Language and Cognition **7** (2015), 546–562. doi:10.1017/langcog.2015.22

What you see is what you do: on the relationship between gaze and gesture in multimodal alignment

BERT OBEN*

AND

GEERT BRÔNE

University of Leuven, Department of Linguistics

(*Received* 21 *December* 2012 – *Revised* 12 *September* 2014 – *Accepted* 14 *April* 2015)

ABSTRACT

Interactive language use inherently involves a process of coordination, which often leads to matching behaviour between interlocutors in different semiotic channels. We study this process of interactive alignment from a multimodal perspective: using data from head-mounted eye-trackers in a corpus of face-to-face conversations, we measure which effect gaze fixations by speakers (on their own gestures, condition 1) and fixations by interlocutors (on the gestures by those speakers, condition 2) have on subsequent gesture production by those interlocutors. The results show there is a significant effect of interlocutor gaze (condition 2), but not of speaker gaze (condition 1) on the amount of gestural alignment, with an interaction between the conditions.

KEYWORDS: interactive alignment, eye-gaze, gesture, multimodal, face-to-face conversation.

1. Introduction

When viewing language in its most natural setting, viz. face-to-face conversation, it inherently involves both an interactive and multimodal dimension. Despite the tacit agreement on the primacy of dialogue, (psycho) linguists interested in the cognitive underpinnings of language have only recently started to pay systematic attention to the INTERACTIVE grounding of the language system and its interaction with other semiotic modes. In the

[*] This research was supported by the Research Foundation - Flanders (project 3H110718: "Modelling Interactive Alignment Processes. A Multimodal and Multifocal Approach", granted to Kurt Feyaerts & Geert Brône). The authors would like to thank Kurt Feyaerts for extensive discussions on the set-up of this study and Koen Jaspaert for his support in statistically processing the data. e-mail: bert.oben@arts.kuleuven.be

broad paradigm of Cognitive Linguistics, this has led to significant adaptations of existing models such as Cognitive Grammar (Cienki, this issue; Langacker, 2001; Verhagen, 2005) and construction grammar (Fried & Östman, 2005; Goldberg, 2006). These models have incorporated features of the interactional context in their theoretical modelling and discuss the trade-off between different semiotic channels (most notably the question of (co-speech) gesture; e.g., Cienki & Müller, 2008; Mittelberg, 2007; Sweetser, 2007; a.o.).

The present paper is intended as a contribution to the cognitive analysis of multimodal interaction. More specifically, we take as a starting point Garrod and Pickering's (2004, p. 11) programmatic statement on key questions for future research into the cognitive and linguistic mechanisms of dialogic language use. One of these questions, according to the authors, concerns the multimodal analysis of interactive alignment processes in dialogue: "What is the relationship between linguistic and non-linguistic alignment processes?" We approach this question through an empirical analysis of the interrelation between interlocutors' gaze behaviour and interactionally co-present gestures. More specifically, following up on Gullberg and Kita (2009) and Wang and colleagues (Wang & Hamilton, 2014; Wang, Newport, & Hamilton, 2011), we address the question whether a speaker's visual focus on his own gestures when speaking, and/or an addressee's focus on that same gesture, has an influence on gestural alignment between the interlocutors. We base our analysis on the InSight Interaction Corpus (Brône & Oben, 2015), a multimodal video corpus consisting of targeted and free-range dyadic interactions, with head-mounted scene cameras and eye-trackers providing a unique 'speaker-internal' perspective on the conversation.

In what follows, we briefly review relevant literature on the role of eye-gaze in interactional discourse (Section 2), and on interactive alignment as a key feature in setting up successful communication (Section 3). We then combine these two features (eye-gaze and interactive alignment) into the central research question for the present paper: 'What is the influence of eye-gaze on multimodal interactive alignment?' (Section 4). In the methods section (5) we introduce the InSight Interaction Corpus and explain how we measured gaze and gesture behaviour. Finally, we present (Section 6) and discuss (Section 7) the main results of the study, and close with some perspectives for future work.

2. Eye-gaze in interaction

Some early seminal works focused on the role of eye-gaze in interactional settings, inspired by the work on paralinguistic (intonation, volume, pitch) and non-linguistic (gesture, gaze) signalling in conversation analysis. Kendon (1967), one of the pioneers in multimodal analysis (and more particularly gesture studies), presented a first detailed empirical account of the direction

https://doi.org/10.1017/langcog.2015.22 Published online by Cambridge University Press

of speakers' and hearers' gaze during a face-to-face conversation, based on snapshots of a video-recording (1 frame per 500 ms interval was used as a basis for the analysis) and corresponding transcription of speech. The analysis focused on the role of eye-gaze as an instrument of perception and processing on the one hand, and on its potential as a communicative signal on the other (e.g., for directing attention, managing the organization of the conversation, etc.). The pioneering work by Kendon fostered the interest of researchers across disciplines in the role of gaze behaviour in interactional discourse. In another landmark publication, Argyle and Cook (1976) list a range of functions of gaze (cited in Raidt, 2008, p. 49), including the signalling of grammatical breaks, attention or disapproval, providing feedback (cf. also Bavelas Coates, & Johnson, 2002), etc. In more recent studies, some of these eye-gaze patterns have been studied in more detail and as part of a broader framework, as, e.g., the work on eye-gaze in conversational turn management and information structure by Cassell, Torres, and Prevost (1999), Jokinen (2010), and Novick, Hansen, and Ward (1996).

