Alignment in multimodal interaction: an integrative framework

Marlou Rasenberg[1,2], Asli Özyürek[1,2,3] & Mark Dingemanse[1,2] for the CABB team[4]*

*Corresponding author: marlou.rasenberg@mpi.nl
[1]Centre for Language Studies, Radboud University, Nijmegen, The Netherlands
[2]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
[3]Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, the Netherlands
[4]Language in Interaction Consortium, The Netherlands

Abstract

When people are engaged in social interaction, they often repeat each other's communicative behavior, such as words or gestures. This kind of *alignment* has been studied across a wide range of disciplines and has been accounted for by diverging theories. In this paper, we review various operationalizations of lexical and gestural alignment. We reveal that scholars have fundamentally different takes on when and how behavior 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 behavior overlap in form), yet underspecify, conflate or neglect other dimensions. This stands in the way of proper 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 formalize the relationship between any pair of behavior: *sequence*, *time*, *semantics*, *form* and *modality*. We show how assumptions regarding the underlying mechanism of alignment (categorized into *priming* versus *grounding*) pattern together with the operationalization in terms of the five dimensions. This conceptual framework can help researchers in the field of alignment and related phenomena (including behavior matching, mimicry, entrainment and accommodation) to formulate their hypotheses and operationalizations in a more transparent and systematic manner. The framework also enables us to discover unexplored research avenues and derive new hypotheses from existing theories.

Keywords: social interaction; alignment; mimicry; behavior matching; accommodation; entrainment; co-speech gestures

1. Introduction

In social interactions, speakers coordinate their actions in an effort to incrementally and interactively reach their communicative goals. A pervasive component of such joint actions is cross-participant repetition of communicative behavior. Work across a wide range of fields shows that when people are engaged in interaction, their behaviors 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 behavioral *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 behavior 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 behaviors 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 Cordon (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]$_{A}$ Chad?
3 C: Yeah the meat [eating]$_{B}$ Chad guy

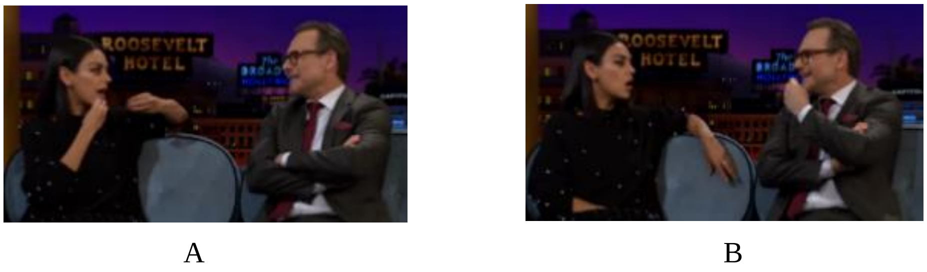


Fig 1. Alignment of speech and gestures produced by the talk show guests Mila Kunis and Christian Slater.[1]

Square brackets indicate the start and end points of a gesture, and the capital letters correspond to the pictures shown in Fig. 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.[2] 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.,

[1] Stills are taken from the video “Mila Kunis and Christian Slater are addicted to dating shows” by The Late Late Show with James Cordon, 2018 (https://www.youtube.com/watch?v=B2Pcc_CSaK4, 2:40-2:41).

[2] 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.

Holler & Wilkin, 2011; Pickering & Garrod, 2004). To narrow down the scope of this article, while still adopting a multimodal perspective on alignment, we will zoom in on lexical and gestural alignment.

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. Intuitively speaking, we can easily identify alignment here, since both participants produce the same lexical phrase ("meat-eating Chad") and gestures which look highly similar. This focus on alignment of *form* ties in with the traditional notion of 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 would 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 would restrict analyses to alignment in speech *or* gestures, while others would look at both. Some would only focus on these cases of alignment in adjacent speech turns, while others also look for alignment of behavior which are temporally further apart. Design choices and measurement techniques tend to 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 behavior in interactions.

Given the diversity of work in the interdisciplinary area of social interaction, some notes on terminology are in order. First, alignment in the narrow sense used here refers to observable similarities in communicative behavior. Scholars sometimes use alignment in a wider sense, as they use these observable similarities in behavior to make inferences about people's conceptual representations. In this paper we generally adhere to the narrow sense of the term, and will explicitly mention it when we move to the wider sense (e.g., when addressing the *semantic* relation between behaviors). Second, what we call alignment here has been studied under a range of terms, and is part of a larger array of phenomena variously labelled behavior matching, entrainment, accommodation, repetition, imitation, and mimicry. Though all of these terms target contingent behavioral similarities in socially interacting agents, each of them comes with its own disciplinary history and therefore carries its own commitments and implications with regard to the kinds of behavior in focus, the embodied and interactional mechanisms at play, and the cognitive or socio-affective functions involved. A key goal in this paper is to set out clear terms of comparison that will enable cumulative progress and principled comparison.

With many fields now working towards empirical and theoretical accounts of alignment, it is crucial to have a shared conceptual framework that allows us to capture the space of possibilities of what can be considered alignment. By systematically tracking five dimensions along which communicative behaviors 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 operationalizations tend to pattern together with assumptions regarding underlying mechanisms (for instance, *priming* versus *grounding*), resulting in a situation where some areas of the space of possibilities are much better explored than others. We single out the interrelation of lexical and gestural alignment as one of the promising areas for future studies.

2. Two Theoretical Approaches: Priming and Grounding

Today, we can distinguish two broad theoretical approaches to alignment in dialog, which we will call *priming* and *grounding* (cf. Oben, 2018; also denoted automatic versus strategic alignment (Kopp & Bergmann, 2013); and related to the distinction between Aggregate and Interactive approaches (Healey, Mills, Eshghi, & Howes, 2018)). Priming approaches postulate that alignment comes about through an automatic and non-intentional process, which requires little cognitive resources. Examples are *ideomotor theory* in cognitive psychology (for a review, see Shin, Proctor, & Capaldi, 2010) or the *perception-behavior link* in social psychology (e.g., Dijksterhuis & Bargh, 2001). This work builds on the premise that the observation of a certain action (such as foot waggling or a finger movement) automatically leads to the activation of the corresponding motor plan, which increases the likelihood or ease of subsequently producing the same behavior (Genschow et al., 2017). Within psycholinguistics, a related and highly influential proposal has been the *interactive alignment account* by Pickering and Garrod (2004). According to this account, there is a parity between the linguistic 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), makes listeners likely to subsequently use it in their own speech production as well. This 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 higher-level representations or *situation models*. Though Pickering and Garrod's account does not suppose the involvement of motor activation, it clearly corresponds to the earlier mentioned psychology theories; both make a case for priming as an automatic process confined to the cognitive system of one individual.

