (2002) argues that such inter-personal coordination of speech and gesture (which happens between mechanically "disconnected" bodies) poses challenges for intra-personal speech-gesture production models, because "these theories typically, or almost always, require or at least assume all of the sub-systems or sub-components of the whole production system of speech and gesture to be mechanically connected" (p. 349).

The above example can help us to unpack this further. According to gesture production models, overt gestures are generated from an abstract representation (called "imagistic thinking" or spatio-motoric imagery; McNeill, 1992; Kita & Özyürek, 2003) which is argued to involve a complex process of selecting information and assigning a perspective (de Ruiter, 2007). This is why, generally speaking, iconic gestures are idiosyncratic and variable in form (though not entirely, see e.g., Ortega & Özyürek, 2020). In this example however, it appears that gesture B4 is a deliberate reproduction of A1. Note that participant B's initial depiction of the referent (gesture B3) was different both in form (one-handed instead of two-handed, tracing the outline with the index finger), and in the specific visual properties it emphasised through the speech-gesture combination (shape and size rather than orientation). Gesture B4 is used to realise the interactional work of *confirming* the restricted offer, which it does by copying the form of the partner's gesture (A1), rather than by repeating his own prior gesture (B3) or establishing iconic mappings afresh (Holler, 2022).

The main question this raises is: how is gesture B4 produced? Does it still involve the generation of an action plan on the basis of personal action schemata or spatial imagery (Kita & Özyürek, 2003)? Could this mechanism be modulated by the perception of the partners gesture? Or is action generation bypassed, such that the perception of the partners gesture can be directly mapped onto (or "prime") the execution of one's own gesture (see e.g., Brass & Heyes, 2005; Dijksterhuis & Bargh, 2001; Heyes, 2011; Rizzolatti et al., 2001)? And what about those cases of gestural alignment where there is overlap in some form features, but variation in others (as we saw in chapter 3)? Note that these questions do not apply to the same extent for coordination of speech trough lexical repetition, because people already have (lexical, semantic and phonological) representations of the words available, at least, when they both are speakers of the same language.

As such, we would benefit from a reconsideration and perhaps expansion of theoretical gesture production models, as well as more empirical work on how people coordinate their use of speech and gesture. To commensurate the findings of such studies, the integrative framework of chapter 2 can be of help, as the dimensions allow for a precise description of the (cross-modal) relation between behaviours of different interactants.

Social interaction as joint action

Researchers working on joint action ask the question how people are able to act together to achieve joint goals, which involves "predicting what others are going to do next, adjusting one's behaviour to complement another's task, and achieving precise temporal coordination" (Vesper et al., 2010, p. 998). This has mostly been studied in terms of the short-term adaptations and predictions that underly the coordination of bodily movements, while linguistic coordination has been cast as qualitatively different, being more suited to situations involving long-term planning or when people cannot directly perceive each other's behaviours (Vesper et al., 2010). Meanwhile, in (psycho)linguistic, language use and linguistic structure are often studied from the perspective of individual language users producing and comprehending linguistic material. But what if both fields would start to consider language use in social interaction as a principle form of joint action? This thesis provides some pointers.

Specifically, in chapter 5 I asked how people coordinate their multimodal turns to address and resolve trouble with hearing and understanding. I found that they adhere to a joint action principle of *coefficiency* (Santamaria & Rosenbaum, 2011; Török et al., 2019, 2021), or as it has been introduced with regard to language use, the principle of *least collaborative effort* (Clark & Brennan, 1991; Clark & Schaefer, 1987; Clark & Wilkes-Gibbs, 1986). That is, rather than minimizing individual efforts, people share the multimodal workload in predictable ways, opting for the use of restricted offers where possible to minimise the overall multimodal effort invested by the dyad as a social unit. As such, I have shown that the micro-environment of interactive repair—which occurs at frequencies and timescales commensurate with "typical" cases of non-linguistic joint action—provides a unique window onto the real-time negotiation of distributed agency.

If we turn the question around, we can ask how joint action perspectives (and in particular the notion of *coefficiency*) might inform (psycho)linguistic theory. In linguistic research, production and comprehension effort is assessed on an individual level, for example by computing word or dependency lengths (Gibson et al., 2019; Levshina & Moran, 2021), or reaction times or brain activity patterns in experiments (see e.g., Jaeger & Tily, 2011; Kutas et al., 2006). Though scholars have considered how processing efforts of the addressee can be prioritised over those of the speaker (see e.g., Trott & Bergen, 2022; or Rasenberg, Rommers, et al., 2020 on how intersubjective discourse markers affect comprehension efforts), the question if and how people design their turns in order to minimise *joint* efforts has not been asked yet. Furthermore, psycholinguistic research on efficiency is mostly concerned with processing efforts of isolated utterances, overlooking how interactional strategies can alleviate the computational or articulatory efforts involved in production or comprehension (Lerner, 1991; van Arkel et al., 2020). This presents challenges for theories of language efficiency, as well as speech and gesture production models, as discussed in the previous section.

Finally, to arrive at a comprehensive understanding of social interaction as efficient joint actions, we need to consider efforts across different types of (communicative) behaviours which may be used or combined in social interactions. Chapter 5 showed that speech and gesture efforts rise and fall together across repair types and sequential positions, in line with other studies showing parallels in speech and gesture efforts (Holler & Bavelas, 2017; Trujillo et al., 2021). One way to think about this is that speech and gesture contributions may be similar in terms of information density. For example, in open requests, there is generally little semantic content expressed in speech (e.g., just "What?" or "Sorry, come again?") and in gesture (manual gestures are rarely used in such turns). But there are exceptions to this pattern, where gestures carry more of the communicative burden to convey specific information. For example, in chapter 4 I showed that in restricted offers speech can merely serve an indexing function ("like this?"), drawing attention to an iconic gesture which depicts a candidate understanding (cf. Jokipohja & Lilja, 2022 who show that speech can also be absent all together). And in repair solutions in response to restricted offers, a candidate understanding can be confirmed by means of gestural repetition (which iconically depicts the referent), along with a short verbal response ("yes exactly"). While in chapter 5 we have assessed the division of labour *across people,* these examples show that we would also benefit from an investigation of the intra-personal division of effort or information density *across modalities* (see e.g., de Ruiter et al., 2012) and how this impacts coordinated language use in joint actions.

In sum, I believe we would benefit from softening the conceptual boundary between non-linguistic and linguistic interactions, by considering how both are primary cases of joint action. This also brings to the forefront the recognition that many joint actions in our everyday lives involve the coordination of language *and* other actions. Take the beloved moving-a-couch example in the joint action literature; rather than artificially keeping language use out of the picture, we need a theory that can account for the simultaneous coordination of lifting and exchange of communicative signals such as "higher", "lower" or head movements to indicate directions. Similarly, research on language use (this thesis included) will eventually need to incorporate how language use interplays with other actions. For example, when an addressee is engaged in a parallel activity, or when interactants are engaged in an activity such as cooking together, this can have consequences for how repair is initiated (Blythe, 2015; Jokipohja & Lilja, 2022; Rossi, 2015). That is, "a theory of action must come to terms with both the details of language use and the way in which the social, cultural, material and sequential structure of the environment where action occurs figure into its organization" (C. Goodwin, 2000, p. 1489).

Methodological contributions and limitations

Combining data sources and disciplines

I will discuss the methodological contributions of the three empirical chapters (3-5), focusing on the combination of data sources and the interdisciplinary character of the research.

For chapters 3-5, I collected a dataset of task-based interactions. To this end, 20 dyads (of unacquainted participants) came to the lab, where they performed a referential communication task in a face-to-face setting. Head-mounted microphones and three cameras were used to collect high-quality audio and video recordings. Furthermore, I collected motion tracking data using Kinect V2 devices. In this thesis, I have transcribed the speech and coded co-speech gestures manually. Subsequently, the motion tracking data has been used to analyse kinematic features of the (manually annotated) gesture strokes. In line with the team science environment in which this research was carried out, I benefitted from the help of others, most notably Sara Bögels (stimuli and task development), Lotte Eijk (audio-video recording set-up), James Trujillo (Kinect data collection), Maarten van der Heuvel (Kinect-video synchronisation), Wim Pouw (Kinect data analysis), Emma Berensen (speech transcription, gesture coding), and the technical staff of the Max Planck Institute for Psycholinguistics and Donders Centre for Cognitive Neuroimaging.

Manually coding gestures is very time-consuming, and the field is moving towards more automated techniques in order to analyse larger datasets (e.g., Beugher et al., 2018; Ripperda et al., 2020). The manual annotations created for this thesis project were also used to develop and test an automatic movement detection algorithm which identifies potential iconic gestures (as described in detail in Pouw, de Wit, et al., 2021). Though this gesture detector cannot fully replace human coding, it can be used as a tool to reveal meaningful kinematic patterns of (auto-coded) gestures when applied to a larger dataset.[25]

In this thesis, I have analysed speech transcriptions, gesture annotations and motion tracking data. Gesture studies traditionally involve human coders who make decisions about the meaning and functions of bodily movements. Conversely, in behavioural sciences and joint action studies, automated techniques are used to capture the precise spatial and temporal dynamics of bodily movements. In this thesis I have combined these two approaches, constituting a more holistic investigation of language use, where I have

[25] Proof of concept is provided in the manuscript presenting the *CABB dataset* (Eijk et al., 2022). This dataset contains audio, video and motion tracking (Kinect) recordings from a (more extended version of a) referential communication task, along with behavioural and neuroimaging data collected before and after the referential task (for a total of N = 71 dyads). This dataset has been made available to the scientific community; all data and code necessary for automatically detecting gestures and analysing kinematic features is shared (courtesy of Wim Pouw).

considered how people use both categorical and gradient features of semiotic resources in multiple modalities to make meaning together. Specifically, in chapter 5, I operationalised multimodal effort with a) a linguistically informed quantification of speech in terms of orthographic characters (Dingemanse, Roberts, et al., 2015; Piantadosi et al., 2011) and b) a kinematic measure of submovements derived from continuous manual movement (Namboodiripad et al., 2016; Pouw, Dingemanse, et al., 2021; Trujillo et al., 2019; Vesper et al., 2021). This approach represents a novel methodological contribution towards a truly multimodal linguistics that takes semiotic diversity seriously (Kendon, 2014).

Considering the multimodal, continuous and dynamic properties of language use also comes with challenges, for example if we wish to study if people align their behaviour. How can we judge whether people use the "same" behaviour if it is not part of a discrete symbolic system? As discussed in chapter 2, prior work has often focused on form as the crucial dimension on which behaviour is compared; if two gestures look alike, they are considered aligned. But form overlap is gradient, and in chapter 3 (Appendix B) I presented a detailed analysis of form overlap in gestures which referred to the same referent (i.e., that are related in meaning and produced in the same modality, but without imposing restrictions on time or sequence). This analysis revealed that partial overlap is the norm, while complete form overlap occurs for less than 5% of referentially aligned gestures. For this analysis I used binary scoring on discrete form properties (handedness, handshape, movement, orientation and position). But in a different study with the same dataset, we derived gradient measures of form overlap from the Kinect motion tracking data, and showed that these "kinematic distance" scores correspond to meaning overlap (Pouw, de Wit, et al., 2021).

The questions and methods in this thesis have been inspired by various (partially overlapping) research fields: gesture studies, psycholinguistics, cognitive science, interactional linguistics and conversation analysis (CA). In chapters 3-5, a combination of quantitative and qualitative analyses has been used. Disciplinary integration of this kind is challenging due to conflicting ideas about how social interaction should be studied. Most notably, researchers in CA are wary of quantifying phenomena that are part of the intricate, multimodal process of interaction, because "any sort of formal coding risks a massive reduction and flattening of complex human behaviour to simplistic codes" (Stivers, 2015, p. 1). How to avoid this risk? According to conversation analysts, "we need to know what the phenomena are, how they are organised, and how they are related to each other as a precondition for cosently bringing methods of quantitative analysis to bear on them" (Schegloff, 1993, p. 114). Fortunately, other-initiated repair has been well-attested in CA (Drew, 1997; Kitzinger, 2013; Schegloff, 2000, 2004; Schegloff et al., 1977; to name a few prominent sources), which led to the development of a general coding scheme (Dingemanse et al., 2016) that I could use in this thesis. For alignment however, I agree that we should be careful when quantifying it. This is exactly why one of the objectives of this thesis was to conceptualise alignment as a phenomenon embedded in multimodal, sequential talk-in-interaction, and to create tools that allow researchers to operationalise alignment in a transparent and systematic way in empirical studies.

Another source for disagreement between disciplines is the study of task-based interactions in the laboratory versus spontaneous interactions in the field. But there are a number of good reasons for studying social interaction in the lab, as summarised by Kendrick (2017):

(1) the ability to generate convergent evidence;
(2) the ability to use cutting-edge technologies;
(3) the ability to address concerns about ecological validity (if needed);
(4) the possibility to disseminate conversation analytic knowledge to the wider scientific community.

In this thesis, point (1) holds for chapter 5 in particular, which included a replication of earlier findings on other-initiated repair in spontaneous interactions (Dingemanse, Roberts, et al., 2015). Point (2) becomes clear from the technology-heavy lab setting (including head-mounted microphones, multiple cameras and Microsoft Kinect (V2) devices), which would be hard to deploy in the field. As for point (3), many of the qualitative findings of this thesis are in line with those of studies of more naturalistic interactions (e.g., Chui, 2014 on gestural alignment; Jokipohja & Lilja, 2022 on manual gestures in other-initiated repair), though to be more certain of the generalisability of the newly-attested quantitative patterns, we would benefit from parallel analyses on naturalistic data (as suggested by Schegloff, 1991; Kendrick, 2017). Some specific concerns related to ecological validity are discussed in the Limitations section below.

Finally, with respect to point (4), I took into consideration that psychologists and cognitive scientists might not be familiar with qualitative, micro-analytic approaches to interactional data, let alone the complex transcription conventions for multimodal interaction (Mondada, 2018). Hence, I have opted for simplified transcripts in this thesis, making it as easy as possible for readers to get an impression of the multimodal behaviour, by presenting screenshots of the gestures directly next to the transcribed speech.

Generalisability of findings

To what extent can we generalise the findings of this thesis, which are based on task-based interactions taking place in the lab, to naturalistic language use settings? The referential communication task with Fribbles is peculiar, but we can find parallels with everyday interactions. Though we do not usually talk about strange blue objects that we have never seen before, we do refer to (new or known) objects, ideas or people for which we do not have a conventional label or name at hand. And while we do not usually look at separate screens with some information only being available to one person, we can drive a car while a friend is figuring out the route on their phone, giving directions while we keep our eyes on the road. Importantly, the task-based setting meets the ten features of "basic" language use settings defined by Clark (1996), pertaining to *immediacy, medium* and *control*. That is, participants can see and hear each other and their surroundings without delay; produce and receive simultaneously without a track record; and "speak for themselves, jointly determine who says what when, and formulate their utterances as they go" (Clark, 1996, p. 10). As such, we can expect the techniques and practices used by participants to be of the sort that we find in "basic" co-present conversational settings more generally.

There is however a particular feature of the task-based setting that warrants our attention: participants spend a lot of time looking at their screen, as opposed to the default of gazing at the face of one's conversational partner that we find in face-to-face conversations in similar cultures (Hessels, 2020). The height and viewing angle of the screens was set-up as to ensure visibility of the partner's main gesture area (McNeill, 1992) and to facilitate alternated gazing at the screen versus the partner. There are times where it is clearly visible on the video recordings that people gaze at their partner, but when they look down at their screen, it is unclear to what extent people perceive gestural and facial signals from their partner in their peripheral vision. This means that on the one hand, we should be careful to generalise the findings obtained in this setting to co-present, face-to-face conversations more generally. But on the other hand, we can take this task-based setting to be one instance of the many different conversational settings that humans take part in (across cultures and across digital technologies) and ask how people adapt their behaviour accordingly.

As for the generalisation of the current findings to face-to-face settings, it is important to note that we cannot be sure if alignment with form variation is due to participants not having seen their partners gesture, and whether alignment with full form overlap is accidental. Therefore, rather than drawing conclusions from the quantitative patterns alone, we benefit from looking at the qualitatively discussed examples (which show the verbal context and typically include notes on gaze) for additional information. For example, in Transcript 4.2 of chapter 4, gestural alignment occurred with the utterance "oh with the opened book you mean...", where the definite article and speech-gesture combination suggest that the form overlap follows from deliberate copying. Secondly, in naturalistic face-to-face conversations head movements and facial signals play an important role as signals of understanding and non-understanding (see e.g., Hömke, 2019). It might be that in this study, participants have missed these signals while they were searching their screen. It might also be that they have minimised their use of these signals (because they know that their partner would not perceive them), relying more heavily on verbal or gestural strategies to compensate for it.

As for the diversity in conversational settings, we can think about how the task-based setting in this thesis relates to some "non-basic" settings (Clark, 1996). For example, in multi-party video calls we can also experience issues with eye gaze and the lack of a shared visual space; while you can see your partner's face and eyes, you cannot know whose video they are looking at (George et al., 2022). Future research can investigate how people deal with such features, and how this impacts the coordinated use of speech and co-speech gesture in alignment and other-initiated repair. Another exciting research avenue is the field of human-computer interactions. When interacting with a computer, establishing "mutual understanding" can be more of a continuous struggle rather than a tacit process. This calls for the development of systems that are sensitive to signals of understanding and non-understanding (see e.g., Buschmeier & Kopp, 2018; Corti & Gillespie, 2016; Dingemanse & Liesenfeld, 2022), which could be modelled after the strategies that people use when mutual understanding is not a given, such as the multimodal repair strategies that are used to negotiate joint reference to Fribbles.

Turning to another generalisability concern, we should keep in mind that the participant group of this thesis is WEIRD (Henrich et al., 2010) and rather homogeneous; participants were all native speakers of Dutch, with the majority being students in their early twenties. How does the multimodal realisation of alignment and other-initiated repair as characterised in this study generalise to other cultures and languages? As discussed in chapter 4, for other-initiated repair there is quite some cross-linguistic work on lexicosyntactic resources, and the gestural strategies that I found are in line with earlier examples reported in the literature. Furthermore, though there is cultural variation in how people gesture, the use of iconic gestures to depict referents appears to be a universal strategy (see e.g., Cooperrider, 2019; Kita, 2009). For alignment this question is harder to answer. Alignment is found in all sorts of ways in conversations across cultures and languages, for example to respond to a turn (see e.g., Norrick, 1987; and discussion in Gipper, 2020), or to initiate repair (Dingemanse & Enfield, 2015). Yet it has been argued that the use and frequency of alignment might vary as a function of cultural differences, with repetition being regarded as "non-creative" in English, yet used to create interpersonal involvement in Japanese (Fujii, 2012). Finally, alignment can also differ across languages due to grammatical properties. For example, in many languages repetitional responses can be used instead of or in addition to particles like "yes" or "no" to respond to a prior turn (Gipper, 2020), and repetition of the verb can even be the default way to respond to polar questions, as is the case in "echo languages" such as Finnish and Welsh (Enfield et al., 2019; Holmberg, 2015).