Despite the detailed empirical analyses presented in the early work on eye-gaze in interaction, these studies typically had to deal with the technological–methodological shortcomings of video-recordings. Although gaze estimations on the basis of video data may be useful for a basic segmentation of the distribution of visual attention (e.g., looking at an interlocutor vs. looking away), they are notoriously coarse-grained and unreliable for more detailed analysis (Kendon, 2004; Streeck, 2008). For instance, video-based analysis does not provide useful information on short fixations (of 200 ms or less), saccades (i.e., fast movements of the eye), and visual scan paths. In order to be able to include this level of detail in the analysis, a different methodological paradigm is needed, viz. eye-tracking. Eye-tracking, or the measuring of gaze points and eye movements during ongoing behaviour, allows for a highly detailed analysis of gaze patterns (see Rayner, 1998, for an overview). With the development of unobtrusive eye-tracking equipment in the last decade, in the form of remote or head-mounted systems, it is possible to measure subjects' gaze patterns while they are involved in interactive or collaborative tasks. This opens up a vast area of research in (cognitive) interaction studies, including the role of gaze as a directive instrument, the correlation between gaze and gesture, gaze as a disambiguation instrument, interactive alignment in various semiotic channels, etc. In what follows, we provide an overview of some recent work exploring eye-gaze in interaction, using eye-tracking technology.

One of the basic features of interaction is the joint focus of attention of co-participants in the process of establishing the coordinated action that is language use (Clark, 1996). One correlate of this basic feature is gaze coordination, or the joint visual focus on relevant aspects of the context (e.g., referents that are the current topic of conversation), also referred to

https://doi.org/10.1017/langcog.2015.22 Published online by Cambridge University Press

as SHARED GAZE. In a series of experiments, Richardson and colleagues (Richardson, Dale, & Kirkham, 2007; Richardson, Dale, & Tomlinson, 2009) measured the coupling of eye-movements between participants in mediated settings, i.e., with participants viewing a screen rather than a face-to-face setting. The results show a strong tendency towards joint visual attention, as well as an impact of shared background information (common ground) on this coupling. Brennan, Chen, Dickinson, Neider, and Zelinsky (2008) and Neider, Chen, Dickinson, Brennan, and Zelinsky (2010) used collaborative visual search tasks for remotely located pairs of people (both wearing head-mounted eye-trackers) to study the relative impact of shared gaze and speech on performance efficiency. The results show that the condition with shared gaze (with participants seeing the gaze cursor generated by the eye-tracker of the other) scored significantly better than (i) solitary search, (ii) a condition with only shared voice, and even (iii) the condition with both shared gaze and shared voice.

A final dimension that needs mentioning in this general overview, as it is of particular relevance to the present study, is the distribution of visual attention across the interactional space of the conversation. Co-participants do not only focus on specific referents that are the current topic of (verbal) communication, or on the speakers and/or addressees (see above). Rather, taking into account the active communicative role that gaze may play as a cue for referent assignment, turn management, etc., we obtain a complex interplay of gaze-as-cause and gaze-as-effect. One much-discussed feature is the so-called GAZE CUEING EFFECT, which can be defined as the effect that cueing a target (e.g., by looking at it) has on the gaze behaviour of an addressee. Studies on the gaze cueing effect, which date back to early work by Posner, Snyder, and Davidson (1980), stress its role for joint attention in interaction (Emery, 2000; Frischen, Bayliss, & Tipper, 2007). Lachat, Conty, Hugueville, and George (2012) are the first to test the gaze cueing effect in a spontaneous face-to-face setting (rather than in on-screen experiments), however, without the use of the eye-tracking paradigm. Gullberg and Holmqvist (2006) and Gullberg and Kita (2009) focus on one specific case of gaze cueing, using head-mounted eye-trackers, viz. the effect of a speaker focusing on his/her own gesture on the addressee's gaze behaviour. The studies reveal that a speaker's gaze at own gestures is a powerful cue for addressees to leave the dominant fixation position (i.e., the face of the speaker) and give overt visual attention to the speaker's gesture.

3. Interactive alignment: a multimodal approach

At various points in the previous section, it was stressed that efficient communication depends to a large extent on joint (visual) attention and coordination. In fact, coordinated actions in interaction have been observed

https://doi.org/10.1017/langcog.2015.22 Published online by Cambridge University Press

for a variety of language and bodily parameters, which have been addressed in several disciplines and paradigms. Or, as Richardson et al. (2007, p. 407) aptly put it:

> When people talk, they coordinate whose turn it is to speak (Sacks, Schegloff, & Jefferson, 1974). They also implicitly agree upon names for novel objects (Brennan & Clark, 1996; Clark & Brennan, 1991), align their spatial reference frames (Schober, 1993), and use each other's syntactic structures (Branigan, Pickering, & Cleland, 2000). Their accents become more similar (Giles et al. 1992), they sway their bodies in synchrony (Condon & Ogston, 1971; Shockley, Santana & Fowler, 2003), and they even scratch their noses together (Chartrand & Bargh, 1999).

One central effect of coordinated action in dialogue is a high degree of convergence across speakers and their turns. The imitative nature of face-to-face communication has been dealt with in great detail in linguistics and conversation analysis (e.g., Bolinger, 1961; Carter, 1999; Halliday & Hasan, 1976; Sacks et al., 1974; Tannen, 1987, 1989), and has been labelled alternatively as 'accommodation' (Giles et al., 1992), 'entrainment' (Brennan & Clark, 1996), 'resonance' (Du Bois, 2010), 'convergence' (Lewandowski, 2012), 'interactive alignment' (Pickering & Garrod, 2004, 2006), etc. Rather than presenting an overview of these different accounts (which would be well outside the scope of the present paper), we zoom in on the interactive alignment theory developed by Pickering and Garrod, as it has generated a substantial amount of theoretical and empirical work over the last decade. It focuses on the basic underlying mechanism driving the tendency towards convergence or accommodation, and argues that this is a largely automatic and unconscious process.