The interactive alignment account, as originally conceived (Pickering & Garrod, 2004), is restricted to the spoken modality, and has been corroborated by experimental, task-based studies on lexical and syntactic alignment (e.g., Branigan, Pickering, & Cleland, 2000; Garrod & Anderson, 1987). However, the theoretical line of reasoning could be extended to include nonverbal behavior. For example, Louwerse, Dale, Bard and Jeuniaux (2012) analyzed the multimodal interactions of participants engaged in a route communication task (Map Task; cf. Anderson et al., 1991). They investigated cross-speaker imitation of verbal, facial and gestural behavior (“behavior matching”), as well as the temporal dependencies of those matched behaviors (“synchronization”). They found that participants exhibited a significant degree of temporal organization in different types of matched behavior, including deictic pointing gestures, smiling and the use of digits in descriptions. These synchronization effects increased the longer people talked to each other and the more difficult the task became. The authors argue that this synchronization of behavior matching is a low-level and low-cost resource for communication.

Whereas priming accounts thus argue for an automatic and non-intentional process, grounding accounts emphasize the interactive, coordinative efforts underlying alignment. For example, Bavelas (2007) has argued that “motor mimicry” (e.g., of facial expressions) is not an automatically triggered response, but instead an intentional, social act to communicate understanding to a communicative partner. Within the field of psycholinguistics, grounding approaches have their roots in the work of Clark and colleagues, who have studied the interactive, coordinative efforts underlying lexical alignment. Producing and comprehending lexical phrases is not enough for communication; interlocutors try to make sure that what has been said has been understood. That is, they actively and incrementally *ground* their contributions, using such interactive tools as feedback and repair (Clark & Wilkes-Gibbs, 1986). Once a shared conceptualization has been established, interlocutors can repeatedly refer to these *conceptual pacts* with the same words (Brennan & Clark, 1996), thus yielding sustained lexical alignment (denoted “entrainment”).

Though this earlier work of Clark and colleagues was solely focused on speech, more recently it has also been applied as a conceptual framework in studies on gestural alignment. For example, Holler and Wilkin (2011) analyzed the iconic and metaphoric co-speech gestures of participants engaged in a referential communication task (cf. Clark & Wilkes-Gibbs, 1986). In this task, the participants (playing the roles of Director and Matcher), have to refer to cards with geometrical figures (tangrams). They find that gestural “mimicry” occurs in these interactions, and show how it serves various communicative functions, such as asserting acceptance or signaling that a reference has been understood. The authors subsequently conclude that mimicry of gestures plays a core role in the process of grounding. Furthermore, certain gestures appeared to be used persistently over the course of the interaction to depict a particular figure (that is, there was gestural “entrainment”), as they were part of a conceptual pact (cf. Brennan & Clark, 1996). The authors regard this kind of gestural mimicry as intentional and conscious in nature, which is directly involved in the establishment of mutual understanding between interlocutors.

These two theoretical accounts (denoted priming and grounding from here onwards) thus have a different focus, which results in diverging assumptions, predictions and methodological approaches – making them hard to reconcile. To illustrate this point, take again the studies by Louwerse et al. (2012) and Holler and Wilkin (2011). At first glance, the studies seem to report fairly similar results, namely, that there is cross-speaker repetition of co-speech gestures in a face-to-face task-based interaction (be it for iconic/metaphoric gestures depicting tangram figures in Holler & Wilkin, 2011, or deictic pointing gestures in a Map Task in Louwerse et al. 2012). However, on closer inspection it turns out these are quite different observations. In the study by Holler and Wilkin (2011) iconic or metaphoric gestures are considered to be "mimicked" when they represent the same meaning and have some similarity in their form. For Louwerse et al. (2012), "synchronized matching behavior" for gestures boils down to the following: both interlocutors used a gesture of the category *deictic* (i.e., pointing gestures), with an average time interval of 25 seconds. This means that, in contrast to Holler and Wilkin (2011), interlocutors are considered to match or mimic each other's gesture, even when they point to different locations on the map and thus refer to different things. Louwerse et al. thus consider the act of two people pointing to be an instance of gestural matching behavior, whereas for Holler and Wilkin mimicry means that the interlocutors' gestures also represent the same meaning and refer to the same entity.

The respective theoretical frameworks seem to tie up with the results of these studies, but it is unclear whether this would also have been the case if the empirical approaches had been different. That is, could an automatic, non-intentional priming mechanism account for the kind of alignment Holler and Wilkin found in gestures that refer to complex, novel objects? And vice versa, could a more intentional grounding process explain why interlocutors are likely to align and synchronize certain kinds of behavior (such as pointing), irrespective of the meaning of that behavior? This highlights the need to look closely at the operationalization of what counts as "mimicry" and "matching behavior". Moreover, it is important to be aware of the extent to which such choices (implicitly) follow from the theoretical assumption and research traditions of the researchers.

3. A Framework for Understanding and Investigating Alignment

In order to arrive at a comprehensive understanding of the literature on alignment, we have to outline the space of possibilities of how alignment is conceptualized and measured across studies. Generally speaking, all studies on alignment compare an instance of behavior from person A with an instance of behavior from person B, which are considered to be "the same" or "matched" in one way or the other. That is, the units of analysis are cross-speaker *paired behaviors*, where A's behavior is similar to B's behavior on one or several dimensions. Though *prime-target pairs* is a more commonly used term in the field, we intentionally use the more neutral term paired behaviors here, as to remain agnostic to the origin or mechanism behind the pairing.

Variation exists among empirical studies with respect to the dimension(s) which are taken into consideration, and how it is operationalized. Generally speaking, most studies use similarity in form as a criterion for alignment, though there is plenty of variation in the definition of form overlap and the way it is measured. However, the relation between the two instances of behavior on other dimensions is often taken for granted or simply neglected. 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 operationalize experimental studies accordingly.

In an effort to clarify and reveal (often times implicit) differences in what is considered to be aligned, we introduce a common conceptual framework to decompose the notion of alignment into its constituent dimensions. We consider five key dimensions on which the relation between any pair of behavior can be characterized: *sequence*, *time*, *semantics*, *form* and *modality*. The framework is presented in Table 1, where we outline the dimensions in clear and generic terms which are applicable to all kinds and levels of verbal and non-verbal behavioral alignment, be it posture or gesture, phonetics or syntax. For illustration purposes, we use rectangular shapes as instances of behavior, which are produced by two interlocutors (A and B), as shown in Fig. 2.

