Radboud University

Faculty of Arts

Study year: 2019 – 2020

Testing the prerequisites for hand gestures playing a potential predictive role in communication

A corpus study of the speech-gesture timing in English conversation

Name: Nguyen Phuc Cam Nhi

Student number: s1039840

Course: Master's Thesis (LET-TWM 400-2019-JAAR-V)

Supervisor: Judith Holler, ph.D

Secondary supervisor: Mareike Geiger, MA

July 2020

Declaration on plagiarism and fraud

The undersigned

Nhi-Cam Nguyen Phuc, student number s1039840

Master's student at the Radboud University Faculty of Arts,

declares that the assessed thesis is entirely original and was written exclusively by herself. The undersigned indicated explicitly and in detail where all the information and ideas derived from other sources can be found. The research data presented in this thesis was collected by the undersigned himself/herself using the methods described in this thesis.

Place and date:

Nijmegen, the Netherlands

13 July, 2020

Signature:

Acknowledgements

The completion of this research paper, for me, has been a long, arduous, yet intriguing and rewarding experience. In retrospect, I am extremely thankful for the abundant guidance and support (both professionally and emotionally) that I have received, without which this thesis would have never been completed. One page of acknowledgement is not enough to describe my gratitude for this valuable assistance.

First of all, I would like to express my deepest gratitude towards my two wonderful supervisors, Dr. Judith Holler and Mareike Geiger. As my internship and MA thesis supervisors, you have sparked my interest in the field of multimodal communication, thus guiding me during the enttire process of conducting this research. Judith, thank you so much for your patience, optimism, understanding, and for all your feedbacks despite your super busy schedule. The Covid-19 outbreak forced me to change my thesis topic, which was extremely stressful and frustrating for me. However, you still calmly helped me to solve the situation with positive attitude and great advice. Mareike, thank you so much for being super responsive to my constant Slack messages, also for spending time and effort to help me with the coding process, even though you are having lots of projects at hand. I have learned so much from both of you during the past 6 months at CoSI lab. I never regret sending that email to you in November, Judith, to ask for an internship at your lab.

I also want to send my deepest gratitude to Marlijn ter Bekke. Thank you for your advice on my coding procedure, statistics and literature aspects when I was super panic. I hope that my thesis could be of some help for your research project.

Last but not least, I have been forever indebted to my friends and family. They have been amazing emotional backups, who have always loved, sympathized and supported me in every stage of doing this research. Without them, the past couples of months would have been more stressful and challenging.

Mom and Dad, thank you for your love and determination to let me stay in the Netherlands despite the current situation. I know you are extremely worried, yet you both still encouraged me each day to finish my thesis.

Erik – thank you for being my greatest cheerleader, who symphathized, motivated and brought joy to me throughout this difficult period. I very much appreciate you withstanding my constant mood swing and sudden depression during the past few months.

And to Julia, Sophie and Marie, I feel super lucky to have you guys as friends during my study here, in the Netherlands. Had it not been for our board game nights, sharing sessions and gossips, I could not have found my motivation to survive the quarantine period and complete my thesis.

This accomplishment would not have been possible without all of you, and I can not think of another better word to express my gratitude beside: Thank you!

Abstract

Natural, face-to-face communication often involves rapid turn-taking sequences, with the average transition time of 200 ms between these turns. This is remarkably fast considering the 600 ms duration allocated for speech production. Therefore, language processing in communication must be both fast and predictive to resolve this timing constraint. For this to be possible, speakers must rely on both verbal and visual signals to predict the message and plan for the upcoming turn. This has lead to the assumption that gestures should be produced in anticipation of speech to facilitate predictive language processing. This study sets out to 1) investigate the speech-gesture asynchrony for both representational and non-representational gestures, and 2) prove the role of speech-gesture timing in predictive language processing. Based on 10 dyadic conversations of an English corpus, manual gestures associated with question-response (QR) sequences were annotated, along with the verbal information that are closest in meanings with these gestures. The researcher calculated the time gap for gesture onset – speech onset to examine the anticipative effect of gesture. Next, the relationship between speech-gesture asynchrony and response time in QR pairs with gestures was tested to provide evidence for the potential ability of preceding gestures in predictive language processing. The findings revealed that representational gestures and their strokes started before their lexical affiliates, yet non-representational gestures would follow their corresponding speech. Furthermore, no predictive effect was detected for speech-gesture asynchrony and the response time in QR sequences. These results thus provided further evidence for the timing relationship between gestures and their corresponding speech. However, further studies are needed to verify speech-gesture asynchrony in both representational and non-representational gestures, also the link between speech-gesture asynchrony and language processing time.