According to the interactive alignment theory, interlocutors are primed to use the linguistic input of immediately preceding utterances they have just processed: "We propose that this works via a priming mechanism, whereby encountering an utterance that activates a particular representation makes it more likely that the person will subsequently produce an utterance that uses that representation" (Pickering & Garrod, 2004, p. 173). Alignment is defined as a state in which two or more dialogue partners have an identical (or at least highly similar) representation at a particular linguistic level. Alignment at different levels of linguistic representation enhances alignment of situation models, with similarly constructed situation models being a precondition for successful communication. The theory proposes a layered account of interconnected levels of representation (see Figure 1), with alignment at one level leading to alignment at other levels. The growing body of literature on interactive alignment has uncovered the cognitive reality of the phenomenon, most notably at the lexical and syntactic level (see, e.g., Garrod & Pickering, 2009, for an overview of the literature). Some recent work has also focused on aspects of alignment at the level of

https://doi.org/10.1017/langcog.2015.22 Published online by Cambridge University Press

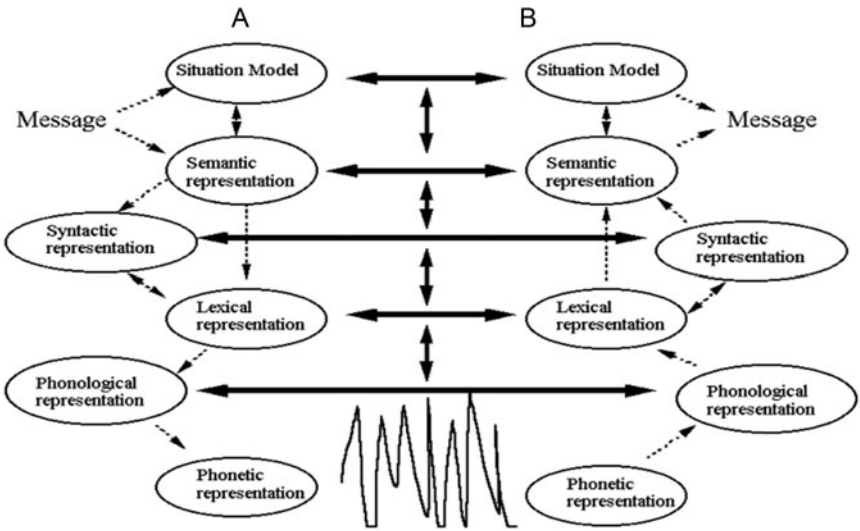

Fig. 1. Interconnected levels of linguistic representation in the interactive alignment model of Pickering and Garrod (2004).

phonetics (Lewandowski, 2012; Lewandowski & Schweitzer, 2010; Szczepek Reed, 2010) and pragmatics (Roche, Dale, & Caucci, 2012).

Although the interactive alignment model was construed as a mechanistic model of dialogue in psycholinguistics, and thus focuses primarily on linguistic levels of representation, the basic phenomenon can also be observed in other semiotic channels. Several studies have focused on the strong (and largely unconscious) tendency of dialogue partners towards convergence in non-linguistic behaviour, including gesture (Kimbara, 2006), posture (Shockley, Richardson, & Dale, 2009; Shockley, Santana, & Fowler, 2003) and eye-gaze (Richardson & Dale, 2005, cf. references in Section 2). In fact, Garrod and Pickering (2009, p. 296) acknowledge that "interlocutors construct aligned non-linguistic representations. Just as linguistic alignment at one level can enhance alignment at other levels, so non-linguistic alignment can also enhance alignment."

For the purpose of the present paper, a number of recent studies are of particular interest, as they are instructive for a multimodal approach that also takes into account the temporal dynamics of interactive alignment. Dale, Kirkham, and Richardson (2011) studied eye-gaze patterns in a cooperative tangram task. Both participants were given the same set of puzzle pieces, but in a different order, and were asked to arrive at a matching order. A cross-recurrence analysis on the eye-movement data for both participants reveals growing synchronization over time: eye-gaze patterns become more coordinated.

https://doi.org/10.1017/langcog.2015.22 Published online by Cambridge University Press

McNeill (2006) presents an exploratory study of the coordination of individual cognitive states in multi-party exchanges. He does so through the study of language, gesture, and gaze as packages of multimodal information (referred to as 'hyperphrases') that present a so-called 'growth point' or a speaker's cognitive state, and which are crucial for achieving synchrony between speakers. McNeill explicitly presents his account as a multimodal extension of the interactive alignment model, albeit without buying into the 'mechanistic' account of a solely priming-based mechanism.

4. The present study: eye-gaze and multimodal alignment

The present study combines insights from recent work on eye-gaze in interaction (see Section 2) and interactive alignment (see Section 3) to explore the role of eye-gaze in the emergence of alignment at the gestural level. More specifically, we address the question whether gaze cueing by a speaker on his or her own gesture has an effect not only on the gaze behaviour of the interlocutor (Streeck, 1993), but also on the subsequent gesture production by the latter. Do interlocutors who directly focused on a speaker's gesture exhibit a stronger tendency to use the same (or similar) gestures in subsequent turns than in those cases without direct fixation? We hypothesize that speakers' and interlocutors' attention to gesture not only has an effect on the visual attention of interlocutors and information uptake of gestural information (as shown by Gullberg & Kita, 2009), but also on gestural alignment across speakers. Drawing on insights from the interactive alignment model, we present an experimental study of the relation between gaze cueing, visual processing, and aligned (gestural) representations.