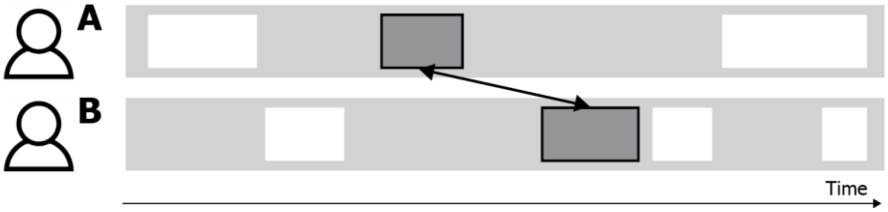
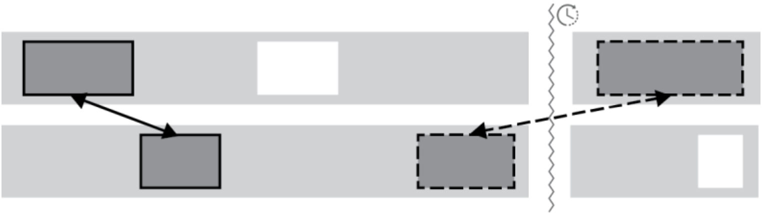
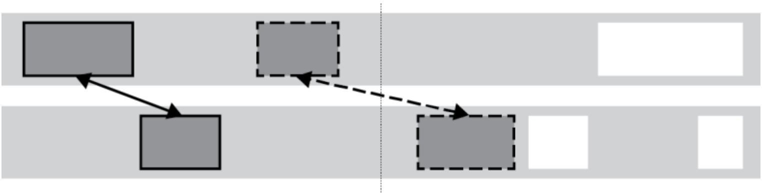
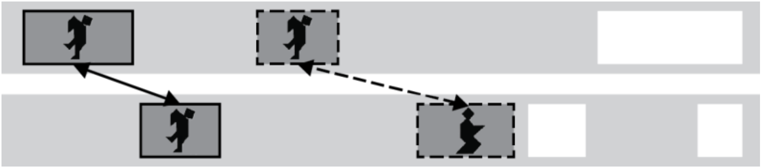
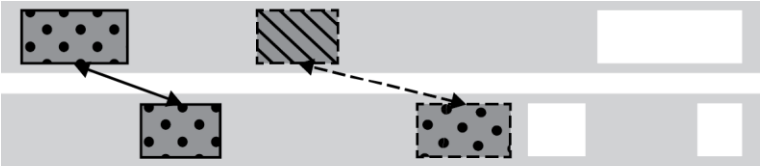
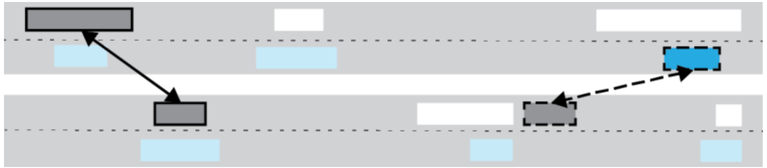


Fig. 2. Visualization of an interaction between two people. Every rectangle represents an instance of behavior. The behavior can be of various types (i.e., the rectangles could represent syntactic constructions, lexical choices, mannerisms etc.) and levels (e.g., the rectangles could represent complete speech turns or individual gestures). The arrow indicates a possible comparison between two instances of behavior.

The dimensions of this framework are distinct concepts, though not completely independent of each other. For example, when two instances of behavior occur within a certain *sequence* (e.g., in adjacent turns), this naturally has consequences for the dimension *time* (i.e., the two pair parts are likely to have only a short temporal lag). And when two instances of behavior are produced in different *modalities* (e.g., a lexical phrase is compared to an iconic gesture), the dimension *form* will become less relevant. Due to these interrelations, certain dimensions can become conflated or taken for granted in both empirical and theoretical approaches. Hence, we will argue here that it is crucial for all work on alignment to treat the dimensions as conceptually distinct from each other, and specify the relationship between two instances of behavior for each dimension separately. We will corroborate this in Section 4, in which we apply the framework to studies on lexical and gestural alignment.

Table 1. A multidimensional framework for understanding and investigating alignment[a]

Sequence	The sequential relation between any pair of aligned behavior can vary from occurring within a certain sequence (e.g., the behavior occurs in adjacent turns or within the same trial), to transcending such sequential boundaries.	
Time	The temporal distance between the first and second part of a pair of behavior 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).	
Semantics	For levels of behavior which convey semantic meaning (e.g. lexical items or gestures), any pair of behavior can vary from conveying the same meaning or referent to conveying different meanings.	
Form	The two parts of a pair of behavior can vary from being exact copies, to having little or no overlap in form or shape.	
Modality	The two parts of a pair of behavior 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 behavior-pair can vary on five dimensions, outlined in this table. For each dimension we visualize two different relationships between instances of behavior; one with a solid arrow and one with a dashed arrow. For the semantics dimension, we use tangram figures to visualize the meaning of speech and/or gestures (cf. Clark & Wilkes-Gibbs, 1996; Holler & Wilkin, 2011).

4. A Review Based on the Framework

This section will illustrate how we can use the five dimensions introduced in the previous section to characterize and compare 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 operationalizations relate to the two theoretical approaches (priming and grounding). By doing so, we explore the space of possibilities; we will show which dimensions are of fundamental importance in various empirical and theoretical accounts, and which dimensions are understudied or neglected. 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 behavior, making the current discussion of relevance to the field as a whole.

We will restrict our focus to studies investigating spontaneous, interactive dialogs (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, Krahmer, Maes, & Swerts, 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 dialog acts, 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.

4.1 Sequence

All interactions have a sequential structure. Conversations are comprised of *adjacency pairs*; pairs of utterances where the latter is functionally dependent on the first, such as offer-acceptance or question-answer (Schegloff & Sacks, 1973). On a higher level, one or several of such pairs together can constitute a *project* or a *course of action* (Levinson, 2013; Schegloff, 2007), such as scheduling a meeting. In addition to these naturally emerging sequential structures, task-based interactions also have an experimentally-imposed structure in terms of *trials* (e.g., in referential communication tasks; Clark & Wilkes-Gibbs, 1986; Holler & Wilkin, 2011) or *games* (e.g., in the Maze task; Garrod & Anderson, 1987).

When studying alignment in conversation, analyses can be restricted to specific sequences. For example, Chui (2014) qualitatively investigated gestural alignment in 12 short stretches of talk in free interaction, in which interlocutors communicated about the meaning of a referent. However, 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 3993 cross-participant gesture comparisons for the analyses.