Outlook: understanding mutual understanding

Some readers might have noticed—perhaps even been disappointed—that I have not explicitly operationalised or quantified "mutual understanding" in this thesis. Instead, I have thought of mutual understanding as a "temporally-bound achievement accomplished through (and embedded in) turns at talk, as a collaborative, 'public' achievement" (Sikveland & Ogden, 2012, p. 167). This take is in line with conversation analytic approaches, which is principally inspired by Ryle (1949) and Wittgenstein (1968; see the special issue edited by Koschmann, 2011). Yet another way to think of mutual understanding, more familiar to psychologists, would be to take understandings to be embedded in situation models (Pickering & Garrod, 2004) or goal and task representations (Sebanz et al., 2006; Vesper et al., 2010, 2017). People can then be considered to have mutual understanding when they mutually infer that their mental representations are sufficiently aligned or compatible (Stolk et al., 2016; Wheatley et al., 2019). The difference in these perspectives has been pointed out long ago by Garfinkel (1967), noting that in ethnomethodological approaches to social interaction (such as conversation analysis), mutual understanding is thought of as "an operation rather than a common intersection of overlapping sets" (p. 30).

But perhaps the time is ripe to bring these takes on mutual understanding together, and the work by the Communicative Alignment in Brain and Behaviour (CABB) team is a demonstration of that. One route to such integration is a computational approach, for example by studying how people deal with asymmetries in mental representations through pragmatic inferencing and collaborative interactional strategies (Blokpoel et al., 2019; van Arkel et al., 2020). Another route could be to investigate how changes in mental representations are related to social interactions, using neuroimaging methods. This does come with its challenges, as there is no consensus on what representations are (see e.g., Chalmers, 1993; Fodor & Pylyshyn, 1981) and where and how we can find them in the brain (see e.g., Bowers, 2009; Kiefer & Pulvermüller, 2012; Meteyard et al., 2012), though there are some concrete proposals that can be used to make a start (Binder et al. 2006).

In sum, arriving at an integrative understanding of how mutual understanding is embedded in both interactions and people's minds will require interdisciplinary efforts from across the field of cognitive science. An indispensable part of this endeavour should be the detailed study of behaviour in social interaction (Krakauer et al., 2017). With this thesis I hope to have shown that we should direct our attention to sequential environments like those involving other-initiated repair and alignment, bearing in mind their multidimensional nature (chapter 2) and multimodal realisation (chapters 3-5). It is in these environments—where people negotiate meaning turn-by-turn and bit-by-bit—that the causal locus of mutual understanding is most likely to be found.

Conclusion

Scholars have long carved up language use into separate modalities, where speech became the normative and primary locus of investigation. Meanwhile, people in interaction are unaffected by such scholarly dichotomies, and their main concern is to use whatever semiotic resources available to them to do the interactional work at hand. Here I have shown that people negotiate mutual understanding by means of alignment and other-initiated repair, deploying speech and co-speech gestures in a flexible way to form effective multimodal strategies to meet communicative demands. I also found that when people work together to resolve interactional trouble, they distribute the multimodal labour across repair initiations and repair solutions in predictable ways, thereby minimizing collaborative efforts at the level of the dyad. These findings underscore the notion that social interaction is a *joint action*, where I believe that people coordinate their use of speech and gesture "not only because they are trying to communicate effectively, but also because they are trying to communicate enjoyably" (Furuyama, 2002, p. 373).

Altogether, the research reported in this thesis helps us to understand what it means to say that social interaction is collaborative and multimodal, and opens up new horizons in the interdisciplinary study of mutual understanding.

References

References

Abney, D. H., Paxton, A., Dale, R., & Kello, C. T. (2014). Complexity matching in dyadic conversation. *Journal of Experimental Psychology: General*, *143*(6), 2304–2315. https://doi.org/10.1037/xge0000021

Albert, S., & de Ruiter, J. P. (2018). Repair: The interface between interaction and cognition. *Topics in Cognitive Science*, *10*(2), 279–313. https://doi.org/10.1111/tops.12339

Alibali, M. W., Nathan, M. J., Church, R. B., Wolfgram, M. S., Kim, S., & Knuth, E. J. (2013). Teachers' gestures and speech in mathematics lessons: Forging common ground by resolving trouble spots. *ZDM Mathematics Education*, *45*(3), 425–440. https://doi.org/10.1007/s11858-012-0476-0

Allwood, J., & Ahlsen, E. (1999). Learning how to manage communication, with special reference to the acquisition of linguistic feedback. *Journal of Pragmatics*, *31*(10), 1353–1389. https://doi.org/10.1016/S0378-2166(98)00109-X

Ameka, F. K., & Terkourafi, M. (2019). What if…? Imagining non-Western perspectives on pragmatic theory and practice. *Journal of Pragmatics*, *145*, 72–82. https://doi.org/10.1016/j.pragma.2019.04.001

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S., & Weinert, R. (1991). The HRCR Map Task corpus. *Language and Speech*, *34*(4), 351–366. https://doi.org/10.1177/002383099103400404

Andrews, D. (2014). Gestures as requests for information: Initiating repair operations in German native-speaker conversation. *Focus on German Studies*, *21*, 76–94.

Angus, D., & Wiles, J. (2018). Social semantic networks: Measuring topic management in discourse using a pyramid of conceptual recurrence metrics. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *28*(8), 085723. https://doi.org/10.1063/1.5024809

Bangerter, A., Mayor, E., & Knutsen, D. (2020). Lexical entrainment without conceptual pacts? Revisiting the matching task. *Journal of Memory and Language*, *114*, 104129. https://doi.org/10.1016/j.jml.2020.104129

Baranova, J. (2015). Other-initiated repair in Russian. *Open Linguistics*, *1*(1), 555–577. https://doi.org/10.1515/opli-2015-0019

Barry, T. J., Griffith, J. W., De Rossi, S., & Hermans, D. (2014). Meet the Fribbles: Novel stimuli for use within behavioural research. *Frontiers in Psychology*, *5*, 103. https://doi.org/10.3389/fpsyg.2014.00103

Bavelas, J., Black, A., Chovil, N., Lemery, C. R., & Mullett, J. (1988). Form and function in motor mimicry: Topographic evidence that the primary function is communicative. *Human Communication Research*, *14*(3), 275–299. https://doi.org/10.1111/j.1468-2958.1988.tb00158.x

Bavelas, J., & Chovil, N. (2000). Visible acts of meaning: An integrated message model of language in face-to-face dialogue. *Journal of Language and Social Psychology*, *19*(2), 163–194. https://doi.org/10.1177/0261927X00019002001

Bavelas, J., Chovil, N., Coates, L., & Roe, L. (1995). Gestures specialized for dialogue. *Personality and Social Psychology Bulletin*, *21*(4), 394–405. https://doi.org/10.1177/0146167295214010

Bavelas, J., Chovil, N., Lawrie, D. A., & Wade, A. (1992). Interactive gestures. *Discourse Processes*, *15*(4), 469–489. https://doi.org/10.1080/01638539209544823

Bavelas, J., Coates, L., & Johnson, T. (2000). Listeners as co-narrators. *Journal of Personality and Social Psychology*, *79*(6), 941–952. https://doi.org/10.1037//0022-3514.79.6.941

Bekke, M. ter, Drijvers, L., & Holler, J. (2020). *The predictive potential of hand gestures during conversation: An investigation of the timing of gestures in relation to speech*. PsyArXiv. https://doi.org/10.31234/osf.io/b5zq7

Bekkering, H., Wohlschläger, A., & Gattis, M. (2000). Imitation of gestures in children is goal-directed. *The Quarterly Journal of Experimental Psychology: Section A*, *53*(1), 153–164. https://doi.org/10.1080/713755872

Bergmann, K., & Kopp, S. (2012). Gestural alignment in natural dialogue. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 1326–1331). Cognitive Science Society.

Bertrand, R., Ferré, G., & Guardiola, M. (2013). French face-to-face interaction: Repetition as a multimodal resource. In M. Rojc & N. Campbell (Eds.), *Coverbal synchrony in human-machine interaction* (pp. 141–172). Science Publishers/CRC Press.

Beugher, S. D., Brône, G., & Goedemé, T. (2018). A semi-automatic annotation tool for unobtrusive gesture analysis. *Language Resources and Evaluation*, *52*(2), 433–460. https://doi.org/10.1007/s10579-017-9404-9

Blokpoel, M., Dingemanse, M., Woensdregt, M., Kachergis, G., Bögels, S., Toni, I., & van Rooij, I. (2019). *Pragmatic communicators can overcome asymmetry by exploiting ambiguity*. OSF Preprints. https://doi.org/10.31219/osf.io/q56xs

Blythe, J. (2015). Other-initiated repair in Murrinh-Patha. *Open Linguistics*, *1*(1). https://doi.org/10.1515/opli-2015-0003

Bolinger, D. (1968). *Aspects of language*. Harcourt, Brace, and World.

Bowers, J. S. (2009). On the biological plausibility of grandmother cells: Implications for neural network theories in psychology and neuroscience. *Psychological Review*, *116*(1), 220–251. https://doi.org/10.1037/a0014462

Branigan, H. P., Pickering, M. J., & Cleland, A. A. (2000). Syntactic co-ordination in dialogue. *Cognition*, *75*(2), B13–B25. https://doi.org/10.1016/S0010-0277(99)00081-5

Branigan, H. P., Pickering, M. J., McLean, J. F., & Cleland, A. A. (2007). Syntactic alignment and participant role in dialogue. *Cognition*, *104*(2), 163–197. https://doi.org/10.1016/j.cognition.2006.05.006

Branigan, H. P., Pickering, M. J., Pearson, J., McLean, J. F., & Brown, A. (2011). The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers. *Cognition*, *121*(1), 41–57. https://doi.org/10.1016/j.cognition.2011.05.011

Brass, M., & Heyes, C. (2005). Imitation: Is cognitive neuroscience solving the correspondence problem? *Trends in Cognitive Sciences*, *9*(10), 489–495. https://doi.org/10.1016/j.tics.2005.08.007

Brennan, S. E. (1996). Lexical entrainment in spontaneous dialog. *Proceedings of the International Symposium on Spoken Dialogue*, *96*, 41–44.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1482–1493. https://doi.org/10.1037/0278-7393.22.6.1482

Brennan, S. E., Galati, A., & Kuhlen, A. K. (2010). Two minds, one dialog: Coordinating speaking and understanding. *Psychology of Learning and Motivation*, *53*, 301–344. https://doi.org/10.1016/S0079-7421(10)53008-1

Brown, G., Anderson, A., Schillock, R., & Yule, G. (1984). *Teaching talk*. Cambridge University Press.

Buschmeier, H., Bergmann, K., & Kopp, S. (2010). Modelling and evaluation of lexical and syntactic alignment with a priming-based microplanner. In E. Krahmer & M. Theune (Eds.), *Empirical methods in natural language generation: Data-oriented methods and empirical evaluation* (pp. 85–104). Springer.

Buschmeier, H., & Kopp, S. (2018). Communicative listener feedback in human-agent interaction: Artificial speakers need to be attentive and adaptive. *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems*, 1213–1221.

Byun, K.-S., Vos, C. de, Zeshan, U., & Levinson, S. (2019). Repair in cross-signing: Trouble sources, repair strategies and communicative success. In U. Zeshan & J. Webster (Eds.), *Sign multilingualism* (pp. 23–80). De Gruyter Mouton.

Cai, Z. G., Sun, Z., & Zhao, N. (2021). Interlocutor modelling in lexical alignment: The role of linguistic competence. *Journal of Memory and Language*, *121*, 104278. https://doi.org/10.1016/j.jml.2021.104278

Cann, R., Kempson, R., Marten, L., & Otsuka, M. (2005). Right node raising, coordination and the dynamics of language processing. *Lingua*, *115*(4), 503–525. https://doi.org/10.1016/j.lingua.2003.09.013

Capirci, O., Bonsignori, C., & Di Renzo, A. (2022). Signed languages: A triangular semiotic dimension. *Frontiers in Psychology*, *12*, 802911. https://doi.org/10.3389/fpsyg.2021.802911

Cassell, J., McNeill, D., & McCullough, K.-E. (1998). Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information. *Pragmatics & Cognition*, *6*(2), 1–33. https://doi.org/10.1075/pc.7.1.03cas

Chalmers, D. J. (1993). Connectionism and compositionality: Why Fodor and Pylyshyn were wrong. *Philosophical Psychology*, *6*(3), 305–319. https://doi.org/10.1080/09515089308573094

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception–behavior link and social interaction. *Journal of Personality and Social Psychology*, *76*(6), 893–910. https://doi.org/10.1037/0022-3514.76.6.893

Chartrand, T. L., & van Baaren, R. (2009). Human mimicry. *Advances in Experimental Social Psychology*, *41*, 219–274. https://doi.org/10.1016/S0065-2601(08)00405-X

Cheney, D. L., & Seyfarth, R. M. (2005). Constraints and preadaptations in the earliest stages of language evolution. *The Linguistic Review*, *22*(2–4), 135–159. https://doi.org/10.1515/tlir.2005.22.2-4.135

Chui, K. (2014). Mimicked gestures and the joint construction of meaning in conversation. *Journal of Pragmatics*, *70*, 68–85. https://doi.org/10.1016/j.pragma.2014.06.005

Clark, H. H. (1996). *Using language*. Cambridge University Press.

Clark, H. H. (2016). Depicting as a method of communication. *Psychological Review*, *123*(3), 324–347. https://doi.org/10.1037/rev0000026

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127–149). American Psychological Association.

Clark, H. H., & Schaefer, E. F. (1987). Collaborating on contributions to conversations. *Language and Cognitive Processes*, *2*(1), 19–41. https://doi.org/10.1080/01690968708406350

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39. https://doi.org/10.1016/0010-0277(86)90010-7

Clayman, S. E. (2013). Turn-constructional units and the transition-relevance place. In J. Sidnell & T. Stivers (Eds.), *The handbook of conversation analysis* (pp. 151–166). John Wiley & Sons, Ltd.

Cleland, A. A., & Pickering, M. J. (2003). The use of lexical and syntactic information in language production: Evidence from the priming of noun-phrase structure. *Journal of Memory and Language*, *49*(2), 214–230. https://doi.org/10.1016/S0749-596X(03)00060-3

Clift, R. (2016). *Conversation analysis*. Cambridge University Press.

Colman, M., & Healey, P. G. T. (2011). The distribution of repair in dialogue. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 1563–1568). Cognitive Science Society.

Cooperrider, K. (2019). Universals and diversity in gesture: Research past, present, and future. *Gesture*, *18*(2–3), 209–238. https://doi.org/10.1075/gest.19011.coo

Cooperrider, K., Abner, N., & Goldin-Meadow, S. (2018). The palm-up puzzle: Meanings and origins of a widespread form in gesture and sign. *Frontiers in Communication*, *3*, 23. https://doi.org/10.3389/fcomm.2018.00023

Corti, K., & Gillespie, A. (2016). Co-constructing intersubjectivity with artificial conversational agents: People are more likely to initiate repairs of misunderstandings with agents represented as human. *Computers in Human Behavior*, *58*, 431–442. https://doi.org/10.1016/j.chb.2015.12.039

Costa, A., Pickering, M. J., & Sorace, A. (2008). Alignment in second language dialogue. *Language and Cognitive Processes*, *23*(4), 528–556. https://doi.org/10.1080/01690960801920545

Couper-Kuhlen, E., & Selting, M. (2017). *Interactional linguistics: Studying language in social interaction*. Cambridge University Press.

Curl, T. S. (2005). Practices in other-initiated repair resolution: The phonetic differentiation of ‘repetitions’. *Discourse Processes*, *39*(1), 1–43. https://doi.org/10.1207/s15326950dp3901_1

Dale, R., Fusaroli, R., Duran, N. D., & Richardson, D. C. (2013). The self-organization of human interaction. *Psychology of Learning and Motivation*, *59*, 43–95. https://doi.org/10.1016/B978-0-12-407187-2.00002-2

Dale, R., & Spivey, M. J. (2006). Unraveling the dyad: Using recurrence analysis to explore patterns of syntactic coordination between children and caregivers in conversation. *Language Learning*, *56*(3), 391–430. https://doi.org/10.1111/j.1467-9922.2006.00372.x

de Fornel, M. (1992). The return gesture: Some remarks on context, inference, and iconic gesture. In P. Auer & A. Di Luzio (Eds.), *The contextualization of language* (pp. 159–176). John Benjamins Publishing Company.

De Jaegher, H., Di Paolo, E., & Gallagher, S. (2010). Can social interaction constitute social cognition? *Trends in Cognitive Sciences*, *14*(10), 441–447. https://doi.org/10.1016/j.tics.2010.06.009

De Looze, C., Scherer, S., Vaughan, B., & Campbell, N. (2014). Investigating automatic measurements of prosodic accommodation and its dynamics in social interaction. *Speech Communication*, *58*, 11–34. https://doi.org/10.1016/j.specom.2013.10.002

de Ruiter, J. P. (2007). Postcards from the mind: The relationship between speech, imagistic gesture, and thought. *Gesture*, *7*(1), 21–38. https://doi.org/10.1075/gest.7.1.03rui

de Ruiter, J. P. (2013). Methodological paradigms in interaction research. In I. Wachsmuth, J. P. De Ruiter, P. Jaecks, & S. Kopp (Eds.), *Alignment in communication: Towards a new theory of communication* (pp. 11–31). John Benjamins Publishing Company.

de Ruiter, J. P., Bangerter, A., & Dings, P. (2012). The interplay between gesture and speech in the production of referring expressions: Investigating the tradeoff hypothesis. *Topics in Cognitive Science*, *4*(2), 232–248. https://doi.org/10.1111/j.1756-8765.2012.01183.x

Denby, J., & Yurovsky, D. (2019). Parents’ linguistic alignment predicts children’s language development. In A. K. Goel, C. M. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Conference of the Cognitive Science Society* (pp. 1627–1632). Cognitive Science Society.

Dideriksen, C., Christiansen, M. H., Tylén, K., Dingemanse, M., & Fusaroli, R. (2020). *Building common ground: Quantifying the interplay of mechanisms that promote understanding in conversations*. PsyArXiv. https://doi.org/10.31234/osf.io/a5r74

Dideriksen, C., Fusaroli, R., Tylén, K., Dingemanse, M., & Christiansen, M. H. (2019). Contextualizing conversational strategies: Backchannel, repair and linguistic alignment in spontaneous and task-oriented conversations. In Goel, A.K., Seifert, C.M., & Freksa, C. (Eds.), *Proceedings of the 41st Annual Conference of the Cognitive Science Society* (pp. 261–267). Cognitive Science Society.

Dijksterhuis, A., & Bargh, J. A. (2001). The perception-behavior expressway: Automatic effects of social perception on social behavior. *Advances in Experimental Social Psychology*, *33*, 1–40. https://doi.org/10.1016/S0065-2601(01)80003-4

Dingemanse, M. (2015). Other-initiated repair in Siwu. *Open Linguistics*, *1*(1), 232–255. https://doi.org/10.1515/opli-2015-0001

Dingemanse, M. (2020). Recruiting assistance and collaboration: A West-African corpus study. In S. Floyd, G. Rossi, & N. Enfield (Eds.), *Getting others to do things: A pragmatic typology of recruitments* (pp. 369–421). Language Science Press.

Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2015). Arbitrariness, iconicity, and systematicity in language. *Trends in Cognitive Sciences*, *19*(10), 603–615. https://doi.org/10.1016/j.tics.2015.07.013

Dingemanse, M., Blythe, J., & Dirksmeyer, T. (2014). Formats for other-initiation of repair across languages: An exercise in pragmatic typology. *Studies in Language*, *38*(1), 5–43. https://doi.org/10.1075/sl.38.1.01din

Dingemanse, M., & Enfield, N. J. (2015). Other-initiated repair across languages: Towards a typology of conversational structures. *Open Linguistics*, *1*, 96–118. https://doi.org/10.2478/opli-2014-0007

Dingemanse, M., Kendrick, K. H., & Enfield, N. J. (2016). A coding scheme for other-initiated repair across languages. *Open Linguistics*, *2*, 35–46. https://doi.org/10.1515/opli-2016-0002

Dingemanse, M., & Liesenfeld, A. (2022). From text to talk: Harnessing conversational corpora for humane and diversity-aware language technology. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5614–5633. https://doi.org/10.18653/v1/2022.acl-long.385

Dingemanse, M., Roberts, S. G., Baranova, J., Blythe, J., Drew, P., Floyd, S., Gisladottir, R. S., Kendrick, K. H., Levinson, S. C., Manrique, E., Rossi, G., & Enfield, N. J. (2015). Universal principles in the repair of communication problems. *PLOS ONE*, *10*(9), e0136100. https://doi.org/10.1371/journal.pone.0136100

Dively, V. L. (1998). Conversational repairs in ASL. In C. Lucas (Ed.), *Pinky extension and eye gaze: Language use in deaf communities* (pp. 137–169). Gallaudet University Press.

Drew, P. (1997). 'Open' class repair initiators in response to sequential sources of troubles in conversation. *Journal of Pragmatics*, *28*(1), 69–101. https://doi.org/10.1016/S0378-2166(97)89759-7

Duran, N. D., Paxton, A., & Fusaroli, R. (2019). ALIGN: Analyzing linguistic interactions with generalizable techNiques - A Python library. *Psychological Methods*, *24*(4), 419–438. https://doi.org/10.1037/met0000206

Egbert, M. M. (1996). Context-sensitivity in conversation: Eye gaze and the German repair initiator bitte? *Language in Society*, *25*(4), 587–612. https://doi.org/10.1017/S0047404500020820

Eijk, L., Rasenberg, M., Arnese, F., Blokpoel, M., Dingemanse, M., Döller, C., Ernestus, M., Holler, J., Milivojevic, B., Özyürek, A., Pouw, W., van Rooij, I., Schriefers, H., Toni, I., Trujillo, J., & Bögels, S. (2022). The CABB dataset: A multimodal corpus of communicative interactions for behavioural and neural analyses. *NeuroImage*, *264*, 119734. https://doi.org/10.1016/j.neuroimage.2022.119734

Elffers, J. (1976). *Tangram: The ancient Chinese shapes game*. Penguin Books.

Enfield, N. J. (2009). *The anatomy of meaning: Speech, gesture, and composite utterances*. Cambridge University Press.

Enfield, N. J. (2013). *Relationship thinking: Agency, enchrony, and human sociality*. Oxford University Press.

Enfield, N. J., Stivers, T., Brown, P., Englert, C., Harjunpää, K., Hayashi, M., Heinemann, T., Hoymann, G., Keisanen, T., & Rauniomaa, M. (2019). Polar answers. *Journal of Linguistics*, *55*(2), 277–304. https://doi.org/10.1017/S0022226718000336

Engelbrecht, S. E. (2001). Minimum principles in motor control. *Journal of Mathematical Psychology*, *45*(3), 497–542. https://doi.org/10.1006/jmps.2000.1295

Engle, R. A. (1998). Not channels but composite signals: Speech, gesture, diagrams and object demonstrations are integrated. In M. A. Gernsbacher & S. J. Derry (Eds.), *Proceedings of the 20th Annual Conference of the Cognitive Science Society* (pp. 321–326). Psychology Press.

Engle, R. A., & Clark, H. H. (1995). *Using composites of speech, gestures, diagrams and demonstrations in explanations of mechanical devices*. American Association for Applied Linguistics Conference, Long Beach, CA.

Fay, N., Arbib, M., & Garrod, S. (2013). How to bootstrap a human communication system. *Cognitive Science*, *37*(7), 1356–1367. https://doi.org/10.1111/cogs.12048

Fay, N., Lister, C. J., Ellison, T. M., & Goldin-Meadow, S. (2014). Creating a communication system from scratch: Gesture beats vocalization hands down. *Frontiers in Psychology*, *5*, 354. https://doi.org/10.3389/fpsyg.2014.00354

Fay, N., Walker, B., Swoboda, N., & Garrod, S. (2018). How to create shared symbols. *Cognitive Science*, *42*(S1), 241–269. https://doi.org/10.1111/cogs.12600

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low Kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, *43*(6), 543–549. https://doi.org/10.1016/0895-4356(90)90158-L

Feng, G. C. (2014). Intercoder reliability indices: Disuse, misuse, and abuse. *Quality & Quantity*, *48*(3), 1803–1815. https://doi.org/10.1007/s11135-013-9956-8

Ferrara, L., & Hodge, G. (2018). Language as description, indication, and depiction. *Frontiers in Psychology*, *9*, 716. https://doi.org/10.3389/fpsyg.2018.00716

Floyd, S. (2020). Getting others to do things in the Cha'palaa language of Ecuador. In S. Floyd, G. Rossi, & N. Enfield (Eds.), *Getting others to do things: A pragmatic typology of recruitments* (pp. 51–92). Language Science Press.

Floyd, S., Manrique, E., Rossi, G., & Torreira, F. (2016). Timing of visual bodily behavior in repair sequences: Evidence from three languages. *Discourse Processes*, *53*(3), 175–204. https://doi.org/10.1080/0163853X.2014.992680

Fodor, J. A., & Pylyshyn, Z. W. (1981). How direct is visual perception? Some reflections on Gibson's 'ecological approach'. *Cognition*, *9*(2), 139–196. https://doi.org/10.1016/0010-0277(81)90009-3

Foushee, R., Byrne, D., Casillas, M., & Goldin-Meadow, S. (2021). *Differential impacts of linguistic alignment across caregiver-child dyads and levels of linguistic structure*. CUNY 2021.

Fox, B., Hayashi, M., & Jasperson, R. (1996). Resources and repair: A cross-linguistic study of syntax and repair. *Studies in Interactional Sociolinguistics*, *13*, 185–237. https://doi.org/10.1017/CBO9780511620874.004

Fujii, Y. (2012). Differences of situating Self in the place/ba of interaction between the Japanese and American English speakers. *Journal of Pragmatics*, *44*(5), 636–662. https://doi.org/10.1016/j.pragma.2011.09.007

Furuyama, N. (2002). Prolegomena of a theory of between-person coordination of speech and gesture. *International Journal of Human-Computer Studies*, *57*(4), 347–374. https://doi.org/10.1006/ijhc.2002.1021

Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C., & Tylén, K. (2012). Coming to terms: Quantifying the benefits of linguistic coordination. *Psychological Science*, *23*(8), 931–939. https://doi.org/10.1177/0956797612436816

Fusaroli, R., Konvalinka, I., & Wallot, S. (2014). Analyzing social interactions: The promises and challenges of using cross recurrence quantification analysis. In N. Marwan, M. Riley, A. Giuliani, & C. L. Webber, (Eds.), *Translational recurrences* (pp. 137–155). Springer International Publishing. https://doi.org/10.1007/978-3-319-09531-8_9

Fusaroli, R., Rączaszek-Leonardi, J., & Tylén, K. (2014). Dialog as interpersonal synergy. *New Ideas in Psychology*, *32*, 147–157. https://doi.org/10.1016/j.newideapsych.2013.03.005

Fusaroli, R., & Tylén, K. (2016). Investigating conversational dynamics: Interactive alignment, interpersonal synergy, and collective task performance. *Cognitive Science*, *40*(1), 145–171. https://doi.org/10.1111/cogs.12251

Fusaroli, R., Tylén, K., Garly, K., Steensig, J., Christiansen, M. H., & Dingemanse, M. (2017). Measures and mechanisms of common ground: Backchannels, conversational repair, and interactive alignment in free and task-oriented social interactions. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 2055–2060). Cognitive Science Society.

Fusaroli, R., Weed, E., Fein, D., & Naigles, L. (2021). *Caregiver linguistic alignment to autistic and typically developing children*. PsyArXiv. https://doi.org/10.31234/osf.io/ysjec

Gandolfi, G., Pickering, M. J., & Garrod, S. (in press). Mechanisms of alignment: Shared control, social cognition, and metacognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*.

Garfinkel, H. (1967). *Studies in ethnomethodology*. Prentice-Hall.

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, *27*(2), 181–218. https://doi.org/10.1016/0010-0277(87)90018-7

Garrod, S., & Clark, A. (1993). The development of dialogue co-ordination skills in schoolchildren. *Language and Cognitive Processes*, *8*(1), 101–126. https://doi.org/10.1080/01690969308406950

Garrod, S., Fay, N., Rogers, S., Walker, B., & Swoboda, N. (2010). Can iterated learning explain the emergence of graphical symbols? *Interaction Studies*, *11*(1), 33–50. https://doi.org/10.1075/is.11.1.04gar

Gauthier, I., & Tarr, M. J. (1997). Becoming a "Greeble" expert: Exploring mechanisms for face recognition. *Vision Research*, *37*(12), 1673–1682. https://doi.org/10.1016/s0042-6989(96)00286-6

George, J., Mirsadikov, A., Nabors, M., & Marett, K. (2022). What do users actually look at during videoconference calls? Exploratory research on attention, distraction effects and gender. In T. Bui (Ed.), *Proceedings of the 55th Hawaii International Conference on System Sciences* (pp. 4779–4787).

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*, *23*(5), 389–407. https://doi.org/10.1016/j.tics.2019.02.003

Ginzburg, J., & Cooper, R. (2004). Clarification, ellipsis, and the nature of contextual updates in dialogue. *Linguistics and Philosophy*, *27*(3), 297–365. https://doi.org/10.1023/B:LING.0000023369.19306.90

Gipper, S. (2020). Repeating responses as a conversational affordance for linguistic transmission: Evidence from Yurakaré conversations. *Studies in Language*, *44*(2), 281–326. https://doi.org/10.1075/sl.19041.gip

Goldin-Meadow, S. (2017). What the hands can tell us about language emergence. *Psychonomic Bulletin & Review*, *24*(1), 213–218. https://doi.org/10.3758/s13423-016-1074-x

Goodwin, C. (1979). The interactive construction of a sentence in natural conversation. *Everyday Language: Studies in Ethnomethodology*, *97*, 101–121.

Goodwin, C. (1981). *Conversational organization*. Academic Press.

Goodwin, C. (2000). Action and embodiment within situated human interaction. *Journal of Pragmatics*, *32*(10), 1489–1522. https://doi.org/10.1016/s0378-2166(99)00096-x

Goodwin, M. H., & Goodwin, C. (1986). Gesture and coparticipation in the activity of searching for a word. *Semiotica*, *62*(1–2), 51–75. https://doi.org/10.1515/semi.1986.62.1-2.51

Goudbeek, M., & Krahmer, E. (2012). Alignment in interactive reference production: Content planning, modifier ordering, and referential overspecification. *Topics in Cognitive Science*, *4*(2), 269–289. https://doi.org/10.1111/j.1756-8765.2012.01186.x

Graziano, M., Kendon, A., & Cristilli, C. (2011). 'Parallel gesturing' in adult-child conversations. In G. Stam & M. Ishino (Eds.), *Integrating gestures: The interdisciplinary nature of gesture* (pp. 89–101). John Benjamins Publishing Company.

Gries, S. T., & Kootstra, G. J. (2017). Structural priming within and across languages: A corpus-based perspective. *Bilingualism: Language and Cognition*, *20*(2), 235–250. https://doi.org/10.1017/S1366728916001085

Hartsuiker, R. J., Bernolet, S., Schoonbaert, S., Speybroeck, S., & Vanderelst, D. (2008). Syntactic priming persists while the lexical boost decays: Evidence from written and spoken dialogue. *Journal of Memory and Language*, *58*(2), 214–238. https://doi.org/10.1016/j.jml.2007.07.003

Haviland, J. B. (2013). The emerging grammar of nouns in a first generation sign language: Specification, iconicity, and syntax. *Gesture*, *13*(3), 309–353. https://doi.org/10.1075/gest.13.3.04hav

Hayashi, M., Raymond, G., & Sidnell, J. (Eds.). (2013). *Conversational repair and human understanding*. Cambridge University Press.

Healey, P. G. T., Mills, G. J., Eshghi, A., & Howes, C. (2018). Running repairs: Coordinating meaning in dialogue. *Topics in Cognitive Science*, *10*(2), 367–388. https://doi.org/10.1111/tops.12336

Healey, P. G. T., Plant, N. J., Howes, C., & Lavelle, M. (2015). When words fail: Collaborative gestures during clarification dialogues. *Turn-Taking and Coordination in Human-Machine Interaction: Papers from the 2015 AAAI Spring Symposium*, 23–29.

Healey, P. G. T., Purver, M., & Howes, C. (2014). Divergence in dialogue. *PLOS ONE*, *9*(6), e98598. https://doi.org/10.1371/journal.pone.0098598

Healey, P. G. T., Swoboda, N., Umata, I., & King, J. (2007). Graphical language games: Interactional constraints on representational form. *Cognitive Science*, *31*(2), 285–309. https://doi.org/10.1080/15326900701221363

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*(7302), 29. https://doi.org/10.1038/466029a

Hessels, R. S. (2020). How does gaze to faces support face-to-face interaction? A review and perspective. *Psychonomic Bulletin & Review*, *27*(5), 856–881. https://doi.org/10.3758/s13423-020-01715-w

Heyes, C. (2011). Automatic imitation. *Psychological Bulletin*, *137*(3), 463–483. https://doi.org/10.1037/a0022288

Hoetjes, M., Koolen, R., Goudbeek, M., Krahmer, E., & Swerts, M. (2015). Reduction in gesture during the production of repeated references. *Journal of Memory and Language*, *79–80*, 1–17. https://doi.org/10.1016/j.jml.2014.10.004

Hoetjes, M., Krahmer, E., & Swerts, M. (2015). On what happens in gesture when communication is unsuccessful. *Speech Communication*, *72*, 160–175. https://doi.org/10.1016/j.specom.2015.06.004

Holler, J. (2022). Visual bodily signals as core devices for coordinating minds in interaction. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, *377*(1859), 20210094. https://doi.org/10.1098/rstb.2021.0094

Holler, J., & Bavelas, J. (2017). Multi-modal communication of common ground: A review of social functions. In R. B. Church, M. W. Alibali, & S. D. Kelly (Eds.), *Why gesture? How the hands function in speaking, thinking and communicating* (pp. 213–240). Benjamins.

Holler, J., & Wilkin, K. (2011a). Co-speech gesture mimicry in the process of collaborative referring during face-to-face dialogue. *Journal of Nonverbal Behavior*, *35*(2), 133–153. https://doi.org/10.1007/s10919-011-0105-6

Holler, J., & Wilkin, K. (2011b). An experimental investigation of how addressee feedback affects co-speech gestures accompanying speakers' responses. *Journal of Pragmatics*, *43*(14), 3522–3536. https://doi.org/10.1016/j.pragma.2011.08.002

Holmberg, A. (2015). *The syntax of yes and no*. Oxford University Press.

Hömke, P. (2019). *The face in face-to-face communication: Signals of understanding and non-understanding* [PhD Thesis]. Radboud University Nijmegen.

Horton, L. (2020). Representational strategies in shared homesign systems from Nebaj, Guatemala. In O. Le Guen, J. Safar, & M. Coppola (Eds.), *Emerging sign languages of the Americas* (pp. 97–154). De Gruyter Mouton.

Horton, W. S., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition*, *59*(1), 91–117. https://doi.org/10.1016/0010-0277(96)81418-1

Hostetter, A. B. (2011). When do gestures communicate? A meta-analysis. *Psychological Bulletin*, *137*(2), 297–315. https://doi.org/10.1037/a0022128

Howes, C., Healey, P. G. T., & Purver, M. (2010). Tracking lexical and syntactic alignment in conversation. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the Annual Meeting of the 32nd Cognitive Science Society* (pp. 2004–2009). Cognitive Science Society.

Hutchins, E. (1995). *Cognition in the wild*. MIT press.

Iverson, J. M., & Goldin-Meadow, S. (2005). Gesture paves the way for language development. *Psychological Science*, *16*(5), 367–371. https://doi.org/10.1111/j.0956-7976.2005.01542.x

Jaeger, T. F., & Tily, H. (2011). On language 'utility': Processing complexity and communicative efficiency. *WIREs Cognitive Science*, *2*(3), 323–335. https://doi.org/10.1002/wcs.126

Jefferson, G. (1972). Side sequences. In D. N. Sudnow (Ed.), *Studies in social interaction* (pp. 294–338). MacMillan/The Free Press.

Jefferson, G. (2004). Glossary of transcription conventions. In G. H. Lerner (Ed.), *Conversation analysis: Studies from the first generation* (pp. 14–31). John Benjamins Publishing Company.

Jokipohja, A.-K., & Lilja, N. (2022). Depictive hand gestures as candidate understandings. *Research on Language and Social Interaction*, *55*(2), 123–145. https://doi.org/10.1080/08351813.2022.2067425

Kaschak, M. P., Kutta, T. J., & Schatschneider, C. (2011). Long-term cumulative structural priming persists for (at least) one week. *Memory & Cognition*, *39*(3), 381–388. https://doi.org/10.3758/s13421-010-0042-3

Kelly, S. D., Barr, D. J., Church, R. B., & Lynch, K. (1999). Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory. *Journal of Memory and Language*, *40*(4), 577–592. https://doi.org/10.1006/jmla.1999.2634

Kelly, S. D., Özyürek, A., & Maris, E. (2010). Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension. *Psychological Science*, *21*(2), 260–267. https://doi.org/10.1177/0956797609357327

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press.

Kendon, A. (2014). Semiotic diversity in utterance production and the concept of 'language'. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651), 20130293. https://doi.org/10.1098/rstb.2013.0293

Kendrick, K. H. (2015). Other-initiated repair in English. *Open Linguistics*, *1*(1), 164–190. https://doi.org/10.2478/opli-2014-0009

Kendrick, K. H. (2017). Using conversation analysis in the lab. *Research on Language and Social Interaction*, *50*(1), 1–11. https://doi.org/10.1080/08351813.2017.1267911

Kendrick, K. H., & Holler, J. (2017). Gaze direction signals response preference in conversation. *Research on Language and Social Interaction*, *50*(1), 12–32. https://doi.org/10.1080/08351813.2017.1262120

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods*, *42*(3), 643–650. https://doi.org/10.3758/BRM.42.3.643

Keysar, B., Barr, D. J., & Horton, W. S. (1998). The egocentric basis of language use: Insights from a processing approach. *Current Directions in Psychological Science*, *7*(2), 46–50. https://doi.org/10.1111/1467-8721.ep13175613

Kiefer, M., & Pulvermüller, F. (2012). Conceptual representations in mind and brain: Theoretical developments, current evidence and future directions. *Cortex*, *48*(7), 805–825. https://doi.org/10.1016/j.cortex.2011.04.006

Kim, K. (1999). Other-initiated repair sequences in Korean conversations: Types and functions. *Discourse and Cognition*, *6*(2), 141–168.

Kimbara, I. (2006). On gestural mimicry. *Gesture*, *6*(1), 39–61. https://doi.org/10.1075/gest.6.1.03kim

Kimbara, I. (2008). Gesture form convergence in joint description. *Journal of Nonverbal Behavior*, *32*(2), 123–131. https://doi.org/10.1007/s10919-007-0044-4

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102. https://doi.org/10.1016/j.cognition.2015.03.016

Kita, S. (2009). Cross-cultural variation of speech-accompanying gesture: A review. *Language and Cognitive Processes*, *24*(2), 145–167. https://doi.org/10.1080/01690960802586188

Kita, S., & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language*, *48*(1), 16–32. https://doi.org/10.1016/S0749-596X(02)00505-3

Kita, S., van Gijn, I., & van der Hulst, H. (1998). Movement phases in signs and co-speech gestures, and their transcription by human coders. In I. Wachsmuth & M. Fröhlich (Eds.), *Gesture and sign language in human-computer interaction* (pp. 23–35). Springer. https://doi.org/10.1007/BFb0052986

Kitzinger, C. (2013). Repair. In J. Sidnell & T. Stivers (Eds.), *The handbook of conversation analysis* (pp. 229–256). John Wiley & Sons, Ltd.