By linking visual fixation on gestures (on the part of both speakers and addressees) to subsequent gesture production in an interactional setting, we add one important dimension to the study on eye-gaze and information uptake of gesture presented by Gullberg and Kita (2009). Whereas in their study, participants were shown video-recordings of naturally occurring gestures in narratives, after which the information uptake was measured using a drawing task after stimulus presentation, we explore the relationship between visual gesture processing and production in a dialogic setting, using a multimodal dialogue corpus that includes eye-tracking data for both dialogue partners (Section 5).

5. Method

5.1. PARTICIPANTS

All of the participants (9 male, 21 female) are students at the University of Leuven and native speakers of Dutch. Each conversational pair was well

https://doi.org/10.1017/langcog.2015.22 Published online by Cambridge University Press

acquainted before taking part in the experiment, and they were all rewarded with cinema tickets.

5.2. DESIGN

All of the data and analyses in this paper are based on the InSight Interaction Corpus (Brône & Oben, 2015). This corpus consists of fifteen recordings of face-to-face conversations that last about 20 minutes each. Both of the participants are wearing head-mounted eye-trackers, and a fixed camera records the interaction from an external perspective. This recording technique results in a video file that simultaneously shows three perspectives on the speech situation (see Figure 2).

5.3. PROCEDURE

Each 20-minute conversation in the corpus consists of three types of interaction: a picture description task, a collaborative problem-solving task, and a free conversation on a given topic. For this study we only used the second interaction type, i.e., the targeted collaborative task. In this task, the interlocutors were each shown a video animation. They saw the animation at the same time, but they couldn't see each other's animation. The two video animations were identical, except for a few minor details. The goal of the task was to discover those differing details. Once they completed the task, they were shown a new animation (with a total of fifteen animations they had to discuss).

5.4. ANALYSIS

Since this paper is mainly concerned with the coupling of and interaction between gesture and gaze in face-to-face interaction, we will zoom in on those two levels, and not focus primarily on the linguistic or prosodic level. More specifically, we single out one type of gesturing and its relation to gaze, viz. depictive gestures that are used by the participants to represent objects that appear in the animation videos. All other gestures, like emblems, beats, metaphoric gestures, etc. are not part of the dataset for the present analysis. With regard to the eye-tracking data, we focus on the cases where, either in the role of speaker or hearer, participants explicitly focus on those depictive gestures. In other words, we included all the cases of interlocutors looking at their own hand gestures and the hand gestures of the other person.

In testing our hypotheses, we compare two factors: the alignment between adjacent representational gestures, and the eye-gaze of the interlocutors on

https://doi.org/10.1017/langcog.2015.22 Published online by Cambridge University Press

Fig. 2. Recording configuration of the InSight Interaction Corpus (Brône & Oben, 2015).

those gestures. The latter is a case of binary measuring results: either there is visual focus on the gesture or there is not. The former, the factor of alignment, is expressed on a more continuous scale: any two gestures can not only be fully aligned or fully non-aligned, they can also be partially aligned. This partial alignment is due to the fact that gesture is multidimensional by nature. Two gestures can, for example, be identical in handshape and finger orientation, but not in velocity and palm orientation. How we dealt with making both gesture and gaze quantifiable factors is clarified in what follows.

5.4.1. *Quantifying gaze*

Quantifying gaze data is fairly straightforward, albeit time-consuming. For this study we defined six REGIONS OF INTEREST (ROI; see Figure 3). On a frame-by-frame basis, we manually tagged the gaze data from the eye-tracker; i.e., for each frame we determined on which ROI the visual focus of the interlocutor was. This was done for each of the interlocutors during the entire interaction.

Eye-movements are extremely fast and, unless engaged in specific tasks, the human eye jumps from the one object of focalization to the other at a very high speed. Because we want to test the effect of eye-gaze on gesture production, it was important to establish a threshold for explicit visual focus or MINIMAL FIXATION DURATION. In order to regard a gaze event as a fixation, and thus as a cognitive act of perception, it should last at least a predefined number of milliseconds. In the eye-tracking literature there is much debate about a standard minimum fixation duration (Munn, Stefano, & Pelz, 2008), and this duration seems to be highly linked to the specific task of

https://doi.org/10.1017/langcog.2015.22 Published online by Cambridge University Press

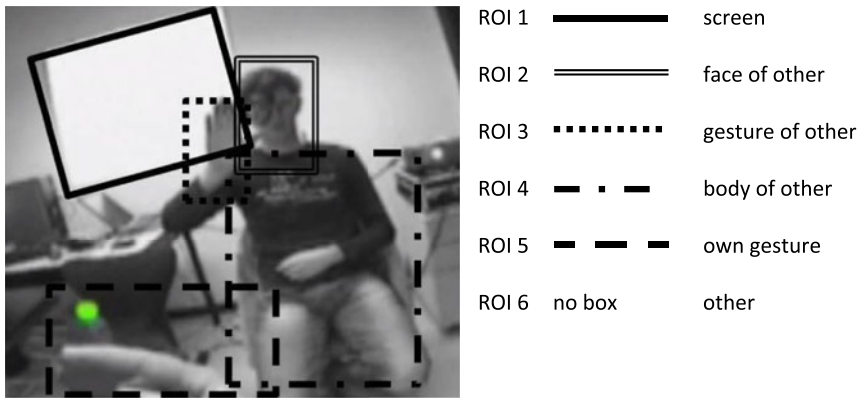

Fig. 3. Regions of interest.

the eye-tracked person. In reading, the minimum fixation duration will typically be shorter, as short as 60 milliseconds, whereas for other tasks this is generally between 150 and 400 milliseconds. For our dataset, we regarded any fixation at a gesture of at least five video frames (or 200 ms) as a genuine fixation (cf. also Gullberg & Kita, 2009).