Once the to-be-analyzed data has been selected, a possible approach is to look at paired behaviors which are in a specific sequential relation to each other. For example, alignment can be analyzed on the speech turn level; i.e., one compares the behavior in turn *x* from speaker A and following turn *y* from speaker B. Thus in this case adjacent speech turns are taken as the unit of analysis[3], 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). However, there are also various approaches possible which do not take speech turns into account but look at adjacent behavior, e.g., 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). Finally, an approach completely independent of sequential structures is also possible, by simply comparing all instances of a kind of behavior category (such as iconic gestures) from both interlocutors (cf. Bergmann & Kopp, 2012; Louwerse et al., 2012) or restricting analyses to pre-defined time-windows (Oben, 2015).

The analytic approaches which are adopted seem to stem from the research tradition or theoretical stance of the researchers. If alignment is approached as an interactive grounding process, which takes place with a certain communicative intention, the sequential context is highly relevant. Immediate repetition of words in the following turn (constituting an adjacency pair) could be used to express surprise, answer a question, or accept a formulation, to name a few (Norrick, 1987). Furthermore, research has shown that alignment often occurs in the context of other-initiated repair. In task-based interactions, there is a significantly larger likelihood of finding alignment in *repair* turn pairs (consisting of a problematic turn followed by a repair-initiation) compared to other adjacent turns (Fusaroli et al., 2017). However, seen from a priming perspective, alignment is argued to result from activation of (linguistic or motor) representations, deeming sequential structures of the discourse to be irrelevant.

We could thus 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 organization (as long as the temporal distance is the same), based on grounding accounts we could expect higher amounts of alignment in turns that are in a specific sequential relation (e.g., repair sequences, such as reported by Fusaroli et al. (2017)). 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 hypothesize equal amounts (as long the temporal distance is matched). This would thus constitute an interesting test bed for

[3] Note that we use the term "adjacency" here in the simple sense of adjacent or neighboring; not to be confused with the Conversation Analytic term *adjacency pairs*, as referred to earlier.

contrasting priming and grounding approaches, though surprisingly there are of yet no empirical investigations on this matter.

4.2 Time

As already became clear above, the temporal relation between two aligned instances of behavior can vary. In some studies, 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, 2011). Another common approach which does not operationalize alignment in terms of a certain temporal relation, is to restrict analyses to paired behaviors which are adjacent, such as comparing adjacent trials (Fusaroli et al., 2012), turns (Fusaroli et al., 2017), gestures (Bergmann & Kopp, 2012) or lexical or gestural references to the same referent (Oben & Brône, 2016). Adjacency is operationalized as the absence of intervening behavior of the same type (i.e., there should be no intervening trials, turns, gestures, or references). Though note that as mentioned earlier, adjacency usually implicates a relatively short temporal distance between the two pair parts.

Conversely, alignment could be operationalized as having to occur within a pre-defined temporal window. Studies on the temporal dynamics of alignment have used a technique called *time-aligned moving averages* (TAMA), where a specific time-window (e.g. of 40 seconds) is shifted across the time-axis in a step-wise manner. However this has mostly been used for analyses of prosody (e.g., De Looze, Scherer, Vaughan, & Campbell, 2014), and is not common for lexical or gestural alignment (but see Oben, 2015).

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. Though note that the paired behaviors that are part of the analyses can still differ considerably across studies: whereas in the work by Oben (2015) all gesturally aligned pairs occur within 40 seconds, for Holler and Wilkin (2011) this can go up to several minutes (though the actual time lags are not reported), as long as they were referring to the same referent (see also Section 4.3 below). Again, these operationalizations can be tightly interwoven with one's theoretical presuppositions regarding the underlying mechanism of alignment. Alignment across large time intervals is less likely to be considered in studies working from a priming approach, as priming effects are hypothesized to decrease over time.[4] 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

[4] Though the term priming generally refers to a short-term, automatic effect (lasting a few seconds at most), a case is also made for the existence of so-called "long-term" priming (cf. Pickering & Garrod, 2004), though note that in practice this amounts to an interval of eight sentences (Hartsuiker & Westenberg, 2000) or ten sentences, with an average of 72 seconds (Bock & Griffin, 2000).

of certain conceptual pacts and the need to keep repeating certain lexical items or gestures, though this would not hold for experimental set-ups like in Holler & Wilkin (2011), as target items re-occur in various trials throughout the experiment. However, with the 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 locate a problematic turn which was produced earlier in the conversation (Sacks, 1992; Schegloff, 2000). Thus in general, both theoretical accounts would argue that over time the likelihood of alignment decreases, though for priming this effect would be mechanistic in nature (due to decreased levels of activation), while for grounding it would be incidental.

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, 2011; 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.

4.3 Semantics

A critical difference in studies is whether they require pairs of gestures or words to convey the same meaning or refer to the same referent in order to count as aligned. There are studies on gestural alignment for which this is not a criterion, as they focus specifically on form-similarity (e.g., Bergmann & Kopp, 2012) or alignment of gesture type (Louwerse et al., 2012). With respect to lexical alignment, a similar approach would be to use scripts to search for cross-recurrence of (lemmatized) words in transcripts (cf. Fusaroli et al., 2017; and the Python package ALIGN by Duran, Paxton, & Fusaroli, 2019).

On the other hand, there are studies which do require a semantic link between the paired behaviors, though this kind of semantic coding is not always trivial – especially for gestures, as they are highly context-dependent and two similar gestures can mean completely different things in distinct context. Yet this holds for lexical items to a certain extent as well, as there is a high level of ambiguity in natural language. Especially in challenging communicative situations (such as a Maze Task), identical words can be used to denote different things (Garrod & Anderson, 1987; Mills & Healey, 2008).

There are various ways to examine the semantic overlap between instances of behavior. 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 (2011); 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, it should be noted that there is no one-on-one correspondence between the semantic meaning of words or gestures, and the referent that it is referring to. On the one hand, participants can talk about the *same referent*, yet still lexically or gesturally single out different semantic properties; e.g. when using the word 'straight' or a gesture to depict the orientation versus shape of (a part of) an object. On the other hand, 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 & Mayor, 2013).

The semantic dimension draws the most distinct line between the priming and grounding approaches. Priming approaches argue that a low-level, automatic mechanism results in form overlap in behavior, which can occur independently from alignment at the semantic level. Grounding approaches on the other hand, regard instances of behavior as means to display (mis)understanding or a mutually established *conceptual pacts*, and thus expect alignment to occur when there is a semantic or referential link between the instances of behavior.