Kockelman, P. (2005). The semiotic stance. *Semiotica*, *2005*(157), 233–304. https://doi.org/10.1515/semi.2005.2005.157.1-4.233

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kopp, S., & Bergmann, K. (2013). Automatic and strategic alignment of co-verbal gestures in dialogue. In I. Wachsmuth, J. de Ruiter, P. Jaecks, & S. Kopp (Eds.), *Alignment in Communication: Towards a new theory of communication* (pp. 87–108). John Benjamins Publishing Company.

Koschmann, T. (Ed.). (2011). Understanding understanding in action [Special issue]. *Journal of Pragmatics*, *43*(2), 435–690. https://doi.org/10.1016/j.pragma.2010.08.016

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., & Poeppel, D. (2017). Neuroscience needs behavior: Correcting a reductionist bias. *Neuron*, *93*(3), 480–490. https://doi.org/10.1016/j.neuron.2016.12.041

Krauss, R. M., & Glucksberg, S. (1969). The development of communication: Competence as a function of age. *Child Development*, *40*(1), 255–266. https://doi.org/10.2307/1127172

Krauss, R. M., & Weinheimer, S. (1964). Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science*, *1*(1), 113–114. https://doi.org/10.3758/BF03342817

Krauss, R. M., & Weinheimer, S. (1966). Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *Journal of Personality and Social Psychology*, *4*(3), 343–346. https://doi.org/10.1037/h0023705

Kraut, R. E., Fussell, S. R., & Siegel, J. (2003). Visual information as a conversational resource in collaborative physical tasks. *Human–Computer Interaction*, *18*(1–2), 13–49. https://doi.org/10.1207/S15327051HCI1812_2

Kutas, M., Van Petten, C. K., & Kluender, R. (2006). Psycholinguistics electrified II (1994–2005). In M. J. Traxler & M. A. Gernsbacher (Eds.), *Handbook of psycholinguistics (2nd edition)* (pp. 659–724). Academic Press.

Laalo, K., & Argus, R. (2020). Linguistic recycling in language acquisition: Child-directed speech and child speech in the study of language acquisition. *AILA Review*, *33*(1), 86–103. https://doi.org/10.1075/aila.00031.laa

Lakin, J. L., & Chartrand, T. L. (2003). Using nonconscious behavioral mimicry to create affiliation and rapport. *Psychological Science*, *14*(4), 334–339. https://doi.org/10.1111/1467-9280.14481

Lerner, G. H. (1991). On the syntax of sentences-in-progress. *Language in Society*, *20*(3), 441–458. https://doi.org/10.1017/S0047404500016572

Lerner, G. H. (2002). Turn-sharing: The choral co-production of talk in interaction. In C. E. Ford, B. A. Fox, & S. A. Thompson (Eds.), *The language of turn and sequence* (pp. 225–256). Oxford University Press.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. MIT Press.

Levelt, W. J. M., & Kelter, S. (1982). Surface form and memory in question answering. *Cognitive Psychology*, *14*(1), 78–106. https://doi.org/10.1016/0010-0285(82)90005-6

Levinson, S. C. (2006). On the human 'interaction engine'. In S. C. Levinson & N. J. Enfield (Eds.), *Roots of human sociality* (pp. 39–69). Routledge.

Levinson, S. C. (2013). Action formation and ascription. In T. Stivers & J. Sidnell (Eds.), *The handbook of conversation analysis* (pp. 103–130). Wiley-Blackwell.

Levinson, S. C. (2015). Other-initiated repair in Yélî Dnye: Seeing eye-to-eye in the language of Rossel Island. *Open Linguistics*, *1*(1), 386–410. https://doi.org/10.1515/opli-2015-0009

Levinson, S. C. (2016). Turn-taking in human communication: Origins and implications for language processing. *Trends in Cognitive Sciences*, *20*(1), 6–14. https://doi.org/10.1016/j.tics.2015.10.010

Levinson, S. C., & Holler, J. (2014). The origin of human multi-modal communication. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651), 20130302. https://doi.org/10.1098/rstb.2013.0302

Levshina, N., & Moran, S. (2021). Efficiency in human languages: Corpus evidence for universal principles. *Linguistics Vanguard*, *7*(s3), 20200081. https://doi.org/10.1515/lingvan-2020-0081

Li, X. (2014). Leaning and recipient intervening questions in Mandarin conversation. *Journal of Pragmatics*, *67*, 34–60. https://doi.org/10.1016/j.pragma.2014.03.011

Lin, S., Keysar, B., & Epley, N. (2010). Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology*, *46*(3), 551–556. https://doi.org/10.1016/j.jesp.2009.12.019

Lister, C. J., & Fay, N. (2017). How to create a human communication system. *Interaction Studies*, *18*(3), 314–329. https://doi.org/10.1075/is.18.3.02lis

Louwerse, M. M., Dale, R., Bard, E. G., & Jeuniaux, P. (2012). Behavior matching in multimodal communication is synchronized. *Cognitive Science*, *36*(8), 1404–1426. https://doi.org/10.1111/j.1551-6709.2012.01269.x

Lücking, A., Bergman, K., Hahn, F., Kopp, S., & Rieser, H. (2013). Data-based analysis of speech and gesture: The Bielefeld Speech and Gesture Alignment corpus (SaGA) and its applications. *Journal on Multimodal User Interfaces*, *7*(1–2), 5–18. https://doi.org/10.1007/s12193-012-0106-8

Lücking, A., Ptock, S., & Bergmann, K. (2012). Assessing agreement on segmentations by means of staccato, the segmentation agreement calculator according to Thomann. In E. Efthimiou, G. Kouroupetroglou, & S.-E. Fotinea (Eds.), *Gesture and sign language in human-computer interaction and embodied communication* (pp. 129–138). Springer.

MacNeilage, P. (2008). *The origin of speech*. OUP Oxford.

Macuch Silva, V., Holler, J., Ozyurek, A., & Roberts, S. G. (2020). Multimodality and the origin of a novel communication system in face-to-face interaction. *Royal Society Open Science*, *7*(1), 182056. https://doi.org/10.1098/rsos.182056

Macuch Silva, V., & Roberts, S. G. (2016). Language adapts to signal disruption in interaction. In S. G. Roberts, C. Cuskley, L. McCrohon, L. Barceló-Coblijn, O. Feher, & T. Verhoef (Eds.), *The Evolution of Language: Proceedings of the 11th International Conference (EVOLANG11).* http://evolang.org/neworleans/papers/20.html

Mahowald, K., James, A., Futrell, R., & Gibson, E. (2016). A meta-analysis of syntactic priming in language production. *Journal of Memory and Language*, *91*, 5–27. https://doi.org/10.1016/j.jml.2016.03.009

Manrique, E. (2016). Other-initiated repair in Argentine Sign Language. *Open Linguistics*, *2*(1). https://doi.org/10.1515/opli-2016-0001

Manrique, E., & Enfield, N. J. (2015). Suspending the next turn as a form of repair initiation: Evidence from Argentine Sign Language. *Frontiers in Psychology*, *6*(1326), 215–235. https://doi.org/10.3389/fpsyg.2015.01326

Masson-Carro, I., Goudbeek, M., & Krahmer, E. (2016). Can you handle this? The impact of object affordances on how co-speech gestures are produced. *Language, Cognition and Neuroscience*, *31*(3), 430–440. https://doi.org/10.1080/23273798.2015.1108448

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago Press.

Meier, R. P., Cormier, K. A., & Quinto-Pozos, D. (Eds.). (2002). *Modality and structure in signed and spoken languages*. Cambridge University Press.

Meteyard, L., Cuadrado, S. R., Bahrami, B., & Vigliocco, G. (2012). Coming of age: A review of embodiment and the neuroscience of semantics. *Cortex*, *48*(7), 788–804. https://doi.org/10.1016/j.cortex.2010.11.002

Meulenbroek, R. G. J., Bosga, J., Hulstijn, M., & Miedl, S. (2007). Joint-action coordination in transferring objects. *Experimental Brain Research*, *180*(2), 333–343. https://doi.org/10.1007/s00221-007-0861-z

Mills, G., & Healey, P. (2008). Semantic negotiation in dialogue: The mechanisms of alignment. In D. Schlangen & B. A. Hockey (Eds.), *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue* (pp. 46–53). Association for Computational Linguistics.

Misiek, T., Favre, B., & Fourtassi, A. (2020). Development of multi-level linguistic alignment in child-adult conversations. In E. Chersoni, C. Jacobs, Y. Oseki, L. Prévot, & E. Santus (Eds.), *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* (pp. 54–58). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.cmcl-1.7

Mithen, S. J. (2005). *The singing Neanderthals: The origins of music, language, mind and body*. Harvard University Press.

Mol, L., Krahmer, E., Maes, A., & Swerts, M. (2012). Adaptation in gesture: Converging hands or converging minds? *Journal of Memory and Language*, *66*(1), 249–264. https://doi.org/10.1016/j.jml.2011.07.004

Mondada, L. (2011). Understanding as an embodied, situated and sequential achievement in interaction. *Journal of Pragmatics*, *43*(2), 542–552. https://doi.org/10.1016/j.pragma.2010.08.019

Mondada, L. (2018). Multiple temporalities of language and body in Interaction: Challenges for transcribing multimodality. *Research on Language and Social Interaction*, *51*(1), 85–106. https://doi.org/10.1080/08351813.2018.1413878

Mortensen, K. (2016). The body as a resource for other-initiation of repair: Cupping the hand behind the ear. *Research on Language and Social Interaction*, *49*(1), 34–57. https://doi.org/10.1080/08351813.2016.1126450

Motamedi, Y., Schouwstra, M., Smith, K., Culbertson, J., & Kirby, S. (2019). Evolving artificial sign languages in the lab: From improvised gesture to systematic sign. *Cognition*, *192*, 103964. https://doi.org/10.1016/j.cognition.2019.05.001

Müller, C. (2017). How recurrent gestures mean: conventionalised contexts-of-use and embodied motivation. *Gesture*, *16*(2), 277–304. https://doi.org/10.1075/gest.16.2.05mul

Namboodiripad, S., Lenzen, D., Lepic, R., & Verhoef, T. (2016). Measuring conventionalization in the manual modality. *Journal of Language Evolution*, *1*(2), 109–118. https://doi.org/10.1093/jole/lzw005

Nenkova, A., Gravano, A., & Hirschberg, J. (2008). High frequency word entrainment in spoken dialogue. *Proceedings of ACL-08: HLT, Short Papers (Companion Volume)*, 169–172. https://doi.org/10.3115/1557690.1557737

Norrick, N. R. (1987). Functions of repetition in conversation. *Text - Interdisciplinary Journal for the Study of Discourse*, *7*(3), 2450264. https://doi.org/10.1515/text.1.1987.7.3.245

Oben, B. (2015). *Modelling interactive alignment: A multimodal and temporal account* [PhD Thesis]. KU Leuven.

Oben, B. (2018). Gaze as a predictor for lexical and gestural alignment. In G. Brône & B. Oben (Eds.), *Eye-tracking in Interaction: Studies on the role of eye gaze in dialogue* (pp. 233–262). John Benjamins Publishing Company.

Oben, B., & Brône, G. (2016). Explaining interactive alignment: A multimodal and multifactorial account. *Journal of Pragmatics*, *104*, 32–51. https://doi.org/10.1016/j.pragma.2016.07.002

Okrent, A. (2002). A modality-free notion of gesture and how it can help us with the morpheme vs gesture question in sign language linguistics (Or at least give us some criteria to work with). In R. P. Meier, K. Cormier, & D. Quinto-Pozos (Eds.), *Modality and structure in signed and spoken languages* (pp. 175–198). Cambridge University Press.

Oloff, F. (2018). "Sorry?"/"Como?"/"Was?" – Open class and embodied repair initiators in international workplace interactions. *Journal of Pragmatics*, *126*, 29–51. https://doi.org/10.1016/j.pragma.2017.11.002

Olsher, D. (2008). Gesturally-enhanced repeats in the repair turn communication strategy or cognitive language-learning tool? In S. G. McCafferty & G. Stam (Eds.), *Gesture: Second language acquisition and classroom research* (pp. 109–130). Routledge.

Ortega, G., & Özyürek, A. (2020). Systematic mappings between semantic categories and types of iconic representations in the manual modality: A normed database of silent gesture. *Behavior Research Methods*, *52*(1), 51–67. https://doi.org/10.3758/s13428-019-01204-6

Özer, D., & Göksun, T. (2020). Gesture use and processing: A review on individual differences in cognitive resources. *Frontiers in Psychology*, *11*, 573555. https://doi.org/10.3389/fpsyg.2020.573555

Özyürek, A. (2014). Hearing and seeing meaning in speech and gesture: Insights from brain and behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651), 20130296. https://doi.org/10.1098/rstb.2013.0296

Özyürek, A. (2018). Role of gesture in language processing. In S.-A. Rueschemeyer & M. G. Gaskell (Eds.), *Oxford handbook of psycholinguistics* (pp. 592–607). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198786825.013.25

Pardo, J. S., Urmanche, A., Wilman, S., & Wiener, J. (2017). Phonetic convergence across multiple measures and model talkers. *Attention, Perception, & Psychophysics*, *79*(2), 637–659. https://doi.org/10.3758/s13414-016-1226-0

Peirce, C. S. (1955). *Philosophical writings of Peirce* (J. Buchler, Ed.). Dover.

Perlman, M. (2017). Debunking two myths against vocal origins of language: Language is iconic and multimodal to the core. *Interaction Studies*, *18*(3), 376–401. https://doi.org/10.1075/is.18.3.05per

Perniss, P. (2018). Why we should study multimodal language. *Frontiers in Psychology*, *9*, 1109. https://doi.org/10.3389/fpsyg.2018.01109

Perniss, P., & Vigliocco, G. (2014). The bridge of iconicity: From a world of experience to the experience of language. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651), 20130300. https://doi.org/10.1098/rstb.2013.0300

Perry, M., Breckinridge Church, R., & Goldin-Meadow, S. (1988). Transitional knowledge in the acquisition of concepts. *Cognitive Development*, *3*(4), 359–400. https://doi.org/10.1016/0885-2014(88)90021-4

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529. https://doi.org/10.1073/pnas.1012551108

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*(2), 169–190. https://doi.org/10.1017/S0140525X04000056

Pickering, M. J., & Garrod, S. (2006). Alignment as the basis for successful communication. *Research on Language and Computation*, *4*, 203–228. https://doi.org/10.1007/s11168-006-9004-0

Pickering, M. J., & Garrod, S. (2021). *Understanding dialogue: Language use and social interaction*. Cambridge University Press.

Pouw, W., de Wit, J., Bögels, S., Rasenberg, M., Milivojevic, B., & Ozyurek, A. (2021). Semantically related gestures move alike: Towards a distributional semantics of gesture kinematics. In V. G. Duffy (Ed.), *Digital human modeling and applications in health, safety, ergonomics and risk management: Human body, motion and behavior* (pp. 269–287). Springer. https://doi.org/10.1007/978-3-030-77817-0_20

Pouw, W., Dingemanse, M., Motamedi, Y., & Özyürek, A. (2021). A systematic investigation of gesture kinematics in evolving manual languages in the lab. *Cognitive Science*, *45*(7), e13014. https://doi.org/10.1111/cogs.13014

Pouw, W., & Dixon, J. A. (2020). Gesture networks: Introducing dynamic time warping and network analysis for the kinematic study of gesture ensembles. *Discourse Processes*, *57*(4), 301–319. https://doi.org/10.1080/0163853X.2019.1678967

Purver, M., Eshghi, A., & Hough, J. (2011). Incremental semantic construction in a dialogue system. In J. Bos & S. Pulman (Eds.), *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011)* (pp. 365–369). Association for Computational Linguistics.

Rasenberg, M., Özyürek, A., Bögels, S., & Dingemanse, M. (2022). The primacy of multimodal alignment in converging on shared symbols for novel referents. *Discourse Processes*, *59*(3), 209–236. https://doi.org/10.1080/0163853X.2021.1992235

Rasenberg, M., Özyürek, A., & Dingemanse, M. (2020). Alignment in multimodal interaction: An integrative framework. *Cognitive Science*, *44*(11), e12911. https://doi.org/10.1111/cogs.12911

Rasenberg, M., Rommers, J., & van Bergen, G. (2020). Anticipating predictability: An ERP investigation of expectation-managing discourse markers in dialogue comprehension. *Language, Cognition and Neuroscience*, *35*(1), 1–16. https://doi.org/10.1080/23273798.2019.1624789

Rasmussen, G. (2014). Inclined to better understanding—The coordination of talk and 'leaning forward' in doing repair. *Journal of Pragmatics*, *65*, 30–45. https://doi.org/10.1016/j.pragma.2013.10.001

Ray, M., & Welsh, T. N. (2011). Response selection during a joint action task. *Journal of Motor Behavior, 43*(4), 329–332. https://doi.org/10.1080/00222895.2011.592871

Reitter, D., Keller, F., & Moore, J. D. (2011). A computational cognitive model of syntactic priming. *Cognitive Science*, *35*(4), 587–637. https://doi.org/10.1111/j.1551-6709.2010.01165.x

Reitter, D., & Moore, J. D. (2014). Alignment and task success in spoken dialogue. *Journal of Memory and Language*, *76*, 29–46. https://doi.org/10.1016/j.jml.2014.05.008

Ripperda, J., Drijvers, L., & Holler, J. (2020). Speeding up the detection of non-iconic and iconic gestures (SPUDNIG): A toolkit for the automatic detection of hand movements and gestures in video data. *Behavior Research Methods*, *52*(4), 1783–1794. https://doi.org/10.3758/s13428-020-01350-2

Rizzolatti, G., Fogassi, L., & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience*, *2*(9), 661–670. https://doi.org/10.1038/35090060

Robinson, J. D., & Kevoe-Feldman, H. (2010). Using full repeats to initiate repair on others' questions. *Research on Language and Social Interaction*, *43*(3), 232–259. https://doi.org/10.1080/08351813.2010.497990

Rossi, G. (2015). Other-initiated repair in Italian. *Open Linguistics*, *1*(1), 256–282. https://doi.org/10.1515/opli-2015-0002

Rossi, G. (2020). Other-repetition in conversation across languages: Bringing prosody into pragmatic typology. *Language in Society*, *49*(4), 495–520. https://doi.org/10.1017/S0047404520000251

Ryle, G. (1949). *The concept of mind*. Penguin Books.

Sacheli, L. M., Tidoni, E., Pavone, E. F., Aglioti, S. M., & Candidi, M. (2013). Kinematics fingerprints of leader and follower role-taking during cooperative joint actions. *Experimental Brain Research*, *226*(4), 473–486. https://doi.org/10.1007/s00221-013-3459-7

Sacks, H. (1992). *Lectures on conversation* (G. Jefferson, Ed.). Blackwell.

Safar, J. (2021, September 11). *Communicative repair strategies in Balinese homesign*.