A second issue regarding eye-gaze, apart from determining the minimum fixation duration, is peripheral vision. For our study we can only base our claims on the positive evidence of explicit eye-gaze fixations. It is, of course, always possible that subjects perceived gestures without focusing on them. As is clear from eye-tracking research in sign language (Muir & Richardson, 2005), signers hardly ever fixate their interlocutors' hands, while they 'see' what their conversational partners are expressing with those hands. The human peripheral vision allows perception without fixation, so we should take care in interpreting our data.

5.4.2. *Quantifying gesture*

Intuitively, gestures seem to be holistic *gestalts* of meaningful hand movements (Mittelberg, 2007). When annotating gesture, however, the multidimensionality of the phenomenon forces researchers to choose a level of granularity at which to describe the gestures at hand. For this study we were not primarily interested in annotating and transcribing gesture in as much detail as possible, but rather in adopting a measure of comparing two gestures, i.e., of expressing the degree of alignment between any two gestures. For this purpose we used a five-point scale of features to determine how alike two gestures are. The five features that were considered are finger orientation, palm orientation, handedness, gesture type, and handshape. The maximum

https://doi.org/10.1017/langcog.2015.22 Published online by Cambridge University Press

score for alignment between two gestures was 1, the minimum 0, with every feature counting for 0.2.[1]

The last two features of our annotation for gestural alignment require some further explanation. The annotation of gesture type is borrowed from Streeck (2008), who proposes a typology of shaping mechanisms for depictive gestures. For example, in representing a door by means of gesture, there are different shaping strategies: one could draw the outlines of the door, use the hand as a token for the object, use the hand to handle as if opening a door (by manipulating a fictional doorknob for instance), etc. Our category 'gesture type' corresponds with Streeck's typology in this respect. For the feature 'handshape' we used the gesture annotation grid developed by Bressem (2008). This notational system is strongly form-based (i.e., it is independent from the speech content) and well trained for co-speech gestures in spontaneous conversations, which made it well suited for our purposes.

6. Results

Before turning to the results on the relation between gaze behaviour and alignment in gesture production, we present two basic observations that are relevant to the interpretation of the data. First, we observe a very strong gaze-cueing effect: 47.6% of the SpeakerGaze+ cases (speakers fixating their own hand gesture) are immediately followed by Interlocutor Gaze+ cases (interlocutors fixating that same hand gesture). This means that in nearly half of the cases the gaze cueing is successful.[2] Second, although the experimental set-up does not allow us to measure a causal link between gaze behaviour and gesture behaviour, we did observe a temporal contingency between the two: for each instance in our dataset, the gaze behaviour precedes the gestural behaviour. We cannot claim that the gestural alignment we measure (see below) happens BECAUSE OF the gaze behaviour, but we do want to stress that at least there is a temporal relation between the two.

[1] Note that we are fully aware that some features may contribute more to the overall perception of gestural alignment than others. For the purpose of manageable operationalization, however, we did not weigh the individual factors for the analysis.

[2] For us, SUCCESSFUL gaze cueing is a mere technical matter of co-occurring speaker and interlocutor fixations. We do not take into account the intentions speakers might have when fixating their own gestures. Those intentions might be to explicitly invite the conversational partner to look at the produced gesture as well, but speakers also might have many different reasons to focus on their own gestures. Regardless of those INTENTIONS, the gaze cueing EFFECT remains real: if speakers fixate their own gestures, the interlocutors also fixate those gestures in 47.6% of the cases.

https://doi.org/10.1017/langcog.2015.22 Published online by Cambridge University Press

6.1. GAZE BEHAVIOUR AND GESTURAL ALIGNMENT

To obtain our final result, we combined the factors InterlocutorGaze and SpeakerGaze, hence creating four possible conditions for which we calculated the average alignment scores. Figure 4 visualises the four conditions and Figure 5 shows the average alignment scores for those conditions (with respective *n* values of 44, 226, 40, and 44).[3] Data were analyzed with a 2 × 2 ANOVA test with factors SpeakerGaze(+/–) and InterlocutorGaze(+/–). This revealed a significant main effect of InterlocutorGaze ($F(1) = 41.19$, $p < .001$), reflecting higher alignment scores in the InterlocutorGaze+ conditions than their InterlocutorGaze– counterparts (i.e., if interlocutors have focused on the speaker's hand gesture, they produce more aligned gestures than if they have not focused on the speaker's gestures). The main effect of SpeakerGaze was not significant.

However, the InterlocutorGaze effect was qualified by an interaction with SpeakerGaze ($F(1) = 10.44$, $p = .001$). Follow-up comparisons indicate that this interaction resulted because the InterlocutorGaze effect was present only in the SpeakerGaze– conditions ($p < .001$). In the SpeakerGaze+ conditions, there was no such difference ($p = .26$) between InterlocutorGaze+ and InterlocutorGaze–. In other words, only when speakers don't fixate their own gestures does it matter for gestural alignment scores whether the interlocutor looks at his partner's gestures. If the speaker does fixate his own gestures, the gaze behaviour of the interlocutor is no longer associated with higher gestural alignment scores.