Key words: multimodal communication, predictive language processing, co-speech gesture, gesture-speech asynchrony

Table of contents

Declaration on plagiarism and fraud ..... *i*

Acknowledgements ..... *ii*

Abstract ..... *iv*

1 Background 1

1.1. Multimodality in communication 1

1.2. Multimodal language processing 3

1.2.1. Stable form-meaning mapping 3

1.2.2. Predictive language processing 4

1.3. Gesture and speech 5

1.3.1. What is a gesture? 5

1.3.2. Gesture-speech relationship 6

1.4. The present study 12

2 Methodology 16

2.1. The corpus 16

2.1.1.Experiment set-up and apparatus *16*

2.1.2. Procedure *16*

2.2. The coding process 17

2.2.1. Gesture coding *18*

2.2.2. Lexical affiliate coding *24*

2.3. Data analysis 26

2.3.1. Statistical measurements 26

2.3.2. Statistical analysis methods 26

3 Results 28

3.1. Gesture-speech asynchrony 28

3.1.1. Do representational gestures and their gesture strokes precede their lexical affiliates in the QR responses? 28

3.1.2. Do non-representational gesture precede their corresponding speech in the QR sequences?.....32

3.2. Does gesture-speech temporal misalignment influence the turn transition gap in question-answer turns? 33

3.2.1. Does the gesture-speech asynchrony of representational gestures influence response time in question-answer sequences? 33

3.2.2. Does speech-gesture asynchrony for non-representational gestures influence response time in question-answer sequence?.....34

4 Discussion 35

4.1. Gestures preceding corresponding speech 36

4.2. Gesture-speech asynchrony and its potential role in predictive language processing 38

4.3. Limitations and future directions 38

4.4. Conclusion 40

References 42

1 Background

This thesis is organized into four main sections, starting with the first part - Background, which is followed by Methodology, Results, and Discussion. The Background section provides a brief description of the research, its motivation, objective, and the important concepts or terminologies. A detailed account of the study's design and analysis procedure are then given in the next chapter – Methodology, which is followed by elaboration on the study's results in the Results section. The thesis is concluded with the Discussion section, which covers further discussion and implications based on the generated findings, as well as pinpoints the existing limitations of the study.

1.1. Multimodality in communication

Human communication is identified as a "highly coordinated activity", in which both speakers attempt to maintain mutual attention and understanding (Clark & Schaefer, 1989, p. 259). Furthermore, communication is considered a complex and multilayered phenomenon as verbal or textual information alone is insufficient in "giving a full picture" of the information exchange process between speakers (Wagner, Malisz & Kopp, 2014). In other words, human face-to-face communication is multimodal, as it exploits several different "articulators" and "modalities" such as eye gaze, facial expressions, body postures, manual gestures, and so forth to formulate a coherent message (Holler & Levinson, 2019, p. 639).