Santamaria, J. P., & Rosenbaum, D. A. (2011). Etiquette and effort: Holding doors for others. *Psychological Science*, *22*(5), 584–588. https://doi.org/10.1177/0956797611406444

Schegloff, E. A. (1979). The relevance of repair to syntax-for-conversation. In G. Talmy (Ed.), *Discourse and syntax* (pp. 261–286). Brill.

Schegloff, E. A. (1991). Reflections on talk and social structure. In D. Boden & D. H. Zimmerman (Eds.), *Talk and social structure: Studies in ethnomethodology and conversation analysis* (pp. 44–70). Polity Press.

Schegloff, E. A. (1992). Repair after next turn: The last structurally provided defense of intersubjectivity in conversation. *American Journal of Sociology*, *97*(5), 1295–1345.

Schegloff, E. A. (1993). Reflections on quantification in the study of conversation. *Research on Language and Social Interaction*, *26*(1), 99–128. https://doi.org/10.1207/s15327973rlsi2601_5

Schegloff, E. A. (2000). When 'others' initiate repair. *Applied Linguistics*, *21*(2), 205–243. https://doi.org/10.1093/applin/21.2.205

Schegloff, E. A. (2004). On dispensability. *Research on Language and Social Interaction*, *37*(2), 95–149. https://doi.org/10.1207/s15327973rlsi3702_2

Schegloff, E. A. (2007). *Sequence organization in interaction: A primer in conversation analysis* (Vol. 1). Cambridge University Press.

Schegloff, E. A., Jefferson, G., & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, *53*(2), 361–382. https://doi.org/10.1353/lan.1977.0041

Schegloff, E. A., & Sacks, H. (1973). Opening up closings. *Semiotica*, *8*(4), 289–327. https://doi.org/10.1515/semi.1973.8.4.289

Schneider, S., Ramirez-Aristizabal, A. G., Gavilan, C., & Kello, C. T. (2020). Complexity matching and lexical matching in monolingual and bilingual conversations. *Bilingualism: Language and Cognition*, *23*(4), 845–857. https://doi.org/10.1017/S1366728919000774

Schubotz, L., Özyürek, A., & Holler, J. (2018). Age-related differences in multimodal recipient design: Younger, but not older adults, adapt speech and co-speech gestures to common ground. *Language, Cognition and Neuroscience*, *34*(2), 254–271. https://doi.org/10.1080/23273798.2018.1527377

Sebanz, N., Bekkering, H., & Knoblich, G. (2006). Joint action: Bodies and minds moving together. *Trends in Cognitive Sciences*, *10*(2), 70–76. https://doi.org/10.1016/j.tics.2005.12.009

Selting, M. (2000). The construction of units in conversational talk. *Language in Society*, *29*(4), 477–517. https://doi.org/10.1017/S0047404500004012

Seo, M.-S., & Koshik, I. (2010). A conversation analytic study of gestures that engender repair in ESL conversational tutoring. *Journal of Pragmatics*, *42*(8), 2219–2239. https://doi.org/10.1016/j.pragma.2010.01.021

Shockley, K., Richardson, D. C., & Dale, R. (2009). Conversation and coordinative structures. *Topics in Cognitive Science*, *1*(2), 305–319. https://doi.org/10.1111/j.1756-8765.2009.01021.x

Sidnell, J. (2007). Repairing person reference in a small Caribbean community. In N. J. Enfield & T. Stivers (Eds.), *Person reference in interaction: Linguistic, cultural and social perspectives* (pp. 281–308). Cambridge University Press.

Sikveland, R. O., & Ogden, R. (2012). Holding gestures across turns: Moments to generate shared understanding. *Gesture*, *12*(2), 166–199. https://doi.org/10.1075/gest.12.2.03sik

Skedsmo, K. (2020). Other-initiations of repair in Norwegian Sign Language. *Social Interaction. Video-Based Studies of Human Sociality*, *3*(2). https://doi.org/10.7146/si.v3i2.117723

Slonimska, A., Özyürek, A., & Capirci, O. (2020). The role of iconicity and simultaneity for efficient communication: The case of Italian Sign Language (LIS). *Cognition*, *200*, 104246. https://doi.org/10.1016/j.cognition.2020.104246

Sterelny, K. (2012). Language, gesture, skill: The co-evolutionary foundations of language. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1599), 2141–2151. https://doi.org/10.1098/rstb.2012.0116

Stivers, T. (2015). Coding social interaction: A heretical approach in conversation analysis? *Research on Language and Social Interaction*, *48*(1), 1–19. https://doi.org/10.1080/08351813.2015.993837

Stivers, T., & Sidnell, J. (2005). *Introduction: Multimodal interaction*. *156*, 1–20. https://doi.org/10.1515/semi.2005.2005.156.1

Stolk, A., Verhagen, L., & Toni, I. (2016). Conceptual alignment: How brains achieve mutual understanding. *Trends in Cognitive Sciences*, *20*(3), 180–191. https://doi.org/10.1016/j.tics.2015.11.007

Streeck, J. (1993). Gesture as communication I: Its coordination with gaze and speech. *Communication Monographs*, *60*(4), 275–299. https://doi.org/10.1080/03637759309376314

Streeck, J. (1994). Gesture as communication II: The audience as co-author. *Research on Language & Social Interaction*, *27*(3), 239–267. https://doi.org/10.1207/s15327973rlsi2703_5

Streeck, J. (2008). Depicting by gesture. *Gesture*, *8*(3), 285–301. https://doi.org/10.1075/gest.8.3.02str

Svensson, H. (2020). *Establishing shared knowledge in political meetings: Repairing and correcting in public*. Routledge. https://doi.org/10.4324/9781003004110

Tabensky, A. (2001). Gesture and speech rephrasings in conversation. *Gesture*, *1*(2), 213–235. https://doi.org/10.1075/gest.1.2.07tab

Tannen, D. (1989). *Talking voices: Repetition, dialogue, and imagery in conversational discourse*. Cambridge University Press.

Török, G., Pomiechowska, B., Csibra, G., & Sebanz, N. (2019). Rationality in joint action: Maximizing coefficiency in coordination. *Psychological Science*, *30*(6), 930–941. https://doi.org/10.1177/0956797619842550

Török, G., Stanciu, O., Sebanz, N., & Csibra, G. (2021). Computing joint action costs: Co-actors minimise the aggregate individual costs in an action sequence. *Open Mind*, *5*, 100–112. https://doi.org/10.1162/opmi_a_00045

Trott, S., & Bergen, B. (2022). Languages are efficient, but for whom? *Cognition*, *225*, 105094. https://doi.org/10.1016/j.cognition.2022.105094

Trujillo, J., Özyürek, A., Holler, J., & Drijvers, L. (2021). Speakers exhibit a multimodal Lombard effect in noise. *Scientific Reports*, *11*(1), 16721. https://doi.org/10.1038/s41598-021-95791-0

Trujillo, J., Simanova, I., Bekkering, H., & Özyürek, A. (2018). Communicative intent modulates production and comprehension of actions and gestures: A Kinect study. *Cognition*, *180*, 38–51. https://doi.org/10.1016/j.cognition.2018.04.003

Trujillo, J., Vaitonyte, J., Simanova, I., & Özyürek, A. (2019). Toward the markerless and automatic analysis of kinematic features: A toolkit for gesture and movement research. *Behavior Research Methods*, *51*(2), 769–777. https://doi.org/10.3758/s13428-018-1086-8

Uskokovic, B., & Talehgani-Nikazm, C. (2022). Talk and embodied conduct in word searches in video-mediated Interactions. *Social Interaction: Video-Based Studies of Human Sociality*, *5*(1). https://doi.org/10.7146/si.v5i2.130876

van Arkel, J., Woensdregt, M., Dingemanse, M., & Blokpoel, M. (2020). A simple repair mechanism can alleviate computational demands of pragmatic reasoning: Simulations and complexity analysis. In R. Fernández & T. Linzen (Eds.), *Proceedings of the 24th Conference on Computational Natural Language Learning* (pp. 177–194). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.conll-1.14

Vesper, C., Abramova, E., Bütepage, J., Ciardo, F., Crossey, B., Effenberg, A., Hristova, D., Karlinsky, A., McEllin, L., Nijssen, S. R. R., Schmitz, L., & Wahn, B. (2017). Joint action: Mental representations, shared information and general mechanisms for coordinating with others. *Frontiers in Psychology*, *7*, 2039. https://doi.org/10.3389/fpsyg.2016.02039

Vesper, C., Butterfill, S., Knoblich, G., & Sebanz, N. (2010). A minimal architecture for joint action. *Neural Networks*, *23*(8–9), 998–1003. https://doi.org/10.1016/j.neunet.2010.06.002

Vesper, C., Morisseau, T., Knoblich, G., & Sperber, D. (2021). When is ostensive communication used for joint action? *Cognitive Semiotics*, *14*(2), 101–129. https://doi.org/10.1515/cogsem-2021-2040

Vesper, C., & Richardson, M. J. (2014). Strategic communication and behavioral coupling in asymmetric joint action. *Experimental Brain Research*, *232*(9), 2945–2956. https://doi.org/10.1007/s00221-014-3982-1

Wagner, P., Malisz, Z., & Kopp, S. (2014). Gesture and speech in interaction: An overview. *Speech Communication*, *57*, 209–232. https://doi.org/10.1016/j.specom.2013.09.008

Wang, Y., & Hamilton, A. F. de C. (2012). Social top-down response modulation (STORM): A model of the control of mimicry in social interaction. *Frontiers in Human Neuroscience*, *6*, 153. https://doi.org/10.3389/fnhum.2012.00153

Wheatley, T., Boncz, A., Toni, I., & Stolk, A. (2019). Beyond the isolated brain: The promise and challenge of interacting minds. *Neuron*, *103*(2), 186–188. https://doi.org/10.1016/j.neuron.2019.05.009

Wittgenstein, L. (1968). *Philosophical investigations*. Basil Blackwell.

Wohlschläger, A., Gattis, M., & Bekkering, H. (2003). Action generation and action perception in imitation: An instance of the ideomotor principle. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *358*(1431), 501–515. https://doi.org/10.1098/rstb.2002.1257

Yap, D.-F., So, W.-C., Yap, J.-M. M., Tan, Y.-Q., & Teoh, R.-L. S. (2011). Iconic gestures prime words. *Cognitive Science*, *35*(1), 171–183. https://doi.org/10.1111/j.1551-6709.2010.01141.x

Yngve, V. H. (1970). On getting a word in edgewise. In M. A. Campbell (Ed.), *Papers from the sixth regional meeting, Chicago Linguistics Society* (pp. 567–578). Department of Linguistics, University of Chicago.

Yurovsky, D., Doyle, G., & Frank, M. C. (2016). Linguistic input is tuned to children's developmental level. In A. Papafragou, D. Grodner, D. Mirman, & J. Trueswell (Eds.), *Proceedings of the 38th Annual Meeting of the Cognitive Science Society* (pp. 2093–2098). Cognitive Science Society.

Zipf, G. K. (1935). *The psycho-biology of language: An introduction to dynamic philology*. MIT Press.

Zlatev, J., Wacewicz, S., Żywiczyński, P., & van de Weijer, J. (2017). Multimodal-first or pantomime-first? Communicating events through pantomime with and without vocalization. *Interaction Studies*, *18*(3), 465–488. https://doi.org/10.1075/is.18.3.08zla

Appendix A

Transcription, coding and inter-rater reliability

Speech transcription

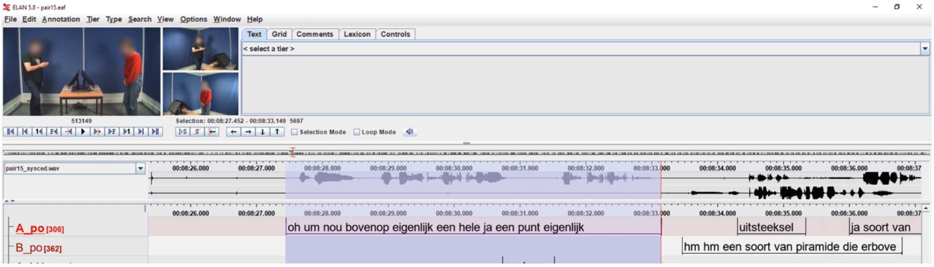
Speech transcriptions of the task-based interactions were used in chapters 3-5. Participants' speech has been segmented and transcribed in ELAN (version 5.8), on two separate tiers: A_po and B_po, where "po" stands for practical orthography (see Figure A1). Speech has first been segmented, where annotations corresponds to turn constructional units (TCUs):

> Each TCU is a coherent and self-contained utterance, recognizable in context as "possibly complete". Each TCU's completion establishes a transition-relevance place (TRP) where a change of speakership becomes a salient possibility that may or may not be realized at any particular TRP. (Clayman, 2013, p. 151)

Consider the highlighted speech annotation of participant A in Figure A1: "oh um nou bovenop eigenlijk een hele ja een punt eigenlijk" [oh um well on top in fact a whole yes a spire in fact]. This is a turn constructional unit that ends in a transition-relevance place, where speaker change indeed occurs. As already becomes clear from the screenshot, spontaneous interaction is complex: speakers can talk at the same time, produce utterances that are contingent on a prior utterance of a partner, repeat or revise their speech, etc. Having high-quality audio recordings from head-mounted microphones helps tremendously when annotating such talk-in-interaction; the speech of one participant can be muted, making it easier to segment and transcribe overlapping speech.

The speech was orthographically transcribed following the standard spelling conventions of Dutch, where the Van Dale dictionary was consulted in case of doubt. This means that pauses and various prosodic features (e.g., speech rate, volume, lengthening etc.) are not included in the transcripts. The following features have been transcribed, mostly following established transcription conventions (Jefferson, 2004), but using various special characters to be able to (semi-)automatically process the transcripts for the analyses:

- Repetitions: e.g., "deze heeft twee twee stokjes"
- Self-repairs: e.g., "aan de zijkant twee ku- kubussen"
- Filled pauses (a.k.a. delay markers): "um" and "uh"
- Reactive tokens: e.g., "hm", "oh", "aha"
- Question prosody: e.g. "en nog een vierkant?" (only used when clearly a question)
- Non-verbal vocal sounds: e.g., #laughs#, #click#, #sigh#
- Task answers: e.g., "dan is het *8*"
- Uncertain transcriptions: e.g., "deze heeft een soort van (bank) aan de zijkant'
- Inaudible speech: e.g., "deze heeft een soort van (?) aan de zijkant"



Selection: 00:08:27.452 - 00:08:33.149 5697

A_po [306]: oh um nou bovenop eigenlijk een hele ja een punt eigenlijk | uitsteeksel | ja soort van

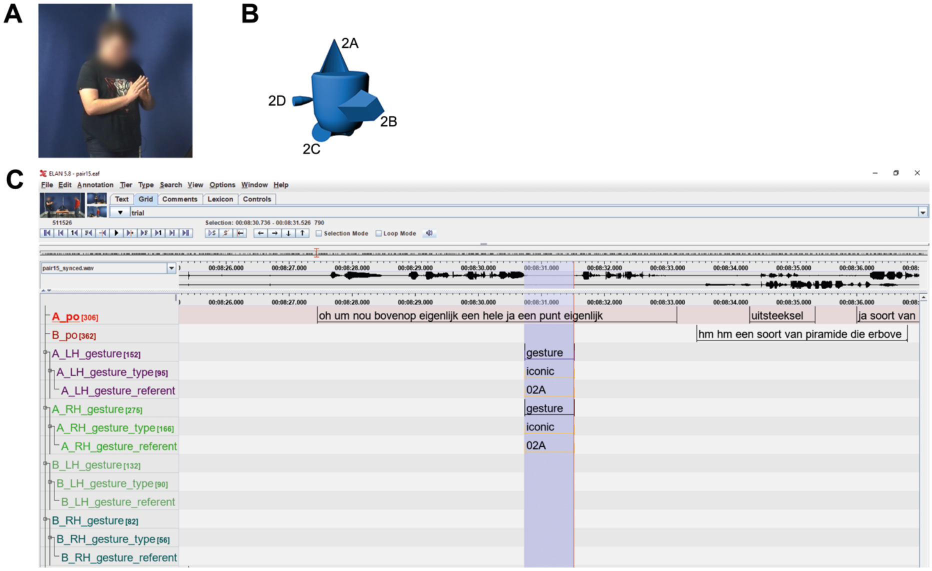
B_po [362]: hm hm een soort van piramide die erbove

Figure A1. Partial screenshot from an ELAN file. The speech tiers include the transcribed speech for the participant on the left (“A_po”) and right (“B_po”; where “po” stands for practical orthography).

Gesture coding

Co-speech gestures annotations were used in chapters 3-5. They were manually annotated using the conventions outlined in Kita et al. (1998; see also e.g., Schubotz et al., 2018). Gestures were annotated in ELAN (version 5.8) for each participant, for the left and right hand separately (see Figure A2, panel C). I focused on manual co-speech gestures, that is, meaningful hand and arm (but not e.g., head) movements. Only the stroke phase of the gesture was annotated, which is the meaningful part of the gestural movement (Kendon, 2004; McNeill, 1992). Gesture segmentation (i.e., defining the on- and offsets of the gesture strokes) was initially performed on the basis of the videos only, without audio. Subsequently, both audio and video were used to check the annotations and for further coding, modifying the boundaries of the annotations when necessary (cf. McNeill, 1992).

For each gesture, the gesture type and referent were coded (on “child” tiers, i.e., tiers that are dependent on the “parent” gesture tier; see Figure A2, panel C). Gestures were categorised into three types: 1) iconic gestures, which depict physical qualities of concrete Fribble referents or movements or actions related to the Fribble referents, 2) deictic gestures, or pointing gestures (with extended finger or hand), and 3) other gestures, which is a heterogeneous group of gestures which do not fit the prior two categories (e.g., beat gestures or interactive gestures). For iconic gestures, the Fribble subpart(s) the gesture referred to was coded, using a pre-defined coding protocol as illustrated for Fribble 2 in Figure A2 (panels A-B). Gesture referents were coded based on the kinematics of the gesture together with the co-occurring speech and overall discourse context. Gestures can refer to one subpart, or to more than one subpart simultaneously. In case of two-handed gestures, both hands can refer to the same subpart(s), or one hand can refer to one subpart, and the other hand to another.



A_po [306]: oh um nou bovenop eigenlijk een hele ja een punt eigenlijk | uitsteeksel | ja soort van
B_po [362]: hm hm een soort van piramide die erbove
A_LH_gesture [152]: gesture
A_LH_gesture_type [95]: iconic
A_LH_gesture_referent: 02A
A_RH_gesture [275]: gesture
A_RH_gesture_type [166]: iconic
A_RH_gesture_referent: 02A
B_LH_gesture [132]
B_LH_gesture_type [90]
B_LH_gesture_referent
B_RH_gesture [82]
B_RH_gesture_type [56]
B_RH_gesture_referent

Figure A2. Panels A-B: a co-speech gesture depicting Fribble subpart 2A. Panel C: partial screenshot from the corresponding ELAN file. It includes speech and gesture tiers. Gestures are annotated for the two participants (A, B), for the left and right hand separately (LH, RH), where gesture type and gesture referent are coded. The annotations in the highlighted time window denote a two-handed iconic gesture, which depicts Fribble subpart 2A, produced by participant A (shown in panels A-B).

Alignment coding

Lexical and gestural alignment has been coded for chapter 3. Below I discuss how they were operationalised, using the dimensions as presented in chapter 2.