4.4 Form

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, viz. the repetition of a particular base word or lemma, thus excluding synonyms or paraphrases (cf. Fusaroli et al., 2017; Howes, Healey, & Purver, 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 nouns and verbs (Bangerter & Mayor, 2013) or only nouns (Oben, 2015).[5]

Regarding gestural alignment, similarly to lexical alignment, "gestural rephrasing" has also been considered a form of "repetition" (e.g., Tabensky, 2001). In these cases, gestures overlap in the meaning they convey,

[5] The exact operationalization 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 recognized 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.

but have a different form or shape. However, most studies on gestural alignment require at least some degree of form resemblance, though studies vary quite substantially with respect to how this is measured. Generally speaking, gestural alignment has been operationalized as overlap in *mode of representation*, specific form features or a combination of those.

Three studies used mode of representation (or representation technique, e.g. the hands can draw the outline of an object, enact a certain action, etc.; Streeck, 2008). 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 (2011) 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 Fig. 3 below:

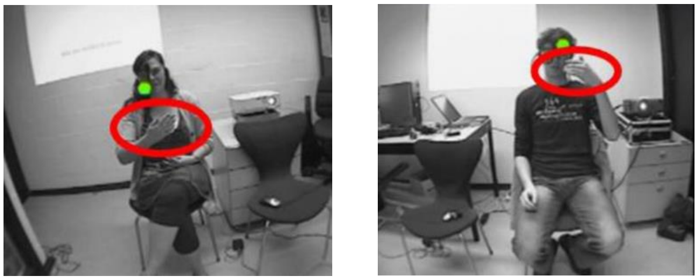


Fig. 3. Gestures with overlap in "modelling" as the mode of representation, reproduced from Oben & Brône (2016). [Image used under Fair Use purposes]

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 (Fig. 4):

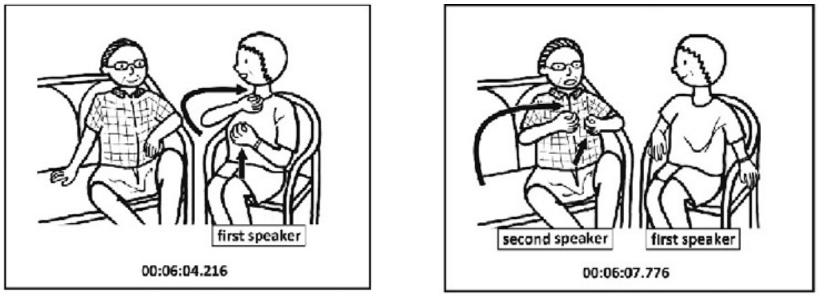


Fig. 4. Gestures with overlap in handedness, handshape, position and motion, reproduced from Chui (2014). [Image used under Fair Use purposes]

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).

Finally, a different approach is to use neither meaning nor form as criteria for alignment, but instead code words or gestures on a more general level. An example of this has already been discussed in Section 3: Louwerse et al. (2012) considered gestures to be aligned when they were of the same gesture type (iconic, deictic, etc.), irrespective of differences in form and/or meaning. In a similar vein speech was categorized into dialog acts (e.g. acknowledgement or clarify) and description types (e.g. use of digits or spatial propositions). However, this approach does not seem to generalize well to investigations of lexical alignment, as that would entail that words would have to be of the same "type" in order to be aligned (for example in terms of part of speech (noun, verb etc.) or constituent type (noun phrase, verb phrase etc.)).

Generally speaking, both priming and grounding perspectives use form as their main criterion for considering behavior 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 lack such a mechanistic explanation. Consequently, priming accounts might have stricter form criteria for selecting paired behaviors, as the theory might not account for form adjustments in terms of e.g. handshape or motion.

Furthermore, the priming account of Pickering and Garrod (2004) hypotheses a fixed relationship between form alignment and mutual understanding: more alignment of linguistic behavior will lead to more alignment on higher levels of representations. According to the grounding account, this relationship is more variable. For example, alignment can be used to signal understanding, thereby grounding a certain referring expression (as is the case with the "meat-eating Chad" in Example 1). However, as previously addressed by others (e.g., Fusaroli et al. (2017) and extensively reported on within Conversation Analysis (e.g., Schegloff, 2000), lexical alignment often occurs in the form of a repair initiation. Lexical alignment can thus also signal *misunderstanding*, and is a means to *get to* higher-level alignment, rather than being an indicator of it (see also Mills & Healey, 2008). Furthermore, (purposefully) using different words or gestures can also be a way to establish mutual understanding (e.g., Clark & Wilkes-Gibbs, 1986; Holler & Wilkin, 2011; Tabensky, 2001). For example, Holler and Wilkin (2011) 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 Section 4.5). 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 conceptualization of the figure. Thus, in grounding accounts there is no fixed relationship between alignment of behavior in terms of form on the one hand, and mutual understanding on the other.

4.5 Modality

A prevalent assumption in the work on alignment is that for any pair of behavior which is considered aligned, the behavior is produced within the *same* modality. That is, the relation between the two pair parts of behavior is considered to be a unimodal one. However, from a theoretical point of view, the two parts of aligned behavior 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 behavior should be similar in form to at least some degree, which is impossible when produced in different modalities.[6] However, we argue that two instances of behavior which are in a certain sequential, temporal, and/or semantic relation to each other, can still be considered aligned.

From both the priming and grounding frameworks it would follow that, for cross-speaker speech-gesture alignment, the semantic relation would be decisive. Though this is not part of Pickering and Garrod's original model (2004), there is evidence that iconic gestures prime semantically related words (e.g., Yap, So, Yap, Tan, & Teoh, 2011). Seen from the grounding framework, 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 frameworks thus build on the assumption that the matching of public behavior in interaction ultimately must be related to some sort of convergence in private

[6] Though *ideophones* might be the exception to this rule, as these words evoke an impression of a certain sensory perception. In principle, such words could thus overlap in "form" with behavior produced in other modalities, such as an action or movement expressed through gesture.

conceptualizations (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 would be challenging, especially in free conversation. To our knowledge, there is only one study which has investigated lexical and gestural alignment with this 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, can be repeated by means of gestures, 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).

The work on alignment from a cross-modal perspective is thus scarce, and such 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 behavior 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 centered 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), discussed earlier in Sections 2 and 3, many kinds of behavior in multiple modalities (linguistic expressions, facial expressions, manual gestures and non-communicative 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 operationalized 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 on 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 (2011) 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. At times interlocutors even dropped speech altogether while slowly and synchronously mimicking the interlocutor’s gestures in an effort to signal incremental understanding. 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.