Considering the multimodality of human communication, it is hypothesized that there must be an underlying mechanism that arrange and intergrate the communicative modalities for sucessful message production and comprehension (Holler & Levinson, 2019; Pouw & Hostetter, 2016). The process of deciphering and responding to the communicative message; on the other hand, needs to be "both fast and predictive" under the tight temporal constraints of conversation (Levinson, 2016; Stivers et al., 2009). A gap of 200 ms is normally detected between two speaking turns in a conversation, which is exceptionally fast given the 600 ms duration required for speech production (Levinson, 2016; Stivers et al., 2009). This means that a speaker should be able to anticipate the delivered message and quickly prepare for his/ her response prior to the upcoming turn. Such restricted processing time is assumed to be challenging for processing multimodal communicative messages (Holler & Levinson, 2019). However, studies have shown that combining both visual and auditory signal leads to faster processing in conversation than relying solely on the verbal mode (Holler, Kendrick &

Levinson, 2018; Wu & Coulson, 2015). Especially, several of these studies pointed to the facilitative effect of gesture-speech combination, as the response time for questions accompanied by hand or head movements was significantly shorter than for questions without gestures (Holler et al., 2018; Kelly et al., 2010; Nagels et al., 2015). These findings could significantly support the predictive effect of multimodal integration, particularly the gesture-speech coordination in language production and comprehension. The question then remains over how gestures produced alongside speech can act as a stimulus for the speakers to predict and understand the underlying messages. One explanation can be that gestures should demonstrate certain semantic relation with the speech, also precede their corresponding verbal utterances (Holler & Levinson, 2019). Findings from previous studies could provide potential evidence to this assumption (e.g. Bergmann, Aksu, & Kopp, 2011; Chui, 2005; Kendon, 1993; Levelt, Richardson, & La Heij, 1985; Schegloff, 1984). Early studies upon the gesture-speech temporal relation involved observations of gesture use during conversations. For example, Schegloff (1984) investigated hand gesture and speech organization in natural English conversations, by closely comparing the hand gesture initiation with their associated speech. He then found that gestural movements were produced in relation to the rhythmic organization of the talk, and their lexical association with the verbal message (p. 273). That is, beat gestures often synchronized with the speech components which were stressed or emphasized, while iconic gestures tended to start before their affiliated lexical elements (Schegloff, 1984). Later on, quantitative studies upon gesture-speech coordination in English, Chinese, French, Portugese, and Dutch face-to-face conversations (e.g. Bergman, et al., 2011; Chui, 2005; Ferre, 2010; Rochet-Capellan, 2008; Ter Bekke, Drijvers, & Holler, 2020) further revealed that representational hand gestures (iconic and deictic) were produced in anticipation of their lexical affiliates – the words or phrases that are related to the gestures in meaning. Instead of relying on observation, these studies employed systematic annotation scheme and statistical analysis to examine the temporal occurrence of representational hand gestures and their lexical affiliates in natural conversation. The use of annotation tools in these studies could not only allow for comparison of different communicative modalities, but could also generate precision and representativeness of the speech-gesture temoral alignment, according to Bergmann et al. (2011) and Ferre (2010). Specifically, Ferre (2010) analyzed a French corpus of 6 speakers using an annotation software called Praat (Boersma & Weenick, 2009) to examine the timing relationship between iconic gestures and their lexical affiliates. Results of this study demonstrated that iconic gestures and their stroke phases often started before the lexical components depicted by these hand gestures. Leonard and Cummins (2009) also discovered

gesture-speech asynchrony for beat gestures and their lexical affiliates in the English corpus, by also adopting gesture-speech annotation techniques.

However, so far there has been no research attempt to examine gesture-speech temporal coordination and link it to with predictive language processing. This study then aims to bridge this gap, by further investigating the predictive potential of manual gestures in language production and comprehension, via examination of the gesture-speech timing in face-to-face conversations. The researcher hypothesized that gestures preceding speech in a conversational turn might provide clues for speakers to grasp the intended message then plan for their responses while the upcoming turn is still in progress. Along with this argument, it is suggested that speech-gesture asynchrony could play a potential role in predictive language processing (Holler & Levinson, 2019).

In the following sections, detailed descriptions of the important definitions and concepts for this research are covered. These include the predictive language processing theory, gesture definition and their semantic/ temporal interrelatedness with speech. Finally, previous studies upon speech-gesture timing will be carefully reviewed to provide grounds for this study.