Lexical alignment

Lexical alignment was coded per Fribble subpart, using the same referent coding procedure as described for gestures (see the section "Gesture coding" in this appendix). As for the form criterium: we considered words to be aligned if they have the same root form (or "lemma"), so diminutive or plural forms counted as aligned, but synonyms or paraphrases did not (cf., Oben & Brône, 2016). Participants sometimes aligned on multiple words (e.g., both refer to a subpart with "flat nose"), but lexical alignment at the level of the Fribble subparts was computed as a binary variable where alignment of *one* lemma sufficed. The specific categories of words that were included and excluded are listed in Table A1.

Gestural alignment

Prior work has used various form criteria for considering gestures as aligned. For instance, gestures should have the same representation technique (e.g., drawing or handling; Oben & Brône, 2016) and/or the same "overall form" (Bertrand et al., 2013; Holler & Wilkin, 2011a; for a review, see Rasenberg, Özyürek, et al., 2020). Based on an explorative analysis on overlap in gesture form features in our data (as described in the supplementary materials in Appendix B), we have decided to include all referentially aligned gestures, irrespective of the degree of form similarity. The reasoning, in a nutshell, is that for a well-motivated set of basic form features, a great majority of candidate aligned gestures in our data showed similarities on one or more features, making the set of all candidate gestures a reasonable proxy for form-aligned gestures.

Table A1. Categories of lexical items included and excluded in the analysis of lexical alignment

Category	Examples (English translations)
Included	
Shape	*circle, cone, hook, trunk, round, elongated*
Size	*small, big, mini*
Orientation	*upright, diagonal, downwards*
Manner of attachment	*against, through, sticking out, surrounding it*
Similarities/differences between subparts	*two, three, the same, different*
Excluded	
Non-referential speech	meta-speech about the task, such as "*oh we're getting better at this!"*
Highly frequent words*	verbs *to have* and *to be*, as well as most pronouns, determiners and conjunctions
Hedging	*sort of, kind of, little bit, like*
Non-informative speech that applies to all Fribbles	words related to general positions, such as *left, right, on top of;* as well as generic words to describe subparts such as *shape, figure, thing*.

* Frequency was determined on the basis of the SUBTLEX-NL corpus (Keuleers et al., 2010), where we used three standard deviations from the mean lemma frequency as the cut-off for "high frequency".

For a subset of the referentially aligned gesture pairs (for 8 dyads in round 1 and round 2, n = 389 gestures), gestures were coded for their similarity in form. This study is the first to provide such a quantitative analysis of form similarity for a relatively large set of gestures which are related in meaning (see Chui, 2014 for a similar approach with a small sample; and Bergmann & Kopp, 2012 for a large-scale quantification of form similarity for gestures based on their temporal rather than semantic relation). Similarity was coded in terms of five form features: handedness, handshape, movement, orientation, and position in a binary fashion. Coding was done by a trained assistant who was naive to the study's rationale. That is, the coder saw the gesture stroke annotations along with the videos, but had no access to the co- occurring speech or referent coding, and was blind to the selection procedure of the gesture pairs.

Repair coding

Annotations of other-initiated repair sequences were used in chapters 4 and 5. We coded *other-initiated self-repair*, which are practices that interrupt the ongoing course of action to attend to possible trouble in speaking, hearing or understanding the talk (Schegloff, 2000; Schegloff et al., 1977). The coding for this study is a simplified version of the coding scheme by Dingemanse, Kendrick and Enfield (2016). We distinguish the following elements (henceforth "repair turns"):

- TROUBLE SOURCE (T-1)
- REPAIR INITIATION (T0)
- REPAIR SOLUTION (T+1)

For a turn to be considered a T0, it needs to be preceded by a T-1 and followed by a T+1. T0s are coded as being one of three (mutually-exclusive) formats:

- OPEN REQUEST. An expression that requests clarification of a prior turn, leaving open where or what the problem is. Often an interjection or "What?"-like form; typically results in repetition.
- RESTRICTED REQUEST. An expression that requests specification or clarification, restricted to a specific element of the trouble source. Often includes WH-question word and/or repetition.
- RESTRICTED REQUEST. A polar question that offers a candidate understanding and invites confirmation or correction in the next turn. Can include repetition and/or new material.

In order to identify and code other-initiated repair, we rely on the whole multimodal interactional context. That is, besides speech productions (what people say and how they say it), we also rely on participants' co-speech gestures, visual-bodily behaviour (e.g., eye gaze) and the stimulus items during coding.

A T0 may occur on its own or as part of a more extended (non-minimal) repair sequence. For minimal sequences, we coded T-1, T0 and T+1. For non-minimal sequences, we coded only the first/original T-1 and all T0s and T+1s. All T0s and T+1s were incorporated in the analyses.

We created annotations for the repair turns in ELAN (version 5.8), in such a way that they temporally corresponded to the speech annotations. Speech was segmented on the level of turn constructional units (see the section "Speech transcription" in Appendix A). A repair turn could consequently correspond to a single TCU or span multiple TCUs (see examples below). The repair annotations were created in such a way that the boundaries of each repair turn correspond to the onset and offsets of the speech annotations. For information on how gestures were linked to these repair annotations, see the section "Linking gestures to repair annotations" in Appendix A.

Below we present some examples. Underlined text is the speech which temporally overlapped with the gesture stroke; square brackets indicate overlapping speech onsets. Each line corresponds to a single TCU. As such, these examples show that some repair annotations consist of multiple TCUs (e.g., T-1 in Example 1), while others consist of a single TCU (which is a subset of the complete speaker turn, e.g., T-1 in example 2).

Example 1: OPEN REQUEST		
T-1	A (director):	um is een kopje met een n- neus aan de rechterkant en een vierkante piercing er eigenlijk *um is a cup with a n- nose on the right side and a square piercing there actually*
	A:	halve wijnglas *half wineglass*
	A:	en een kleine antenne bovenop *and a small antenna on top*
	B	uh uh uh uh
To	B:	#laughs# wat? *what?*
T+1	A:	dus gewoon één balletje zit aan de voorkant ((gesture)) *so just one ball is on the front ((gesture))*
	B:	j[a *y[es*
	A:	[en die heeft zo'n ((gesture)) eigenlijk wel zo'n ((gesture)) piercing erdoorheen *[and that has such a ((gesture)) actually such a ((gesture)) piercing through it*

Example 2: RESTRICTED REQUEST		
T-1	A (director):	dit is de hoofdvorm waarbij um er rechts ((gesture)) uh een cirkeltje is gewoon zo plat ((gesture)) *this is the main shape where um on the right ((gesture)) uh a circle is just flat (like this) ((gesture))*
	A:	links heb je die vorm die half uitgesneden is met een soort spitse punt erin *on the left you have that shape that is half cut out with a sort of pointed point in it*
To	B	rechts is een ? *on the right is a ?*
T+1	B:	ja een cirkeltje ((gesture)) die eraan vast is geplakt waar je iets aa- op ((gesture)) kan zetten *yes a circle ((gesture)) that is pasted on it where you can put something a- on ((gesture))*

Example 3: RESTRICTED OFFER		
T-1	A (director):	die met die hoek zegmaar ((gesture)) aan de onderkant *the one with that hook so to say ((gesture)) on the bottom*
To	B:	ja die zo ((gesture)) [uh *yes that (like this) ((gesture)) [uh*
T+1	A:	[ja ja ja *[yes yes yes*

Inter-rater reliability

Gesture coding

To establish inter-rater reliability for gesture coding, we focused on the first two rounds of the interaction (where presumably the most (diverse) gestures would occur). Two coders independently coded 96 trials (i.e., 5% of the 1920 trials in the total dataset; and 15% of the trials in round 1 and 2), yielding a comparison of *n* = 296 gesture annotations. Inter-rater agreement on gesture identification was 89.2%. For this measure, we scored how many annotations overlapped, where we disregarded differences in handedness, the length of the annotations and/or the number of segments (e.g., one stroke annotation from one coder spanning two stroke annotations of the other coder). To also assess these aspects of the degree of organisation of the coder's segmentations, we used the Staccato algorithm (Lücking et al., 2012, 2013). We applied this to the left and right hand of each participant separately, which resulted in scores of 0.77, 0.71, 0.80 and 0.75 (on a scale from -1 to 1), indicating that the coders had similar understandings of how the observed gestures had to be segmented. Inter-rater agreement for gesture type was substantial (agreement = 95.1%, Cohen's kappa = .64), and for gesture referent high (agreement = 92.8%, Cohen's kappa = .93).

Gesture form similarity

Inter-rater reliability for gesture form similarity coding (for chapter 5, see the section "Alignment coding" in this appendix) was assessed separately for the five form features (handedness, handshape, movement, orientation and position). Agreement for handedness was computed based on 15% of the initial gesture annotations in rounds 1 and 2 (see above), and resulted in high agreement (agreement = 94.7%, Cohen's kappa = .91). For the other features, a second trained, naive assistant coded 25% of the referentially aligned gesture pairs (*n* = 103) for overlap in handshape, movement, orientation, and position. Substantial agreement was obtained for handshape (agreement = 88.3%, Cohen's kappa = .71) and movement (agreement = 85.4%, Cohen's kappa = .63), and moderate to substantial agreement for orientation (agreement = 75.7%, Cohen's kappa = .54). For position, the score was on the lower side of the moderate category (agreement = 77.7%, Cohen's kappa = .47), and so this feature was excluded from further analyses, which are reported in the supplementary materials of chapter 3 (see Appendix B).

Repair coding

To establish inter-rater reliability for repair coding, two coders independently coded 20% of the complete dataset (384 trials, n = 74 repair comparisons). When inspecting agreement on other-initiated repair identification on the trial level, we found that coders identified the same amount of repair initiations in 94.3% of the trials. The inter-rater reliability for the number of identified repair initiations per trial (ICC = 0.88) was deemed adequate (Koo & Li, 2016). When inspecting agreement on a case-by-case level, we found that in 45 cases the coders agreed on the identification of a repair initiation (i.e., both coders independently coded a turn as an initiation), but disagreed in 29 cases (i.e., one coder considered a turn an initiation while the other did not). So, initial inter-rater agreement on the identification of repair initiations was 60.8%. We examined the underlying pattern of divergence and found that many disagreements were related to a specific category of the coding scheme. These are cases where coding comments indicated that it was hard to judge whether a prior turn of the partner is treated as problematic or whether the participant is instead requesting additional information or demonstrating understanding. For example:

Example (1)

TROUBLE SOURCE A: uh bovenop het kopje heb je een soort driehoek of eigenlijk is het een vierkantje maar je ziet niet helemaal meer het puntje ervan en daar doorheen zit er iets doorheen gestoken

uh on top of the cup you have a sort of triangle or actually it is a square but you cannot fully see the top of it anymore and then through that there is something put through it

REPAIR INITIATION B: ja dus op de bovenkant zit zo'n ruitvormige

yes so on top there is a diamond shaped

REPAIR SOLUTION A: ja

yes

The coders jointly identified these cases in the reliability coding set (relying on the coding comments in ELAN). When excluding these 14 cases from the reliability set, the agreement on the identification of other-initiated repair was 75%. Subsequently, to obtain reliable and systematic coding for cases belonging to this category, we decided to opt for an inclusive approach; i.e., include all of these cases as other-initiated repair whilst (re)coding the data.

For the previous reports on agreement on identification, we scored how many repair initiation annotations overlapped, where we disregarded differences in the length of the annotations and/or the number of segments for trouble source, repair initiation and repair solution. For example, one repair initiation annotation from one coder could span two repair initiation annotations of the other coder, or a trouble source annotation from one coder could consist of two TCUs while the other coder's annotation consisted of only one TCU. For the subset of 45 repair initiations that were identified by both coders, the degree of organisation of the coder's segmentations of the repair sequences were assessed with the Staccato algorithm (Lücking et al., 2012, 2013). This resulted in scores of 0.90, 0.96 and 0.94 (on a scale from -1 to 1) for trouble sources, repair initiations and repair solutions respectively, indicating that the coders had highly similar understandings of how the observed repair sequences had to be segmented. In terms of percentages:

- for trouble sources there was 75,9% complete, 24,1% partial and 6,9% no overlap in the annotations;
- for repair initiations there was 95,6% complete and 4,4% partial overlap in the annotations (no overlap is NA);
- for repair solutions there was 84,4% complete, 13,3% partial and 2,2% no overlap in the annotations.

Inter-rater agreement for repair initiator type (agreement = 84.4%, Cohen's kappa = .58) was deemed adequate. Note that this variable has a skewed distribution, yielding a lower Kappa value despite a relatively high percentage agreement score—known as the "high agreement, low consistency" paradox (Feinstein & Cicchetti, 1990; Feng, 2014). The main discrepancy in coding repair type could be attributed to a difficulty in differentiating two types of restricted formats, as shown by the fact that agreement on *restricted* versus *open* formats was almost perfect (agreement = 97.8%, Cohen's kappa = .88). The divergence within the restricted formats largely resulted from a different understanding of a small set of repair initiations in which the trouble source was (partially) repeated, which affected whether these cases were allotted to *restricted offer* (the largest category to begin with) or *restricted request* (relatively rare). This was resolved through discussion; the coders arrived at a common understanding of these cases as "trouble-presenting repeats" (Dingemanse et al., 2014), which were subsequently (re)coded as restricted offers.

Linking gestures to repair annotations

Since we annotated gestures and repair annotations separately (see the sections above), we still needed to "link" these, in order to quantify the use of gestures in repair sequences in chapters 4 and 5. Note again that the repair annotations corresponded to the speech annotations (which were segmented into TCUs, see the section "Speech transcription"), and thus we will need to identify which gestures correspond to the spoken repair annotations (i.e., trouble source, repair initiation and repair solution). Though in principle it is possible for people to initiate or resolve repair through gesture alone (i.e., without any speech production), we did not encounter this in our dataset.

Co-speech gestures tend to have a tight temporal link to speech, meaning that gestures are usually produced simultaneously with or slightly before the production of the co-expressive speech (ter Bekke et al., 2020; McNeill, 1992; Wagner et al., 2014). For our dataset this entails that when people gesture while initiating or resolving repair, that these gestures are likely to overlap in time with the spoken repair utterances. However, it is also possible for a gesture to only partially overlap with the co-expressive spoken utterance, or to even completely precede or follow a spoken utterance, i.e., to be produced in silence (Healey et al., 2015; Holler & Wilkin, 2011a). Consequently, gestures that correspond to a particular repair annotation might be produced shortly before or after the spoken turn, and should thus also be included in the present study (i.e., linked to the corresponding repair annotation).

Figure A3 visualises which gestures we have included in the dataset. As the general rule, we considered gestures to be part of a repair turn when the gesture stroke completely overlapped with the repair annotation (see row 1 in Figure A3). In case of partial overlap, we included the gestures when the stroke overlapped more than 50% (rows 2 and 3). If the overlap was smaller (<50%), but the stroke *did not* overlap with any prior or next speech, we also included the gesture (rows 4 and 5). In case of small overlap (<50%) where the stroke *did* overlap with a prior or next speech turn, we manually inspected those cases to see which turn was most co-expressive with the gesture, and decided whether or not to link those gestures to the repair annotation (rows 8 and 9). Finally, there might be no overlap at all between the gesture stroke and a repair annotation. For those gestures, if they preceded or followed the repair turn by maximally 2000 milliseconds, and if they *did not* overlap with any prior or next speech, we again manually inspected whether they corresponded to the repair annotation and (de)selected them for inclusion (rows 6 and 7). If they *did* overlap with other speech turns, we excluded them.

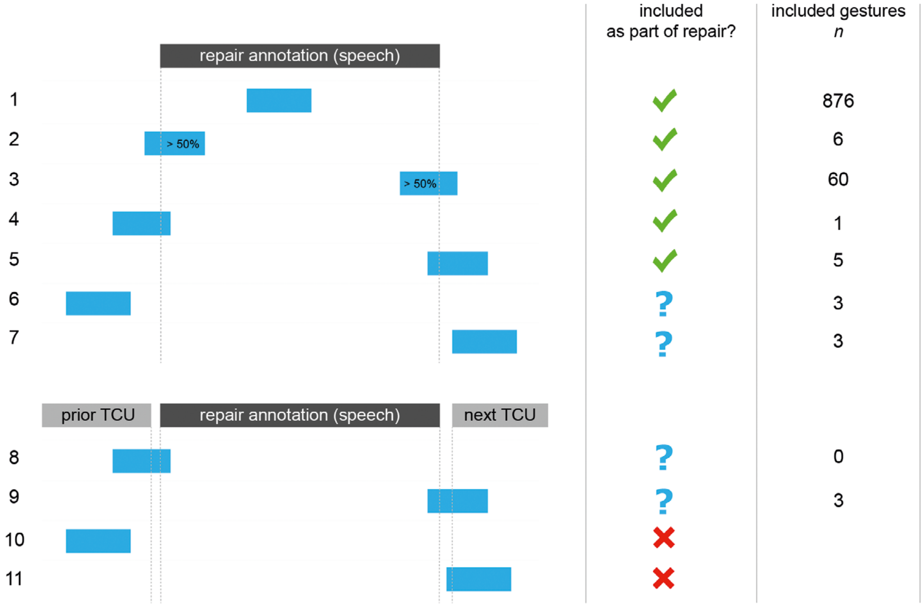


Figure A3. Schematic visualisation of gestures (not) considered to be part of repair annotations

Appendix B

Supplementary materials

Supplementary materials – Chapter 3

Analysis of gesture form similarity

The results of the explorative analyses of gesture form similarity are shown below. As becomes apparent from Figure B1 (panel A), overlap in handedness appears to be most frequent (which naturally follows from the limited degree of freedom: gestures are either left-handed, right-handed or two-handed). Panel B shows that for the number of features that overlap in each gesture pair, overlap in one feature is most common and "complete" form overlap (similar on all four form features) is rare for most dyads. Overall, 80% of all gesture pairs have partial form overlap (similar on one or more features), while only 4% has complete form overlap. Figure B2 shows a combination of the plots in Figure B1: it shows *which* features are most likely to overlap for gesture pairs with a particular number of overlapping features (1, 2 or 3).

In conclusion, based on the fact that the majority of gesture pairs show at least *partial* form overlap,[26] we included all referentially aligned gesture pairs irrespective of their form features.

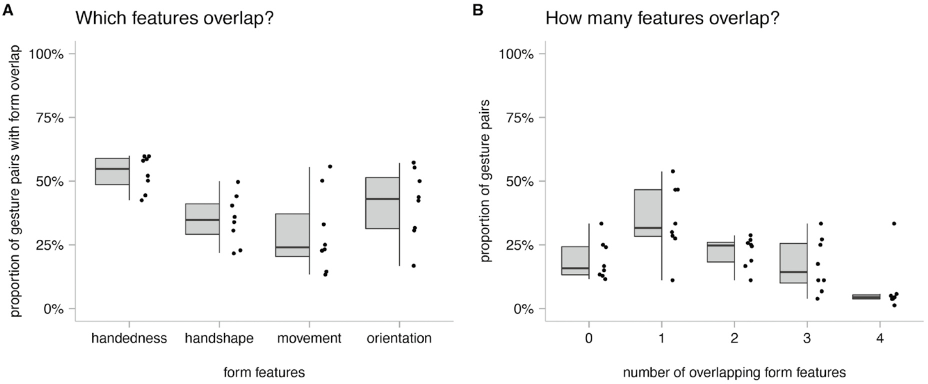


Figure B1. Form similarity of referentially-aligned gestures. Panel (A) shows the relative frequencies with which each form feature overlaps. Panel (B) shows the relative frequencies of the number of features that overlap. Dots represent dyads (*N* = 10).