4.6 Review summary

By unpacking the multidimensional nature of alignment into five dimensions, we captured the space of possibilities of what can be considered alignment. We distinguished two prominent theoretical approaches (priming and grounding), and showed how their assumptions regarding the underlying mechanism of alignment pattern together with methodological choices. A summary of the two approaches is presented in Table 2. Broadly speaking, those working from the premise that communication is (at least partly) driven by automatic, lower-level processes (the priming approach) tend to use quantitative analyses to compare instances of behavior irrespective of their sequential relation, prioritize form resemblance (rather than semantic overlap), and restrict analyses to one modality. In contrast, researchers working from the idea that communication is an interactive, collaborative undertaking (the grounding approach) are more likely to conduct or build on qualitative analyses, focus on the semantic information conveyed by the potentially aligned behavior, thereby taking the (multimodal) discourse context and its sequential structure into account.

As this summary shows, the five dimensions differ in terms of their relative importance. The dimension *form* seems to be most prominent in the literature, as this is by most considered to be the "core property" of what makes behavior aligned (though operationalizations vary). The dimensions *time* and *semantics* are also deemed important; priming accounts predict that priming effects decrease over time, and those working from the grounding approach only consider behaviors to be aligned when they have the same meaning or referent. However, the review highlights that there is of yet limited theorizing and empirical work with respect to the dimensions *sequence* and *modality* – yielding promising avenues for future research, which we will come back to in the Discussion.

Table 2. Schematic summary of relations between empirical and theoretical approaches (with references as examples).

	Priming	Grounding
Underlying mechanism	Automatic Non-intentional Low-level *Pickering & Garrod, 2004*	Controlled Intentional Higher-level *Brennan & Clark, 1996*
Data collection	Controlled experiments *Branigan et al. 2000* Task-based interactions *Garrod & Anderson, 1987*	Naturalistic interactions *Chui et al., 2014* Task-based interactions *Holler & Wilkin, 2011*
Modes of analysis	Quantitative *Branigan et al. 2000*	Qualitative *Tabensky, 2001*
Dimensions prioritized	Time Form *Louwerse et al. 2012*	Semantics Form *Holler & Wilkin, 2011*

5. The Framework in Practice

To further illustrate the dimensions of the framework, we will apply it to showcase various operationalizations of alignment in an example excerpt. We do not intend to prescribe specific coding practices, but instead offer a conceptual guide for researchers working on alignment and related concepts (such as behavior matching, mimicry or accommodation) to disentangle the range of possible relations between various pairs of behaviors.

The example comes from a large corpus of task-based interactions collected by the Communicative Alignment in Brain and Behavior group (CABB; see the Acknowledgments). In this referential task, Dutch participants communicate about unfamiliar 3D objects called *Fribbles* on a trial-by-trial basis. A short excerpt is presented in Fig. 5, with the abstract visualization of the observable behavior on top.

Starting with the dimension *sequence*, on a global level we could regard the entire excerpt to be part of one project, which, in this case, is to describe and find one particular target figure. However, on a smaller scale, we can identify so-called *adjacency pairs*, such as the question-answer pair of lines 2 and 3, where repair is initiated and resolved. Other pairs of utterances, though also adjacent, do not have such a strong cohesive relationship (e.g. lines 3 and 4). Consequently, the speech and gestures in lines 2 and 3 are in a different (more interdependent) sequential relation than the speech and gestures in line 3 and 4.

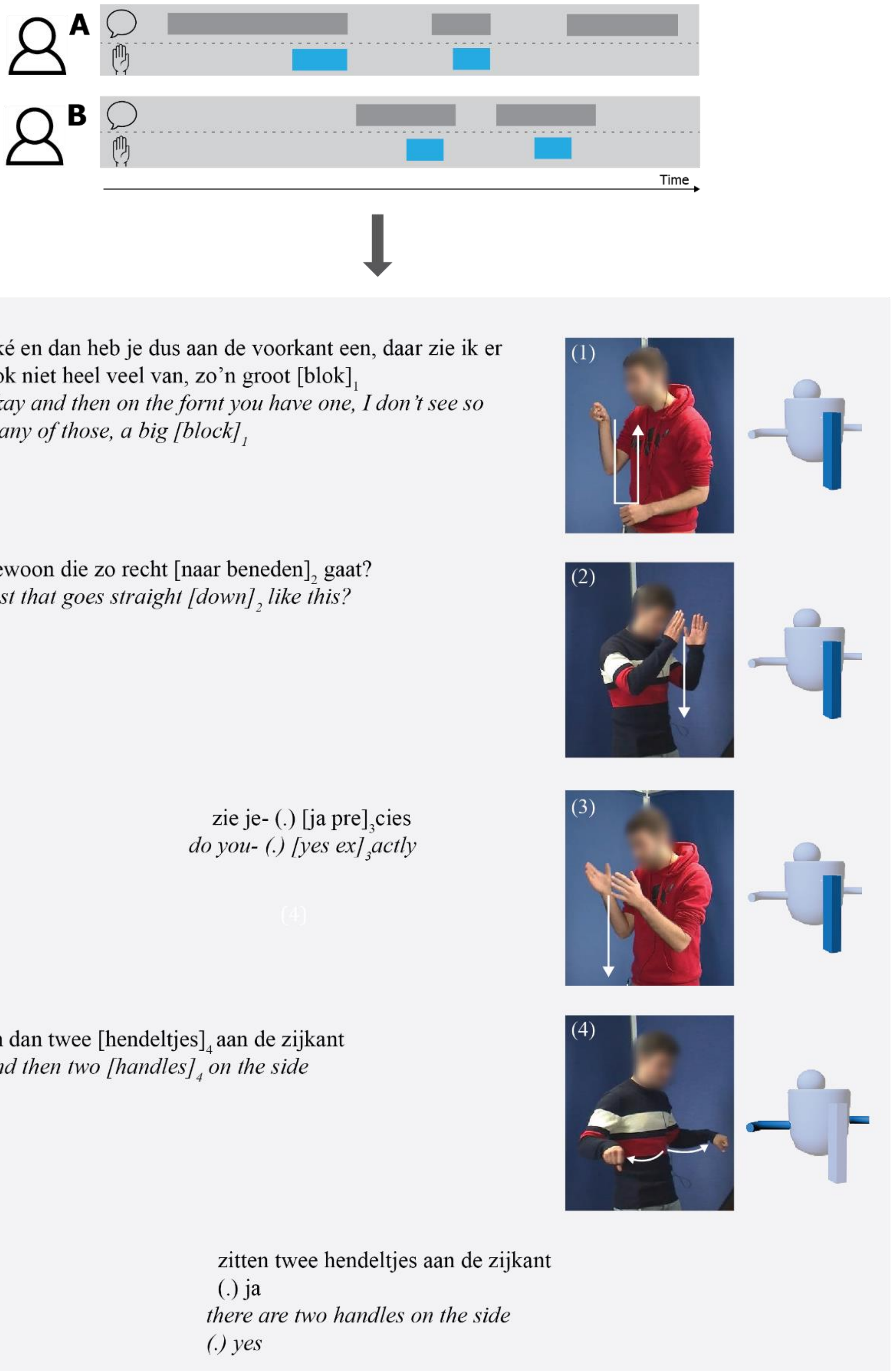
In terms of the dimension *time*, a possible approach would be to compare adjacent turns (in the example above: line 1 vs. 2, line 2 vs. 3, etc.) or adjacent gestures (gesture 1 vs. 2, gesture 2 vs. 3, etc.). Alternatively, behaviors which are further apart (such as gesture 1 and 4, or lexical references '*blok'* ('block') and '*hendeltje'* ('handles')) could also be compared, though these particular paired behaviors are less closely related in terms of sequence or semantics. Another operationalization could be to apply a strict temporal restriction, such as using a time window of 40 seconds, which for this excerpt would result in the inclusion of all behavior (as the excerpt has a total length of 12 seconds).