1.2. Multimodal language processing

Holler and Levinson (2019) proposed two possible mechanisms of multimodal binding in communication to address the "tight time frames allowed in conversation", which are based on the gestalt-like principles (p. 641). The core idea behind the gestalt model is that intelocutors should engage in a "multimodal binding process", in which they gather pieces of information scattered through different communicative modalities to "derive a holistic message corresponding to a whole turn at talk" (Holler & Levinson, 2019, p. 641). Two proposed mechanisms that operate upon this binding framework are the Stable Form-Meaning Mappings and Predictive Language Processing – the investigative target of this study.

1.2.1. Stable Form-Meaning Mappings

Under this mechanism, multimodal signals co-occur, or synchronize with each other to represent one communicative meaning (Holler & Levinson, 2019). Visual modalities such as eye gaze, facial expression, body posture or hand gestures have been proven to convey specific meanings, which are interpreted alongside verbally-produced information. For example, raised eyebrows are often associated with confusion or questioning attitude (Ekman, 1979); a

combination of facial movements could indicate negation, denial or agreement (Benitez-Quiroz, Wilbur, & Martinez, 2016; Chovil, 1991); and eye movements like blinking might demonstrate listeners' understanding of the message delivers (Homke, Holler, & Levinson, 2018). Especially, meaningful patterns can be detected within the hand gestures, as they either depict the same information as their corresponding speech or carry specific pragmatic functions (Bavelas, Chovil, Coates, & Roe, 1995; Bergmann et al., 2011; Holler & Levinson, 2019). A typical example would be the conduit gesture with palm facing up, which is often used by speakers to signify delivery or receipt of information throughout the conversation (Bavelas et al., 1995; Kendon, 2004). Furthermore, Bavelas et al. (1995) presented the concept of pragmatic/interactive gestures which are utilized to facilitate the speaking turns within a conversation. As these visual channels can function as information holers, they are often executed in association with the verbal channel to convey the complete communicative message. Not only should the visual components be intergrated with the verbal ones, they should also precede speech to faciliate the "fast and predictive" processing in communication (Holler & Levinson, 2019). This means that to resolve the tight time frames in conversation, speakers should be able to guess the delivered information and plan for their response in advance, which is only possible through the early execution of the visual parts (Holler & Levinson, 2019).

1.2.2. Predictive language processing

The second gestalt-based mechanism, predictive language processing, concerns more with the temporal alignment of information delivered by different communicative modalities (Holler & Levinson, 2019). Prediction is assumed to be absolutely fundamental under the tight time constraints in conversations, as the speakers must be able to simultaneously anticipate the intended message and initiate response prior to the upcoming speaking turn (Holler & Levinson, 2019). In other words, a speaker must anticipate the underlying message and its endpoint when listening to the ongoing turn to be able to plan and produce his/her response on time (Holler et al., 2018). For prediction to take place, accordingly, multimodal signals should ideally be temporally misaligned. That is, a single or a combination of signals occur one after another (from lower to higher semantic level) to trigger a continuous bottom-up priming effect, thus feeding clues for the overall prediction. Holler and Levinson (2019) illustrate this theory via the process of asking questions, in which the temporal and semantic arrangement of different modalities generates the predictive effect for language processing. In detail, lip formation of the phonetic sound 'w' triggers anticipation of the possible upcoming word range,

which is followed by raised eyebrow and "a lifted palm-up open hand", indicating the word to be a question-related item (Holler & Levinson, 2019, p. 644). The entire process then enables the speakers to anticipate a question being produced by the other interlocutor at the phonetic and sentential level while the upcoming speaking turn is still in progress.