7. Discussion and conclusion

The main findings of this paper can be summarised in three points. First, SpeakerGaze alone (i.e., the speaker fixating his own gestures) does not affect gestural alignment. This finding ties in with what Wang and Hamilton (2014) reported in their reaction-time experiment: addressees looking at actors in a video were not faster in copying target gestures if those gestures were fixated by the actor, compared to when they were not fixated. Second, InterlocutorGaze (i.e., the interlocutor fixating the speaker's gesture) does significantly co-occur with higher scores for gestural alignment, but it only does so in the SpeakerGaze– cases (i.e., when the speaker does not fixate his own gesture). The correlation between fixation behaviour and gesture behaviour is in line with

[3] Because fixations on gestures are scarce in face-to-face conversations, the number of InterlocutorGaze–/Speakergaze– cases, i.e., the cases where participants are gesturing without either of them fixating those gestures, are overwhelmingly available in the corpus. Therefore, for this condition, we randomly selected a subset of 226 cases. For the other three conditions, which are less frequent because they contain a gesture fixation, we used all of the available cases in the corpus.

https://doi.org/10.1017/langcog.2015.22 Published online by Cambridge University Press

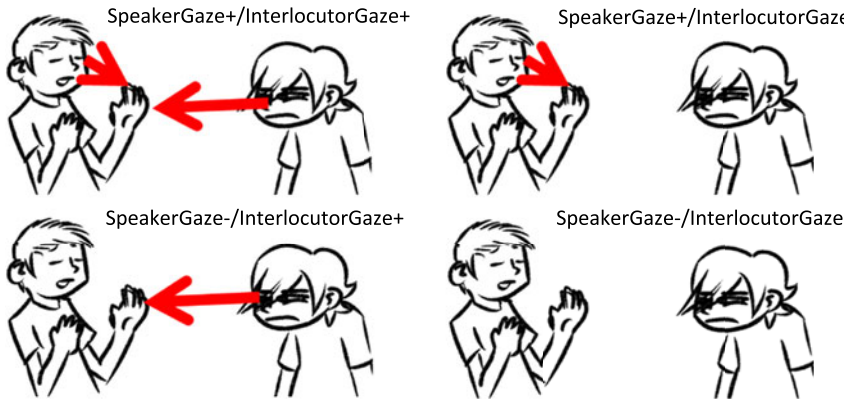

Fig. 4. Four possible conditions when combining the two factors SpeakerGaze and InterlocutorGaze.

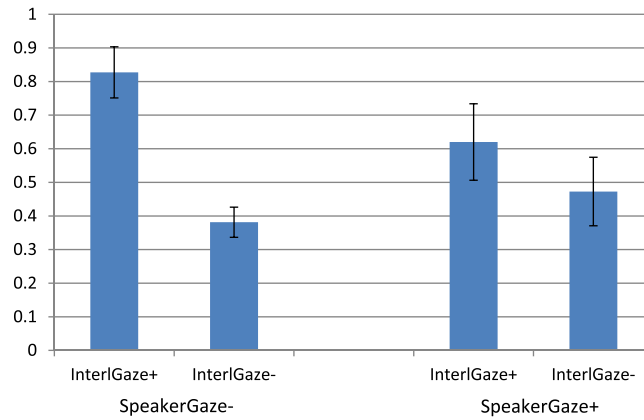

Fig. 5. Average scores for gestural alignment across two factors: InterlocutorGaze (interlocutor has or has not focused on the speaker's hand gesture) and SpeakerGaze (speaker has or has not focused on his own hand gesture). Error bars indicate standard error.

Gullberg and Kita's (2009) claim that "uptake from speaker-fixated trials with addressee fixation did not differ from speaker-fixated trials without addressee fixation" (p. 277). Third, and somewhat surprisingly, our results differ from Gullberg and Kita's observation that "addressees were more likely to retain the directional information in gesture when speakers themselves had first fixated the gesture than when they had not" (p. 268). In our findings, there is no correlation between gaze cueing and gestural alignment.

The differences between the findings reported here and those of Gullberg and Kita (2009) might be due to a number of factors, of which we discuss the

https://doi.org/10.1017/langcog.2015.22 Published online by Cambridge University Press

two most important ones here. First, there is a difference in conversation type: the results in this paper are drawn from eye-tracking participants in face-to-face conversations, whereas Gullberg and Kita start from eye-tracked participants watching videos. Moreover, the specific experimental task is different: targeted collaborative tasks here, versus a story-telling task in Gullberg and Kita. Second, although both studies use mobile eye-tracking to measure visual perception, they differ in how they link perception to the processing of perceived gestures. In other words, they differ in determining the dependent variable: gestural alignment in this paper and information uptake in Gullberg and Kita. The former measures the similarity between the gesture fixated and the subsequent gesture produced (using a five-point similarity scale); the latter measures the directional information retained from the fixated gestures (using drawings that participants made after they watched the target gestures on a video screen).

What appear to be contradictory results may, in fact, be indicative of the many functions of eye-gaze in communication and of the intricate relationship between gaze fixation and cognitive attention. Interlocutors in conversations can fixate their own or their partners' gestures for many different reasons: disambiguating, gaze cueing, signalling uncertainty, deictic referencing, etc. The eye-tracking device is only able to measure visual fixations without, of course, differentiating between these different conversational functions. The issue with interpreting the data boils down to the following: with measuring gaze data we want to tap into low-level cognitive processes (i.e., the correlation between visual perception and gesture production); however, gaze behaviour is not only driven by low-level but also by high-level processes. Marked events in the ongoing interaction or very local, specific communicative goals will just as well drive the (aligned or not) gesture production. Although there might be an automatic, low-level coupling of perception (looking at gestures) and production (using those same gestures), this can at any time be disrupted by higher-level needs. Moreover, and linked to the issue of peripheral vision (see above), there is no one-to-one relation between fixation and attention. Interlocutors might fixate a gesture without cognitively processing it, and the other way around; they might have processed it without a fixation. These issues are important in linking gaze behaviour (processing) to gesture behaviour (production) in the framework of interactive alignment: we expect gesture production to be influenced by cognitive processing of preceding gestures. However, measuring cognitive processing is only estimated by using gaze fixation, which is a key but not the only factor.