When we approach the excerpt from the dimension *semantics*, we can note that the first three gestures all refer to the same subpart of the object (the rectangular shape at the front side). However, when considered in relation to the co-expressive speech, these three gestures do single out different semantic properties. Gesture 1 highlights the shape and size of the subpart; gestures 2 and 3 emphasize the orientation of it. Regarding speech, lines 1 and 2 both refer to the same subpart, though they are not aligned semantically in the narrow sense (as there are no overlapping words or synonyms). However, the words in the two utterances are semantically related, since a property of blocks is that they are straight-sided. One way to quantify such semantic similarity would be to calculate their distance in a semantic space using distributional semantics.

Regarding alignment of *form*, it is striking that both participants use the exact same words in lines 4 and 5. Even though lexical alignment is evidently present here, the way it is to be measured is not straightforward. First of all, one could compare the complete speech turns (or a somewhat “cleaner” version of those, e.g. by removing words based on frequency or word class). Other approaches would be to compare only noun phrases (‘*twee hendeltjes*’; ‘two handles’) or nouns (‘*hendeltjes*’; ‘handles’). Subsequently, there are various possibilities in terms of computing the amount of overlap between the turns (e.g., absolute versus relative to the total number of words). Regarding form overlap in gestures, one could look for overlap in mode of representation or specific form features (e.g. handedness, handshape, palm orientation, position and motion).

Finally, we do not only find alignment (on various dimensions) *within* modalities, but also *across* modalities. In this excerpt, we find cases where cross-modal alignment complements within-modality alignment, for example the gesture on line 3 which aligns to both the speech and gesture of line 2.

By disentangling the multitude of possible relations between paired behaviors in this short excerpt, we showed the complex and multidimensional nature of alignment and how this can be captured using the five dimensions of the framework. We furthermore illustrated the numerous ways in which alignment can be measured, thereby highlighting the importance of explicitly describing those empirical choices and how they follow from theoretical assumptions. The proposed framework can be used by researchers working on alignment and related phenomena (e.g., behavior matching, mimicry and accommodation) to clearly and systematically spell out predictions and operationalizations in terms of the five dimensions.



1 A oké en dan heb je dus aan de voorkant een, daar zie ik er ook niet heel veel van, zo'n groot [blok]$_{1}$
okay and then on the fornt you have one, I don't see so many of those, a big [block]$_{1}$

2 B gewoon die zo recht [naar beneden]$_{2}$ gaat?
just that goes straight [down]$_{2}$ like this?

3 A zie je- (.) [ja pre]$_{3}$cies
do you- (.) [yes ex]$_{3}$actly

4 B en dan twee [hendeltjes]$_{4}$ aan de zijkant
and then two [handles]$_{4}$ on the side

5 A zitten twee hendeltjes aan de zijkant
(.) ja
there are two handles on the side
(.) yes

Fig 5. Example excerpt from a corpus of task-based interactions of the Communicative Alignment in Brain and Behavior (CABB) team. Free English translations are provided in italics below the original Dutch transcript. Square brackets indicate the start- and endpoint of the meaningful part of the co-speech gestures (stroke phase). The specific subparts of the figure that participants refer to are highlighted on the right.

6. 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 understandings on the phenomenon, which results in disparate analyses and results, which are hard to reconcile because they refer to qualitatively different types of alignment. In an effort to enable cumulative progress and principled comparison, we unpacked the complex nature of alignment into five constituent dimensions. We distinguished between priming and grounding as the two most prominent theoretical approaches, and showed that priming approaches prioritize the dimensions *form* and *time*, while grounding approaches mostly focus on *form* and *semantics*. In this section we identify a number of limitations and gaps in the field, and make suggestions for how the framework can benefit future studies.

A shortcoming in the field that the review reveals is that little is known about how alignment at various types and levels of (linguistic) behavior are related. First of all, more work is needed to ascertain whether the current postulated underlying mechanisms (priming versus grounding) generalize to alignment of any behavior, or perhaps only apply to a specific subset. For example, copying each other'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 behavioral alignment might fall somewhere in between, with strategic as well as more automatic components being at play simultaneously (cf. Kopp & Bergmann, 2013). Secondly, very little is known about potential (causal) relations between alignment at various channels or (linguistic) levels of behavior. From the priming perspective it has been argued that alignment at one (linguistic) level can lead to alignment at other levels (Pickering & Garrod, 2004), though empirical evidence so far yields mixed results (see, e.g., Branigan et al., 2000; versus Oben & Brône, 2016). From the grounding perspective, one might argue that different kinds of behavioral 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 conceptualizations, 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, though any systematic inquiry is yet to be undertaken.

These research gaps are largely due to the current challenges involved in comparing findings from studies on various types and levels of (linguistic) behavior. As shown in the current review, there is a large space of empirical possibilities for studying alignment (especially in terms of the dimensions *semantics* and *form*), and these decisions are often guided by theoretical presuppositions. In order to be able to cumulate results in a meaningful way, we recommend researchers to be explicit about their theoretical take on alignment by explaining which dimensions are (not) in focus, and how that has been operationalized. To this end, we believe the proposed framework can be a useful resource; adopting a common terminology across studies will greatly enhance comparability of studies and subsequent theory building in the field.