[26] 80% of all referentially aligned gestures overlap in at least one of the four features considered (handedness, handshape, movement, and orientation), but as many as 94% when also including position, which was excluded due to inter-rater reliability issues (see Appendix A).

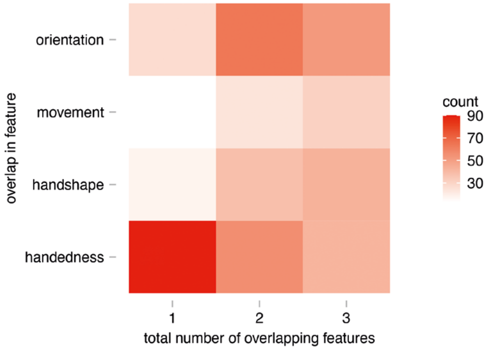


Figure B2. Heatmap displaying *which* features are most likely to overlap for gesture pairs with a particular number of overlapping features (1, 2, or 3).

Relation between alignment and naming similarity scores

In the chapter 3 we state that there is no evident relationship between the degree of lexical and gestural alignment in the interaction and the post naming similarity scores. Here, we present a more detailed inspection of these variables, and the relation between them.

Post-naming similarity

Both before and after the interaction, participants were asked to individually label the Fribbles (target objects) such that their partner could find them. Figure B3 shows the distribution of the naming similarity scores *post* interaction. Though oftentimes dyads provide similar or even exactly the same names, there is also a large amount of naming pairs that have zero similarity after the interaction (we elaborate on this in the section "Shared symbols in the naming task" in chapter 3).

Alignment

First, note that rather than counting the overall frequencies of alignment over the whole interaction, we have specifically tracked the first occurrence of alignment in each modality. That is, we have quantified how often alignment *emerged* in the lexical and/or gestural modality for particular referents (and did not track repeated usage later in the interaction). Second, while naming scores were computed per Fribble (on a scale from 0 to 1), alignment was measured (categorically) for Fribble *subparts*. To be able to relate these variables, we took the relative number of Fribble subparts per Fribble for which alignment emerged as the "degree of alignment" per Fribble. We summed all categories of alignment here (lexical only, gestural only, and multimodal). Including them separately would have resulted in multicollinearity, because the (mutually exclusive) categories are not independent of each other. For example, if for a particular dyad all subparts of a Fribble were grouped in the category multimodal alignment, then it naturally follows that there were zero subparts in the category lexical alignment only. As shown in Figure B4, alignment tended to emerge for almost all subparts (*Median* %).

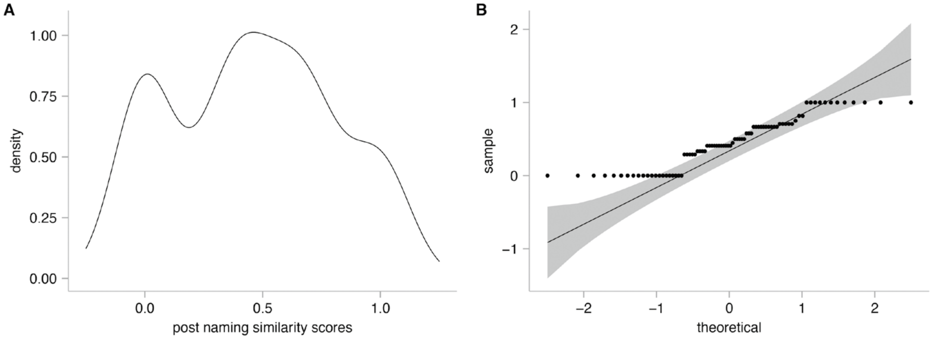


Figure B3. Density plot (panel A) and quantile-quantile plot (panel B) for *post* naming similarity scores (*N* = 80).

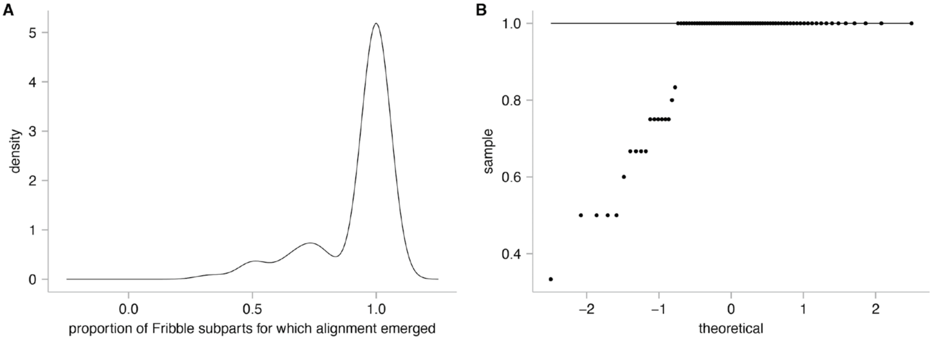


Figure B4. Density plot (panel A) and quantile-quantile plot (panel B) for the proportion of Fribble subparts for which alignment emerged (*N* = 80).

Relation between alignment and post-naming

Figure B5 displays the relation between the relative number of Fribble subparts for which alignment occurred and the *post* naming similarity scores, and shows that there is no evident relationship between the two. The fact that we only measured *emergence* of alignment might explain why we do not find a relation with *post* naming scores, but as becomes clear from Figure B5 this is further complicated by the fact that alignment scores (as measured per Fribble) are near ceiling.

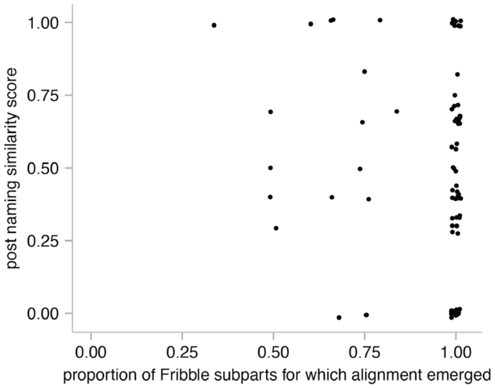


Figure B5. Scatterplot showing the relation between the relative number of Fribble subparts for which alignment emerged and the *post* naming similarity scores (*N* = 80).

Supplementary materials – Chapter 4

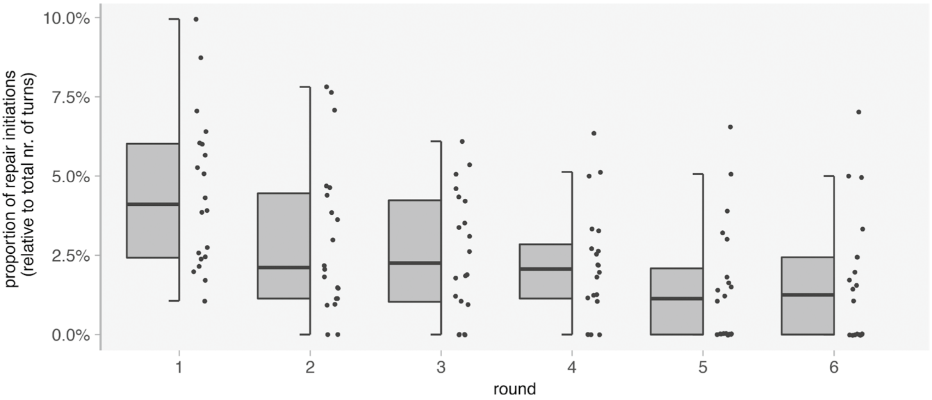


Figure B6. Relative frequency of repair initiations over rounds of the interactive task. Dots represent dyads (*N* = 20). As the task progresses, repair initiations make up a smaller proportion of turns (from an average of 4.5% in round 1, to 1.6% in round 6).

Table B1. Distribution of gesture types (iconic, deictic, other) across repair initiations and repair solutions for different repair formats (open request, restricted request, restricted offer).

	open request (*N* = 24)		restricted request (*N* = 39)		restricted offer (*N* = 315)	
gestures	initiation	solution	initiation	solution	initiation	solution
iconic	-	63 (88.7%)	3 (60%)	64 (95.5%)	202 (94.4%)	98 (81%)
deictic	-	5 (7%)	-	1 (1.5%)	5 (2.3%)	4 (3.3%)
other	1 (100%)	3 (4.2%)	2 (40%)	2 (3%)	7 (3.3%)	19 (15.7%)
total	1	71	5	67	214	121

Supplementary materials – Chapter 5

Operationalisation of gesture effort: more examples

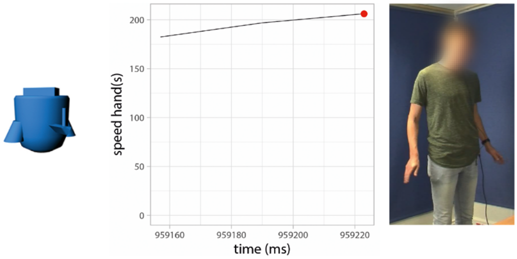


Figure B7. An example of a gesture which is produced by a director as part of a repair solution in response to a restricted offer. The accompanying speech is “ja ja ja” [yes yes yes], thereby confirming the restricted offer which contained a similar gesture. The hands model the subparts on the left (fully visible) and right side (hardly visible) of the Fribble by keeping them still next to either side of his body (number of submovements: 1).

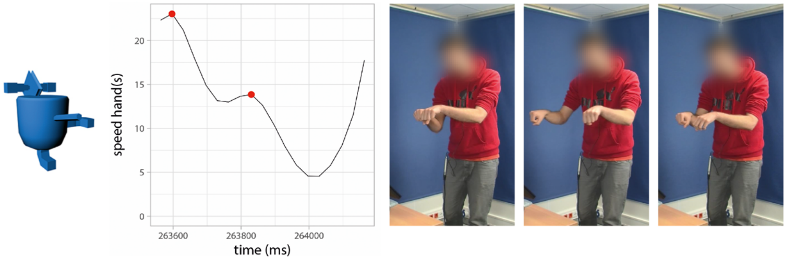


Figure B8. An example of a gesture which is produced by a matcher as part of a repair initiation of the type restricted offer, in response to an unclear turn of the director who described the subparts at the top of the Fribble. The gesture depicts the blocks on the left and right side of the tilted square; the hands start together in the middle, then move sideward and then back to the middle again (number of submovements: 2). The accompanying speech was: “. . . is het een een rechthoek wat in het midden staat waarbij dus uh dat die uh twee objecten met elkaar verbindt?” [. . . is it a a square that stands in the middle where so uh that connects those uh two objects with each other?].

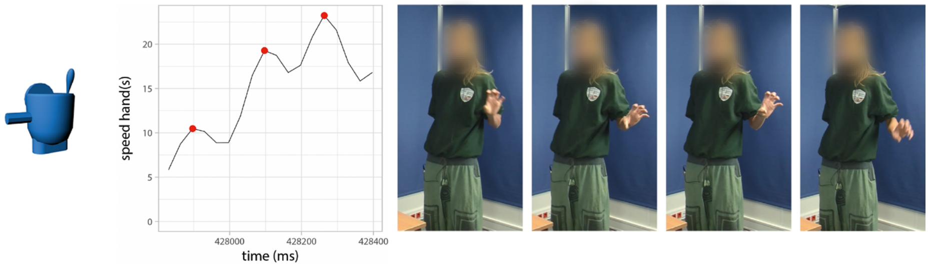


Figure B9. An example of a gesture which is produced by a matcher as part of a repair initiation of the type restricted offer, while saying: “ja redelijk uh blokkerige staaf?” [yes quite uh blocky bar?]. The left-handed gesture depicts the subpart on the left side of the Fribble. The “blocky” aspect is depicted by thrusting the hand sideward in a couple of back-and-forth movements (number of submovements: 3).

Results details

In the main text we have reported the analyses for the multimodal division of effort. Here we elaborate on the potential of analysing the division of effort for the spoken and gestural modality separately. We postulate that, given that people had multiple modalities available to them in which they could express themselves—and the fact that we know that people indeed frequently did so in both modalities—it makes little sense to analyse efforts from a unimodal perspective. For the gestural modality this is further complicated due to the nature of the data. Whereas all repair turns consist of speech (with a minimum of 2 orthographic characters), this is not the case for gestures (across all repair types and sequential positions, the median number of submovements is 0). Thus, we have a smaller set of repair sequences for which we could assess the division of gestural effort at all, and many cases with a highly skewed division of gestural effort (0% versus 100%).

Yet some readers might wonder about the division of verbal effort, and whether those patterns are line with the earlier findings by Dingemanse et al. (2015). Indeed, we too find that the proportional verbal cost paid by the person initiating repair varies as a function of the repair type. The proportional speech effort in the repair initiation was higher for restricted requests compared to open requests ($\beta = 0.17$, $SE = 0.06$, $t = 3.08$, $p = .002$), and higher for restricted offers compared to restricted requests ($\beta = 0.37$, $SE = 0.04$, $t = 9.96$, $p < .001$), as revealed by mixed effects models (with random intercepts for dyads). Thus, this replication shows that the division of labour principle appears to be robust in terms of speech efforts, but it is only now that we have checked the multimodal division of labour for multimodal data that we can take Dingemanse et al.’s (2015) findings to hold water for interactions in their true multimodal form.

Appendix C

Research Data Management

Research data management

Data

For the empirical studies of this thesis (chapters 3-5), data has been collected in 2018 at the Donders Centre for Cognitive Neuroimaging (DCCN) of the Donders Institute, Radboud University Nijmegen. No other datasets have been used for the research reported in this thesis.

Ethical approval and informed consent

This study met the criteria of the DCCN blanket ethical approval for standard studies of the Commission for Human Research Region Arnhem-Nijmegen (CMO 2014/288). Participants were recruited via the Radboud SONA participant pool system. Participants received written information about the study when they signed up, and prior to the testing day they received an email with more detailed information about the study, along with a general information brochure for participants of the DCCN. At the start of the testing session, key information was repeated verbally. Notably, participants were informed that they would take part in an interactive task, and that audio-video recordings would be made. Written informed consent was obtained before data collection started, using two consent forms.

The first consent form was a standardised form for behavioural studies (of the DCCN blanket approval), on which participants agreed to the sharing of the fully anonymised data, and could optionally agree to the sharing of potentially identifiable audio/video data with researchers for scientific purposes. The second consent form was created for this study, and was approved as an amendment to the blanket approval by the Commission for Human Research Region Arnhem-Nijmegen (DCCN CMO 2014/288). On this form, participants could optionally agree to the sharing of audio/video data for educational purposes and/or to promote the research, through a) presentations/lectures (not publicly available), b) newspapers, magazines/journals or other (online) news outlets, c) social media, and d) television.

Data storage

This thesis project is archived in the Donders Repository (https://data.donders.ru.nl/), using three distinct collection types:

- In a Data Acquisition Collection (**DAC**), the data are archived in their original form. Here, original means without any manipulations that limit future analyses of these data.
- A Research Documentation Collection (**RDC**) documents the process via which data are converted into published results.
- A Data Sharing Collection (**DSC**) contains the data that on which published results are based, allowing external researchers to extend scientific findings by reanalysing data with new methods, and/or by addressing new research questions using these data.

An overview of the collections for this thesis is provided in Table C1.

Data sharing

Anonymised data and code. For each empirical chapter in this thesis (chapter 3-5), a Data Sharing Collection (DSC) is created. These DSCs contain (anonymised) data and annotations, as well as the R scripts that were used for processing and analysing the data. These DSCs are (or will be) linked to the (future) publications of chapters 3-5. They are published under a Data Commons license (ODC-ODbL-1.0), meaning that the collections are openly available, ensuring research transparency and reproducibility.

Audio-video data. The audio-video data are stored in private DAC and RDC collections. Note that there are two separate RDCs; RDC 1 contains the complete set of processed data (N = 20 dyads), and RDC 2 contains the data of those participants who consented to their data being shared with researchers that were not involved in the original study (N = 19 dyads). RDC 2 is used to share the data with collaborators when working on future projects.[27] In order for these researchers to get access, they need to sign a custom-made Data Use Agreement, specifying restrictions on data storage and further sharing.

[27] Given the privacy-sensitive nature of the audio-video data, the dataset of this thesis will not be made publicly available. Only the original project team has access and can make RDC2 available to collaborators when working on future projects. However, researchers that are interested in using the data can turn to a similar dataset that has been collected by the Communicative Alignment in Brain and Behaviour (CABB) team (Eijk et al., 2022), for which study-specific data storing and sharing regulations have been devised. The *CABB dataset* (N = 71 dyads) contains data of an interactional task and naming task similar to those reported in this thesis. The dataset is fully documented and archived, and has been made available to the scientific community for research purposes (for instructions on how to access the dataset, see Eijk et al., 2022).

Table C1. Archive of thesis project in the Donders Repository

	Data Acquisition Collection
content	raw data (N = 20 dyads), documentation of set-up, materials and procedure
access	accessible to original project team only (responsible PI: Ivan Toni)

	Research Documentation Collections	
	RDC 1 [internal archive]	RDC 2 [subset for collaborators]
content	processed data (N = 20 dyads), ELAN files, files related to pre-processing, coding	processed data (N = 19 dyads, i.e., those who consented to the sharing of their data with other researchers), ELAN files, files related to pre-processing, coding
access	accessible to original project team only (responsible PI: Ivan Toni)	accessible to original project team (responsible PI: Ivan Toni) and collaborators; subject to Data Use Agreement (specifying restrictions on data storage and further sharing)

	Data Sharing Collections		
	Chapter 3	Chapter 4	Chapter 5
content	speech, gesture and alignment annotations, naming task data, supplementary materials, analysis code	speech, gesture and repair annotations, processing pipeline, analysis code	speech, gesture and repair annotations, motion tracking data, processing pipeline, analysis code
access	publicly available [ODC-ODbL-1.0]	publicly available [ODC-ODbL-1.0]	publicly available [ODC-ODbL-1.0]
DOI	https://doi.org/10.34973/7kbd-5g86	https://doi.org/10.34973/12dp-9q56	https://doi.org/10.34973/jney-n498
status	published	not published	published
linked publication	Rasenberg, M., Özyürek, A., Bögels, S., & Dingemanse, M. (2022). The primacy of multimodal alignment in converging on shared symbols for novel referents. *Discourse Processes*, *59*(3), 209–236. https://doi.org/10.1080/01638 53X.2021.1992235	-	Rasenberg, M., Pouw, W., Özyürek, A., & Dingemanse, M. (2022). The multimodal nature of communicative efficiency in social interaction. *Scientific Reports*, *12,* 19111. https://doi.org/10.1038/s41598-022-22883-w

English summary

People regularly engage in joint actions, such as working together to prepare a meal, play tennis, or move a couch. During such activities, we coordinate our actions to achieve a shared goal. Dialog is no exception. We do not just take turns uttering words, but we work together to make sure that our contributions are understood. How do we do that? In this doctoral thesis I investigated two central phenomena that can help us to achieve mutual understanding: alignment and other-initiated repair. With alignment I mean cross-participant repetition of behaviour, for example when people repeat a word their conversational partner said earlier. Other-initiated repair is when an addressee has trouble perceiving or understanding prior talk, inviting the producer to fix it (e.g., by saying "Huh?"). In this thesis I considered both alignment and other-initiated repair from a multimodal and interactional point of view, that is, by investigating how people coordinate their use of speech and manual gestures in alignment and repair sequences. This brings me to **the main research aim of this thesis**: to contribute towards our understanding of how people work together to negotiate mutual understanding in multimodal interaction.