The results presented in this paper are merely a first step towards a more systematic inquiry into the tight coupling of processing and production at the non-verbal level, with a view to developing a genuine multimodal account of interactive alignment. For the first basic empirical analyses discussed here,

https://doi.org/10.1017/langcog.2015.22 Published online by Cambridge University Press

we necessarily ignored a whole range of parameters that may influence addressees' (overt) attention to an interlocutor's gesture or the likelihood of establishing aligned representations at the non-verbal level. McNeill (2006) and Gullberg and Kita (2009) provide an overview of such factors, including social status, interpersonal stance, speaker information structure, shared common ground, and the physical properties of the gesture. Apart from those factors, also the time difference between the fixation onset and gesture onset, the fixation duration, co-occurring verbal cues, the number of preceding gestures that were or were not fixated, etc. might be parameters with explanatory potential as well. Future work will need to look in more detail into the impact of each of these parameters to provide a more accurate, and fully interactionally grounded account of multimodal alignment.

REFERENCES

Argyle, M., & Cook, M. (1976). *Gaze and mutual gaze*. London: Cambridge University Press.

Bavelas, J., Coates, L., & Johnson, T. (2002). Listener responses as a collaborative process: the role of gaze. *Journal of Communication*, **52**, 566–580.

Bolinger, D. L. (1961). Syntactic blends and other matters. *Language*, **37**(3), 366–381.

Branigan, H., Pickering, M., & Cleland, A. (2000). Syntactic co-ordination in dialogue. *Cognition*, **75**, 13–25.

Branigan, H., Pickering, M., McLean, J., & Cleland, A. (2007). Participant role and syntactic alignment in dialogue. *Cognition*, **104**, 163–197.

Brennan, S., Chen, X., Dickinson, C., Neider, M., & Zelinsky, G. (2008). Coordinating cognition: the costs and benefits of shared gaze during collaborative search. *Cognition*, **106**, 1465–1477.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **22**, 1482–93.

Bressem, J. (2008). Notating gestures – proposal for a form based notation system of coverbal gestures. Unpublished manuscript, European University Viadrina.

Brône, G., & Oben, B. (2015). InSight Interaction: a multimodal and multifocal dialogue corpus. *Language Resources and Evaluation*, **49**, 195–214.

Carter, R. (1999). Common language: corpus, creativity and cognition. *Language and Literature*, **8**(3), 195–216.

Cassell, J., Torres, O. E., & Prevost, S. (1999). *Turn taking vs. discourse structure: how best to model multimodal conversation*. The Hague: Kluwer.

Chartrand, T., & Bargh, J. (1999). The chameleon effect: the perception-behavior link and social interaction. *Journal of Personality and Social Psychology*, **76**, 893–910.

Cienki, A., & Müller, C. (2008). *Metaphor and gesture*. Amsterdam/Philadelphia: John Benjamins.

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

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). Washington, DC: American Psychological Association.

Condon, W., & Ogston, W. (1971). Speech and body motion synchrony of the speaker–hearer. In D. Horton & J. J. Jenkins (Eds.), *The perception of language* (pp. 150–184). Columbus, OH: Charles E. Merrill.

Dale, R., Kirkham, N., & Richardson, D. (2011). How two people become a tangram recognition system. *Proceedings of the European Conference on Computer-Supported Cooperative Work*. Berlin: Springer Verlag. Online: <http://www.researchgate.net/publication/228506291_How_two_people_become_a_tangram_recognition_system>.

https://doi.org/10.1017/langcog.2015.22 Published online by Cambridge University Press