Furthermore, by elucidating the multidimensional nature of alignment, we reveal that there is of yet limited theorizing and empirical work with respect to the dimensions *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. When arguing from the grounding perspective, there are good reasons to believe that alignment rates will be higher within a certain sequence (such as an adjacency pair) than when they transcend such organizational units of conversation. In contrast, presuming that an automatic priming mechanism underlies alignment, we could hypothesize that only the temporal proximity affects alignment, irrespective of the absence or presence of a certain sequential relation.

When it comes to the dimension *modality*, 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 “behavior matching”), because there is a lack of form-resemblance, which is a key characteristic in both grounding and priming accounts. However, research has shown that listeners align to the speaker’s verbal narration in a non-verbal manner, such as wincing or a showing a concerned facial expression when someone tells a close-call story (Bavelas, Coates, & Johnson, 2000) – which could be considered a form of meaning-alignment as well. Yet cross-speaker speech-gesture relationships remain understudied, which is remarkable, given that speech and gesture are semantically co-expressive (McNeill, 1992), and tightly linked in both production and comprehension (Cassell, McNeill, & McCullough, 1998; de Ruiter, Bangerter, & Dings, 2012; Kita & Özyürek, 2003; Kopp & Bergmann, 2013; Mol et al., 2012, for review see Özyürek, 2018). Thus, little is known about whether, and if so how, lexical and gestural alignment are inter-related, making it a promising avenue for further research.

7. Conclusion

By outlining the complex nature of alignment into five constituent dimensions (*sequence, time, semantics, form* and *modality*), we have elucidated the diverging theoretical interpretations and empirical operationalizations of alignment. The novel conceptual framework furthermore enables us to single out unexplored research avenues and derive new hypotheses from existing theories. To conclude, we believe the framework may advance the field of alignment (and related phenomena) as a whole, as it can be used as a conceptual tool to disclose hidden assumptions, further refine theories and cumulate knowledge.

Acknowledgments

This paper came about in the context of a multidisciplinary research project within the Language in Interaction consortium, called Communicative Alignment in Brain & Behavior (CABB). The authors wish to make explicit that the work has been shaped by many (implicit, non-author) contributions of CABB team members (in alphabetical order): Mark Blokpoel, Sara Bögels, Lotte Eijk, Judith Holler, Branka Milivojevic, Iris van Rooij, Ivan Toni, and Marieke Woensdregt. We are also grateful to members of the Multimodal Language and Cognition group for comments on an earlier version of this paper. 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.

References

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., … Weinert, R. (1991). The Hcrc Map Task Corpus. *Language and Speech*, *34*(4), 351–366. https://doi.org/10.1177/002383099103400404

Bangerter, A., & Mayor, E. (2013). Lexical entrainment without conceptual pacts? Revisiting the matching task. *Proceedings of the Workshop Production of Referring Expressions: Bridging the Gap between Cognitive and Computational Approaches to Reference*, 6.

Bavelas, J. (2007). Face-to-face dialogue as a micro-social context: the example of motor mimicry. In S. D. Duncan, J. Cassell, & E. T. Levy (Eds.), *Gesture and the dynamic dimension of language* (pp. 127–146). Amsterdam: John Benjamins.

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

Bergmann, K., & Kopp, S. (2012). Gestural Alignment in Natural Dialogue. *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, 1326–1331.

Bock, K., & Griffin, Z. M. (2000). The persistence of structural priming: Transient activation or implicit learning? *Journal of Experimental Psychology: General*, *129*(2), 177. https://doi.org/10.1037/0096-3445.129.2.177

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

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

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

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. https://doi.org/doi:10.1037/0022-3514.76.6.893

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

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., 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

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

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

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., 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. *Proceedings of the 39th Annual Conference of the Cognitive Science Society*, 2055–2060.

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

Genschow, O., van Den Bossche, S., Cracco, E., Bardi, L., Rigoni, D., & Brass, M. (2017). Mimicry and automatic imitation are not correlated. *PLOS ONE*, *12*(9), e0183784. https://doi.org/10.1371/journal.pone.0183784

Hartsuiker, R. J., & Westenberg, C. (2000). Word order priming in written and spoken sentence production. *Cognition*, *75*(2), B27–B39. https://doi.org/10.1016/S0010-0277(99)00080-3

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

Holler, J., & Wilkin, K. (2011). Co-Speech Gesture Mimicry in the Process of Collaborative Referring During Face-to-Face Dialogue. *Journal of Nonverbal Behavior*, *35*(2), 133–153. https://doi.org/10.1007/s10919-011-0105-6

Howes, C., Healey, P. G. T., & Purver, M. (2010). Tracking Lexical and Syntactic Alignment in Conversation. *Proceedings of the Annual Meeting of the 32nd Cognitive Science Society*, 7.

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

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

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.), *Advances in Interaction Studies* (Vol. 6, pp. 87–108). Amsterdam: John Benjamins Publishing Company.

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: Oxford University Press.

Levinson, S. C. (2013). Action formation and ascription. In T. Stivers & J. Sidnell (Eds.), *The handbook of conversation analysis* (pp. 103–130). https://doi.org/10.1002/9781118325001.ch6

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

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought*. Chicago: University of Chicago Press.

Mills, G., & Healey, P. (2008). Semantic negotiation in dialogue: the mechanisms of alignment. *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, 46–53. Columbus, Ohio: Association for Computational Linguistics.

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

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

Oben, B. (2015). *Modelling interactive alignment: A multimodal and temporal account (Unpublished doctoral dissertation)*. KU Leuven, Leuven, Belgium.

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*. Retrieved from https://benjamins.com/catalog/ais.10.10obe

Oben, B., & Brône, G. (2016). Explaining interactive alignment: A multimodal and multifactorial account. *Journal of Pragmatics*, *104*, 32–51.

Özyürek, A. (2018). Role of Gesture in Language Processing. In S.-A. Rueschemeyer & M. G. Gaskell (Eds.), *Oxford Handbook of Psycholinguistics* (2nd ed., Vol. 1, pp. 592–607). https://doi.org/10.1093/oxfordhb/9780198786825.013.25

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

Sacks, H. (1992). *Lectures on Conversation* (G. Jefferson, Ed.). London: Blackwell.

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. (2007). *Sequence Organization in Interaction: A Primer in Conversation Analysis* (Vol. 1). Cambridge: Cambridge University Press.

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

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

Shin, Y. K., Proctor, R. W., & Capaldi, E. J. (2010). A Review of Contemporary Ideomotor Theory. *Psychological Bulletin*, *136*(6), 943–974. https://doi.org/10.1037/a0020541

Streeck, J. (2008). Depicting by gesture. *Gesture*, *8*(3), 285–301. https://doi.org/10.1075/gest.8.3.02str

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: Cambridge University Press.

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