The predictive language processing is characterized by continuous, bottom-up-top-down interaction among the communicative modalities, which are arranged in temporal misalignment (Holler & Levinson, 2019; Pouw & Hostetter, 2016). Under this regime, it is suggested that visual signals, particularly hand gestures are often produced in anticipation of their affiliated auditory signals. The temporal alignment of gestures and speech in conversation is thus suggested to play a potential role in predictive language processing, as evidenced by the findings that questions with gestures initiated faster responses (e.g. Holler et al., 2018). Indeed, gestures initiation requires no syntactic or grammar rules like speech formation, which means that performing a hand movement is much faster than speaking a sentence (McNeill, 1992). Qualitative and quantitative studies also provided evidence in line with the assumption of gesture preceding speech (e.g. Bergmann et al., 2006; Ferre, 2010; Leonard & Cummins, 2009). That is, manual gestures can have a priming effect upon their corresponding speeh, as hand gestures are produced prior to their semantically-related verbal utterances (Bergman et al., 2006; Ferre, 2010; Kendon, 1993; Levelt et al., 1985; Morrel-Samuals & Krauss, 1992; Schegloff, 1984).

Overall, the main argument centers on the idea that predictive language processing is associated with the temporal distribution of visual and verbal signals (Holler & Levinson, 2019; Pouw & Hostetter, 2016). There are experimental findings to support this assumption; however, they still fail to establish a direct link between multimodal misalignment and prediction in conversation. This research gap would therefore be the primary focus of the study.

1.3. Gesture and speech

To seek evidence for the predictive language processing theory, this study aims to verify the temporal misalignment of communicative modalities, particularly the gesture-speech temporal coordination. This section will therefore elaborate on several gesture-related concepts including its definition and close relationship with speech in conversations.

1.3.1. What is a gesture?

Considering the main focus of this study, only manual gesture and its related concepts are discussed. Gestures generally refer to hand movements that convey certain meanings and often co-occur with speech (Kendon, 2004; McNeill, 1992). Also, hand gestures vary in forms and functions (Alibali, Heath, & Myers, 2001; Kendon, 2004; McNeill, 1992). Co-speech gestures are typically categorized into representational and non-representational gestures (Alibali et al., 2001; Abner, Cooperrider & Goldin-Meadow, 2015). According to this criteria, representational gestures refer to hand movements that represent the semantic information of the speech; therefore, include 1) iconic gestures which demonstrate concrete entities, for example: hand movement depicting the shape of an object; 2) deictic gestures which serves as referential signals, such as a hand arching with a finger pointing towards a direction (Alibali et al., 2001; Kendon, 2004; McNeill, 1992); 3) metaphoric gestures which convey abstract content, for example, a speaker drops the hand while saying "He gets down to business" to illustrate the concept of handling the job (Strabe, Green, Bromberger, & Kircher, 2011, p. 521). The non-representational gestures, on the other hand, "do not present a discernable meaning of the verbal utterance" (McNeill, 1992, p. 80). This gesture type consist of beat gestures which rhythmically synchronize with the speech; and interactive/ pragmatic gestures that function as dialogue coordinators (Alibali et al., 2001; Bavelas, Chovil, Lawrie, & Wade, 1992; Bavelas et al., 1995; McNeill, 1992).

A gesture can be segmented into several phases namely *preparation* (the hand departing from the rest position, stroke to the start of the stroke), *stroke hold* (the hand moves or remains static to express meanings), *pre/post-stroke hold* (static phase that either precedes or follows the stroke), and *retraction* (the hand comes back to the resting position) (Kita, Van Gijn, & Van der Hulst, 1998; McNeill, 2005; Seyfeddinipur, 2006). Furthermore, the stroke phase is considered the most important component while the others are optional (Kita et al., 1998; Bergmann et al., 2011; McNeill, 1992).

1.3.2. Gesture-speech relationship

According to and Kendon (2004), gestures are integrative and inseparable components of communication, as they "synchronize" with speech to "embody a single underlying meaning" (p. 1). In other words, co-speech gestures could have facilitative effect upon different aspects of language, such as language production/comprehension (e.g. Holler et al., 2018; McNeill, Cassell, & McCullough, 1994; Iverson & Goldin-Meadow, 2005), turn-taking (Holler