Du Bois, J. (2010). Towards a dialogic syntax. *Unpublished manuscript*, University of California, Santa Barbara.
Emery, N. J. (2000). The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience and Biobehavioral Reviews*, **24**(6), 581–604.
Fried, M., & Östman, J.-O. (2005). Construction grammar and spoken language: the case of pragmatic particles. *Journal of Pragmatics*, **37**(11), 1752–1778.
Frischen, A., Bayliss, A. P., & Tipper, S. P. (2007). Gaze cueing of attention: visual attention, social cognition, and individual differences. *Psychological Bulletin*, **133**(4), 694–724.
Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences*, **8**, 8–11.
Garrod, S., & Pickering, M. J. (2009). Joint action, interactive alignment, and dialog. *Topics in Cognitive Science*, **1**, 292–304.
Giles, H., Coupland, N., & Coupland, J. (1992). Accommodation theory: communication, context, and consequence. In H. Giles, J. Coupland, & N. Coupland (Eds.), *Contexts of accommodation: developments in applied sociolinguistics* (pp. 1–68). Cambridge: Cambridge University Press.
Goldberg, A. (2006). *Constructions at work*. Oxford: Oxford University Press.
Gullberg, M., & Holmqvist, K. (2006). What speakers do and what addressees look at: visual attention to gestures in human interaction live and on video. *Pragmatics & Cognition*, **14**(1), 53–82.
Gullberg, M., & Kita, S. (2009). Attention to speech-accompanying gestures: eye movements and information uptake. *Journal of Nonverbal Behaviour*, **33**, 251–277.
Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longman.
Jokinen, K. (2010). Non-verbal signals for turn-taking and feedback. *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (LREC), online: <http://www.lrec-conf.org/proceedings/lrec2010/pdf/173_Paper.pdf>.
Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychologica*, **26**, 22–63.
Kendon, A. (2004). *Gesture: visible action as utterance*. Cambridge: Cambridge University Press.
Kimbara, I. (2006). On gestural mimicry. *Gesture*, **6**, 39–61.
Lachat, F., Conty, L., Hugueville, L., & George, N. (2012). Gaze cueing effect in a face-to-face situation. *Journal of Nonverbal Behavior*, **36**, 177–190.
Langacker, R. W. (2001). Discourse in Cognitive Grammar. *Cognitive Linguistics*, **12**, 143–188.
Lewandowski, N. (2012). *Talent in nonnative phonetic convergence*. Unpublished doctoral dissertation, Universität Stuttgart.
Lewandowski, N., & Schweitzer, A. (2010). Prosodic and segmental convergence in spontaneous German conversations. *Journal of the Acoustical Society of America*, **128**(4), 2458.
McNeill, D. (2006). Gesture, gaze, and ground. In S. Renals & S. Bengio (Eds.), *Proceedings of Machine Learning For Multimodal Interaction: second international workshop 2005* (pp. 1–14). Berlin/Heidelberg: Springer Verlag.
Mittelberg, I. (2007). Methodology for multimodality: one way of working with speech and gesture data. In M. Gonzalez-Marquez et al. (Eds.), *Methods in cognitive linguistics* (pp. 225–248). Amsterdam/Philadelphia: John Benjamins.
Muir, L., & Richardson, I. (2005). Perceptions of sign language and its application to visual communications for deaf people. *Journal of Deaf Studies and Deaf Education*, **10**(4), 390–401.
Munn, S., Stefano, L., & Pelz, J. (2008). Fixation-identification in dynamic scenes: comparing an automated algorithm to manual coding. *Proceedings of the 5th Symposium on Applied Perception in Graphics and Visualization* (APGV 08) (pp. 33–42), online: <http://dl.acm.org/citation.cfm?id=1394281&picked=prox>.
Neider, M., Chen, X., Dickinson, C., Brennan, S., & Zelinsky, G. (2010). Coordinating spatial referencing using shared gaze. *Psychonomic Bulletin & Review*, **17**(5), 718–724.
Novick, D. G., Hansen, B., & Ward, K. (1996). In T. Bunnell & W. Idsardi (Eds.), *Proceedings of the International Conference on Spoken Language Processing* (pp. 1888–1891), Philadelphia, 3–6 October.

Pickering, M., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, **27**, 169–226.
Pickering, M., & Garrod, S. (2006). Alignment as the basis for successful communication. *Research on Language and Computation*, **4**, 203–228.
Posner, M. I., Snyder, C. R., & Davidson, B. J. (1980). Attention and the detection of signals. *Journal of Experimental Psychology*, **109**(2), 160–174.
Raidt, S. (2008). *Gaze and face-to-face communication between a human speaker and an animated conversational agent – mutual attention and multimodal deixis*. Unpublished doctoral dissertation, University of Grenoble.
Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, **85**, 618–660.
Richardson, D., & Dale, R. (2005). Looking to understand: the coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive Science*, **29**, 1045–1060.
Richardson, D., Dale, R., & Kirkham, N. (2007). The art of conversation is coordination: common ground and the coupling of eye movements during dialogue. *Psychological Science*, **18**, 407–413.
Richardson, D., Dale, R., & Tomlinson, J. (2009). Conversation, gaze coordination, and beliefs about visual context. *Cognitive Science*, **33**, 1468–1482.
Roche, J. M., Dale, R., & Caucci, G. M. (2012). Doubling up on double meanings: pragmatic alignment. *Language and Cognitive Processes*, **27**(1), 1–24.
Sacks, H., Schegloff, E. A. & Jefferson, G. (1974). A simplest systematics for the organisation of turn-taking for conversation. *Language*, **50**, 696–735.
Schober, M. F. (1993). Spatial perspective-taking in conversation. *Cognition*, **47**(1), 1–24.
Shockley, K., Richardson, D. C., & Dale, R. (2009). Conversation and coordinative structures. *Topics in Cognitive Science*, **1**(2), 305–319.
Shockley, K., Santana, M. V., & Fowler, C. A. (2003). Mutual interpersonal postural constraints are involved in cooperative conversation. *Journal of Experimental Psychology: Human Perception and Performance*, **29**, 326–332.
Streeck, J. (1993). Gesture as communication I: its coordination with gaze and speech. *Communication Monographs*, **60**, 275–299.
Streeck, J. (2008). Depicting by gesture. *Gesture*, **8**, 285–301.
Sweetser, E. (2007). Looking at space to study mental spaces: co-speech gestures as a crucial data source in cognitive linguistics. In M. Gonzalez-Marquez et al. (Eds.), *Methods in cognitive linguistics* (pp. 203–226). Amsterdam/Philadelphia: John Benjamins.
Szczepek Reed, B. (2010). Prosody and alignment: a sequential perspective. *Cultural Studies of Science Education*, **5**(4), 859–867.
Tannen, D. (1987). Repetition in conversation: toward a poetics of talk. *Language*, **63**(3), 574–605.
Tannen, D. (1989). *Talking voices: repetition, dialogue, and imagery in conversational discourse*. Cambridge: Cambridge University Press.
Verhagen, A. (2005). *Constructions of intersubjectivity*. Oxford: Oxford University Press.
Wang, Y., & Hamilton, A. (2014). Why does gaze enhance mimicry? Placing gaze-mimicry effects in relation to other gaze phenomena. *Quarterly Journal of Experimental Psychology*, **67**, 747–762.
Wang, Y., Newport, R., & Hamilton, A. (2011). Eye contact enhances mimicry of intransitive hand movements. *Biology Letters*, **7**, 7–10.