For this research an interactive task has been developed, in which participant pairs take turns to refer to so called "Fribbles": novel 3D figures that are hard to describe or label. This is a challenging task in which participants work together to achieve mutual understanding, relying on such interactional processes as alignment and other-initiated repair. But before starting empirical investigations of these task-based interactions, some clarifications were in order for the concept "alignment", which has been theorised and operationalised in various ways. I therefore created a conceptual framework (presented in **chapter 2**) that enables us to define and study alignment in a precise manner in multimodal interaction. At the heart of the framework is the decomposition of the multidimensional nature of alignment into five core dimensions: *time, sequence, form, meaning,* and *modality*. Reviewing prior work on lexical and gestural alignment with this framework, I have shown that we should look beyond the "priming" versus "grounding" perspectives dichotomy, and that we can benefit from common terminology to enable cumulative process and principled comparison.

In **chapter 3** I used the framework to study when and how people use alignment of words and manual gestures to establish shared symbols for novel figures, such as the Fribbles. I found that the most frequent strategy is to iconically depict the 3D figures with gestures, and to repeat those gestures as well as accompanying words early on in the interaction. An important finding is that identical repetition is rare, and that variation can be wielded as an interactional resource; people can for example change certain form features of the gesture or add an adjective to a noun to further negotiate the shared understanding of the referent.

In the second half of the thesis I turned to other-initiated repair. While other-initiated repair has been studied extensively with respect to speech as well as some facial signals, here I single out the role of manual co-speech gestures as a domain where more progress can be made. In **chapter 4** I take a holistic look at how gestures are distributed in the repair system as a whole, that is, I studied how gestures are used together with speech to initiate repair (i.e., to target trouble in a prior turn) and to resolve repair (i.e., in response to a repair initiation). Qualitative and quantitative analyses revealed that people often use iconic gestures (in combination with speech) to target trouble in a precise way, or to suggest a candidate understanding—a finding that complements the prior literature on embodied repair initiations, which is mostly centred around the role of facial expressions in *signalling* trouble (without specifying it). Gestures were also effectively used to reply to such repair initiations, either to clarify a prior turn, or to confirm the partner's multimodal candidate understanding by repeating their gesture (a form of gestural alignment).

In **chapter 5** I quantitatively studied how people divide their multimodal efforts across repair initiations and repair solutions, when working together to resolve interactional trouble. I found that people divide their speech and gesture efforts in a way that is predicted by the repair format, where the more specific the repair initiation (e.g., "Huh?" < "On which side?" < "You mean like this ((gesture))?"), the more multimodal effort is invested in the repair initiation relative to the repair solution. Furthermore, I found that people overwhelmingly used the most specific format (offering a candidate understanding) to initiate repair, and that this strategy was most *coefficient*: it requires the least amount of multimodal effort for the dyad as a whole.

These findings characterise alignment and other-initiated repair as interactional strategies which are effective by virtue of their semiotic diversity and flexibility, enabling people to adjust to communicative pressures and constraints. By studying alignment and repair as talk-in-interaction rather than as isolated turns, this thesis helps us to understand what it means to say that social interaction is a form of joint action, where people reach their shared goal of mutual understanding through collaborative and multimodal language use.

Nederlandse samenvatting

Het uitgangspunt voor dit proefschrift is dat mensen in sociale interactie samenwerken, en dat ze voor die samenwerking naast gesproken taal ook andere communicatieve modaliteiten gebruiken, zoals handgebaren. Om dat multimodale proces te onderzoeken kijk ik naar twee fenomenen die veelvuldig voorkomen in alledaagse interacties: interpersoonlijke herhaling (*alignment*) en ophelderingsvragen (*other-initiated repair*). Ik bestudeer hoe interpersoonlijke herhaling en ophelderingsvragen worden ingezet om tot wederzijds begrip te komen, en hoe mensen daarvoor spraak en handgebaren combineren.

Voor dit onderzoek is een interactieve taak ontworpen waarin proefpersonen refereren naar zogenaamde “Fribbles”: onbekende 3D figuren die lastig te omschrijven of benoemen zijn. Dit is een uitdagende taak waarin proefpersonen effectieve gespreksstrategieën nodig hebben om ervoor te zorgen dat ze elkaar begrijpen.

Dit proefschrift resulteerde in verschillende nieuwe inzichten in de multimodale processen die gebruikt worden om tot wederzijds begrip te komen. Ten eerste heerst er binnen het vakgebied behoorlijk wat verwarring rondom het fenomeen interpersoonlijke herhaling. Onderzoekers hebben dit op verschillende manieren bestudeerd en hebben uiteenlopende ideeën over wat de functie ervan is. In hoofdstuk 2 van dit proefschrift creëerde ik een conceptueel kader dat ons in staat stelt om interpersoonlijke herhaling in multimodale interactie precies te definiëren en bestuderen. Vervolgens onderzocht ik in hoofdstuk 3 hoe interpersoonlijke herhaling van woorden en handgebaren wordt ingezet door mensen om tot labels te komen voor nieuwe objecten, zoals de Fribbles. Een veelvoorkomende strategie is om eerst de 3D figuren op een iconische manier uit te beelden met handgebaren, en om interpersoonlijke herhaling van die gebaren tegelijk met herhaling van de bijbehorende woorden te gebruiken.

Vervolgens wendde ik me tot ophelderingsvragen. In hoofdstuk 4 vond ik dat mensen handgebaren (in combinatie met spraak) gebruiken om precies aan te duiden wat ze niet goed verstaan of begrepen hebben, of om een oplossing voor te stellen. Handgebaren worden tevens benut om op een effectieve manier te reageren op dat soort vragen, door middel van multimodale verduidelijking of multimodale bevestiging van een voorgestelde oplossing. In hoofdstuk 5 liet ik zien dat dit soort gespreksproblemen op een efficiënte en coöperatieve manier worden opgelost. Mensen stellen meestal specifieke ophelderingsvragen (bijvoorbeeld met “je bedoelt die zo ((gebaar))?” in plaats van “hè?”), om zo de gezamenlijke, multimodale inspanningen te beperken.

De bevindingen van dit proefschrift geven ons een concreet beeld van wat het betekent om op een multimodale manier samen te werken om tot wederzijds begrip te komen in sociale interactie.

Curriculum Vitae

Marlou Rasenberg was born in 1993 in Made, the Netherlands. She obtained a bachelor's degree in Communication in Information Sciences (cum laude) from Utrecht University in 2015, and a research master's degree in Language and Communication (cum laude) from Radboud University in 2017.

During her bachelor's studies, she worked as a teaching and research assistant at the Utrecht Institute of Linguistics (with dr. Bregje Holleman), and during her master's studies as a statistics tutor and organisational assistant for the LOT Winter School at the Radboud University. For her master's degree, she worked as an intern at the Neurobiology of Language Department of the Max Planck Institute for Psycholinguistics, where she also conducted her master's thesis: an EEG study on the role of discourse markers in predictive processing (with dr. Geertje van Bergen and dr. Joost Rommers).

In November 2017, she began her PhD research in the Multimodal Language and Cognition group of the Radboud University with prof. Asli Özyürek and dr. Mark Dingemanse. She was an active member of the *Communicative Alignment in Brain and Behaviour* team: a large-scale team science project of the Language in Interaction consortium that her thesis project was embedded in. During her time as a PhD student, she also worked part-time as a lecturer at the Department of Language and Communication of the Radboud University, teaching courses and supervising bachelor and master thesis projects.

Publications

Thesis chapters

Rasenberg, M., Pouw, W., Özyürek, A., & Dingemanse, M. (2022). The multimodal nature of communicative efficiency in social interaction. *Scientific Reports*, *12,* 19111. https://doi.org/10.1038/s41598-022-22883-w

Rasenberg, M., Özyürek, A., Bögels, S., & Dingemanse, M. (2022). The primacy of multimodal alignment in converging on shared symbols for novel referents. *Discourse Processes*, *59*(3), 209– 236. https://doi.org/10.1080/0163853X.2021.1992235

Rasenberg, M., Özyürek, A., & Dingemanse, M. (2020). Alignment in multimodal interaction: An integrative framework. *Cognitive Science*, *44*(11), e12911. https://doi.org/10.1111/cogs.12911

Other publications

Dingemanse, M., Liesenfeld, A., **Rasenberg, M.**, Albert, S., Ameka, F. K., Birhane, A., Bolis, D., Cassell, J., Clift, R., Cuffari, E., De Jaegher, H., Dutilh Novaes, C., Enfield, N., Fusaroli, R., Gregoromichelaki, E., Hutchins, E., Konvalinka, I., Milton, D., Rączaszek-Leonardi, J., Reddy, V., Rossano, F., Schlangen, D., Seibt, J., Stokoe, E., Suchman, L. A., Vesper, C., Wheatley, T., & Wiltschko, M. (2023). Beyond single-mindedness: A figure-ground reversal for the cognitive sciences. *Cognitive Science, 47*(1), e13230. https://doi.org/10.1111/cogs.13230

Eijk, L.*, **Rasenberg, M.***, Arnese, F., Blokpoel, M., Dingemanse, M., Döller, C., Ernestus, M., Holler, J., Milivojevic, B., Özyürek, A., Pouw, W., van Rooij, I., Schriefers, H., Toni, I., Trujillo, J., & Bögels, S. (2022). The CABB dataset: A multimodal corpus of communicative interactions for behavioural and neural analyses. *NeuroImage*, *264*, 119734. https://doi.org/10.1016/j.neuroimage.2022.119734

*shared first author

Pouw, W., de Wit, J., Bögels, S., **Rasenberg, M.**, Milivojevic, B., & Ozyurek, A. (2021). Semantically Related Gestures Move Alike: Towards a Distributional Semantics of Gesture Kinematics. In V. G. Duffy (Ed.), *Digital human modeling and applications in health, safety, ergonomics and risk management: Human body, motion and behaviour* (pp. 269–287). Springer. https://doi.org/10.1007/978-3-030-77817-0_20

Rasenberg, M., Rommers, J., & van Bergen, G. (2020). Anticipating predictability: An ERP investigation of expectation-managing discourse markers in dialogue comprehension. *Language, Cognition and Neuroscience, 35*(1), 1–16. https://doi.org/10.1080/23273798.2019.1624789

Acknowledgements

This dissertation came about while working together with many amazing researchers and support staff, as well as through the love and support of my friends and family. I would like to thank you all for your help.

First of all, I would like to thank my supervisors **Asli** and **Mark** for the tremendous amount of support the past years. Thank you for allowing me to work independently, and for making me feel comfortable to express my opinion—even when we disagreed. I've come to see that academic success largely depends on the people that guide you. You have opened up so many doors for me, and helped me walk through those doors with your support and enthusiasm.

Asli, you introduced me to the world of gesture and sign—a wonderful world that has forever changed my view of language. Thank you for always making the time to talk when I needed it and for encouraging me to make my own decisions. I would also like to thank you for creating an inspiring and supportive research environment, and for the many social gatherings you initiated or hosted. It was wonderful to be a part of such a vibrant workplace, and I've always felt very welcome.

Mark, thank you for always being there for me when I needed advice—be it about research or otherwise. You remain an inspiring role model, showing me how to do research with enthusiasm and intellectual rigour. Thank you for your many kind and encouraging words throughout the PhD, and for regularly asking me "Why? I want to know what you think". It was, and remains to be, a pleasure to work with you.

Thank you to the manuscript committee, **Iris**, **Marieke**, **Geert**, **Judith** and **Riccardo** for reading this thesis and for their valuable feedback. Riccardo, thank you for the opportunity to visit you at the Interacting Mind Centre in Aarhus; I came back with many new insights after our interesting discussions and your generous input.

A special thanks goes to the CABB team of the Language in Interaction consortium: **Asli**, **Branka**, **Herbert**, **Iris**, **Ivan**, **Judith**, **Lotte**, **Mark B.**, **Mark D.**, **Mirjam**, **Sara**, **Wim** and others who have joined along the way. Our team science journey has been challenging yet rewarding, and convinced me of the importance of doing interdisciplinary work. Sara, thank you so much for all of your efforts in creating the stimuli and setting up the experiments. Lotte, thank you for always being cheerful, helping us get through the long data collection days with your jokes. Judith, thank you for bringing in your invaluable expertise on gesture and social interaction. Iris and Mark B., thank you for the inspiring theory focus sessions. Thank you, **James**, for your help with setting up the Kinect data collection. And thanks to **Flavia**, **Inez**, **Yvonne** and **Lisette**, for your tremendous help with the 'big dataset' collection and transcription.

A massive thank you to the students who worked with me as assistants or interns. **Emma**, I don't know how you managed to transcribe and code for so many hours and yet remain so cheerful—thank you for your incredible amount of work! **Clarissa**, thank you for putting up with my half-baked ideas and changes of mind, and for figuring out an analysis plan together. **Junfei** and **Vicky I.**, thank you for your help with coding gestures and gesture similarity.

Although my research project was part of the CLS, I was supported by many people from the MPI (where I had my office) and the DCCN (where I collected data). At the my MPI I greatly benefitted from the help of the technical group. **Jeroen**, thank you for your help with setting up the audio-video set-up, and for your help with processing the data. **Maarten**, thank you for your help with the Kinect synchronisation. **Alex**, thank you for patiently helping me with my many software and hardware requests. I am also thankful for the practical help of **Jan** and **Robèr**, the lunches provided by **Thea**, and the helpful responses to my questions about books and open access by **Karin** and **Meggie**. Many thanks also go to **Kevin**, for the fun and interesting IMPRS activities, and for providing a listening ear. At the DCCN I would like to thank **Uriel** for installing the complex lab set-up – we never would have managed to collect the data there without you. My gratitude also goes to **Nicole**, **Mirjam**, **Jessica**, **Sabine** and **Steven** for their help with consent forms, administration and data archiving, and to **Mora** for enabling me to start my testing days with a joyful greeting and a cappuccino. Thanks to **Julia** and **Wendy**, for organizing the Language in Interaction activities and for your support to the PhD students. Fortunately at the CLS I had someone check in on me from time to time as well: **Peter**, thank you for making me feel comfortable to be completely honest with you during our GSH progress report meetings, and for organizing the GSH activities.

I would like to thank all past and present members of MLC and MLD: **Asli**, **Andrea**, **Anita**, **Beyza**, **Chiara**, **Christina**, **Clarissa**, **Dilay**, **Els**, **Emma**, **Erce**, **Ezgi**, **Francie**, **Gerardo**, **Huku**, **James**, **Jessica**, **Junfei**, **Kazuki**, **Kimberley**, **Lilia**, **Linda**, **Louise**, **Maddy**, **Marlijn**, **Martina**, **Patrick**, **Pelin**, **Renske**, **Sevilay**, **Sho**, **Tom**, **Uğur**, **Vicky**, **Wim** and **Zeynep**. I have been so fortunate to learn from all of you, and I have thoroughly enjoyed our lab meetings, lunches and spontaneous chats. I remember very well returning to the office after the covid lockdowns, and realizing how much I had missed our stimulating interactions. Working from home took a toll on me, but with every day in the office I felt my excitement for research coming back. Thank you for your support and feedback, and for all the fun times and delicious food. **Tom**, bedankt voor de NGT lessen en de gezelligheid!

A special thank you goes to my paranymphs: Marlijn, Anita and Wim. Thank you for all the defence preparations, and for your support throughout my PhD. **Marlijn**, thank you for being such an awesome friend and colleague. You were always there for me to reassure me, or to help me out with a graph, title or statistics. **Anita**, thank you for letting me barge into your office with my worries and questions, and for being there alongside me during both the fun and stressful times while finishing up our theses. **Wim**, your excitement for research is contagious and I've been lucky enough to catch some of that. I hope one day I'll be a researcher who is just as versatile and independent as you (preferably with the same amount of energy because I wonder if you ever sleep with all the projects you've got going on!). Thank you for your help with the Kinect analyses, it was a pleasure to work with you.

I'm grateful to have met many more bright and kind people during my PhD. **Francie**, thank you for always being there for me when I needed some advice. **Teun**, thank you for the stimulating discussions during our long coffee breaks and for initiating Friday-drinks. **Naomi**, thank you for the fun times and the sincere conversations. **Ryan**, I still find it funny how we met and became friends – thank you for accepting that super random coffee invitation that one morning at the MPI. **Elena** and **Margot**, thank you for our weekly "writing" sessions that helped me get through the last months of my PhD. **Konstantinos** and **Sebastian**, thank you for being my friends when I just moved to Nijmegen. **Ilona**, thank you for the lovely walks and dinners, and for letting me stay over so many times before I moved. **Hanno**, thank you for inspiring me to make time to chill and go out into nature, and for our many open and honest conversations. **Marieke**, thank you for your support and kindness; I am so happy I got to know you better in the last year of my PhD. And thank you to those who brightened my days at work, Language in Interaction events, or the Cultuur Café: **Ine**, **Tiziana**, **Marco**, **Alessio**, **Silvia**, **Ezgi**, **Louise**, **Limor**, **Sara**, **Joe**, **João**, **Guilherme**, **Ionna**, **Filiz** and many others.

Aan mijn vrienden buiten academia heb ik ook veel te danken. Te beginnen met mijn lieve oud-huisgenoten **Jelle**, **Maarten**, **Loek**, **Daan**, **Judith** en **Jet**: wat fijn dat jullie er voor mij waren toen ik begon aan mijn PhD avontuur. Dank ook aan **Lysanne, Leonie, Andy, Felicia, Michiel, Sam** en **Anne** voor een fijn thuis in Nijmegen waar ik volledig mezelf kan zijn. En **Merel**, wat was het fijn om thuis te komen en jou zingend in de keuken aan te treffen. Ik hoop dat je nog vaak komt aanwaaien. **Bas**, wat een geluk om zo'n loyale vriend in mijn leven te hebben. Bedankt voor de onvergetelijke hike- en kayak avonturen. En dan de GR5: bedankt voor al je hulp met de voorbereidingen toen ik door de bomen het bos niet meer zag aan het eind van mijn Phd. Ik ben je ongelofelijk dankbaar dat je me die droom hebt helpen realiseren. Tevens wil ik mijn lieve vriendinnen bedanken die me door dik en dun gesteund hebben tijdens mij PhD, in het bijzonder Lisane, Elke en Sophie. **Lisane**, bedankt dat je altijd je leuke, gekke zelf bent en me eerlijk vertelt wat je denkt. **Elke,** wat fijn dat we ondanks de afstand elkaar toch regelmatig kunnen zien en dan écht kunnen praten. Merci voor alles. En aan **Alex**: heel erg bedankt voor de cover, ik ben er super blij mee! **Sophie**, ik weet niet waar ik moet beginnen, want als er iemand mee heeft gekregen wat voor rollercoaster de afgelopen vijf jaar zijn geweest dan ben jij het wel. Wat voelt het goed om te weten dat jij altijd voor mij klaar staat.

Aan mijn ouders, **Jos** en **Leny**, en aan **Bram** en **Marscha**: bedankt voor jullie goede zorgen en alle steun de afgelopen jaren. Het was altijd fijn om weer even thuis te kunnen komen.

Mouaz, you made me experience the feeling of ‘someone being there for me’ on a whole new level. Thank you for supporting me big time towards the end of my PhD—I think I’ve never worked so hard, yet it was so much fun because I got to do it with you by my side. It was hard, but I mostly remember us having so much fun.

MAX PLANCK INSTITUTE
FOR PSYCHOLINGUISTICS

VISITING ADDRESS
Wundtlaan 1
6525 XD Nijmegen
The Netherlands

POSTAL ADDRESS
P.O. Box 310
6500 AH Nijmegen
The Netherlands

CONTACT
T +31(0)24 3521 911
F +31(0)24 3521 213
E info@mpi.nl
Twitter @MPI_NL
www.mpi.nl

CLS | Centre for Language Studies
Radboud University