et al., 2018), and second language acquisition (Gullberg, 2006). Specifically, Holler et al. (2018) in her study showed that questions accompanied by gestures generated faster response, lending support to the potential role of gesture-speech coordination in turn-taking, and in communicative language processing. The facilitative effect of speech-gesture intergration was also found for speakers with visual impairment, as Iverson and Goldin-Meadow (2005) discovered that blind speakers produced gestures while speaking at the same rate as their sighted counterparts. These findings indicate the potential role of co-speech gestures in "faciliating the thinking that underlies speeaking" (Iverson & Goldin-Meadow, 2005, p. 228). Besides, gesture production is believed to be potentially beneficial for second language acquistion according to Gullberg (2006). Explanation for such assumption is that gestures are critical components of language alongside speech, and that the increasing complexity of gesture initiation might reflect the coginitive process of accumulating a language (Gullberg, 2006, p. 104). Therefore, the facilitative role of gestures in communication has lead to the assumption that gesture and speech must be semantically and temporally related (e.g. Bergmann et al., 2011; McNeill & Duncan, 2000; Kirchof, 2011; Schegloff, 1984). That is, gestures should be able to convey meanings associated with the speech, and they should be temporally aligned with the verbal information (Bergmann et al., 2011). The following sections would elaborate more on these two relationships.

1.3.2.1. The semantic relationship between speech and gesture

First of all, gesture and speech are semantically related, which means they are connected in meaning with each other (McNeill & Duncan, 2000). This semantic connection can be either "redundant" or "complementary" (Bergmann et al., 2006, p. 1). On the one hand, the verbal and gestural information can overlap each other, such as when the speaker mentions the "cutting" action while his/her hand creates a movement of holding a knife in one of the dyads. This could be referred to as gesture-speech redundancy (Bergmann et al., 2006; McNeill & Duncan, 2000). On the other hand, the gestures representing information that adds to the comprehension process of the speech, or both modalities complement one another to facilitate language processing (Bergmann et al., 2006; McNeill & Duncan, 2000). For instance, in one dyad, the speaker was talking about the "virtual set" while shaping his hands like a glass surrounding the eyes. With this depiction, the addressee could understand that the "virtual set" referred to a set of glasses. Therefore, in this case, the gesture added further meaning to the

corresponding speech, thus enhancing the message production and comprehension process (Bergmann et al., 2006; McNeill & Duncan, 2000).

Studies have provided evidence for gesture-speech semantic connection, and its facilitative effect upon language production and comprehension (Kendon, 2004; McNeill, 1992). Specifically, it has been found that representational gestures could add further clues to help the listeners identify and grasp the intended message (Beattie & Shovelton, 1999; Kelly, Ozyurek, & Maris, 2010; Holler, Shovelton, & Beattie, 2009; Riseborough, 1981). Driskell and Radtke (2003) compared the comprehension level of 84 speakers in conversation with gestures and without gestures. They discovered that listeners oftend intergrated gestures with speech to enhance their understanding of the targeted lexical items. The integration of speech and gesture in communication was later proven to be compulsory in a study by Kelly et al. (2010), which discovered that gesture-speech integrated messages initiated faster and more accurate understanding as compared to mesages constructed by only verbal or gestural components. Furthermore, representational gestures could improve the understanding of speech produced in unfavorable communicative contexts, such as in conditions with loud noise (Hoskin & Herman, 2001; Kendon, 2004; Drijvers & Ozyurek, 2017). Drijvers and Ozyurek (2017) examined to what extent iconic gesture – speech intergration could facilitate comprehension in situations with different noise-vocoding levels. Their study discovered higher understanding for speech-gesture intergrated messages as compared to messages without visual information, indicating the benefit of multimodality in communication (Drijvers & Ozyurek, 2017). Studies also reveal a facilitative effect upon language comprehension for non-representational gestures, suggesting semantic relation between these gesture types and speech. For example, beat gestures help listeners to better grasp the message by highlighting and emphasizing the important information (Wang & Chu, 2013). LlanesCoromina et al. (2018) later confirmed the influence of beat gestures with the finding that storytelling accompanied by beat gestures could significantly improve children's overall understanding.

Beside facilitating the language comprehension process, gestures play a crucial role in language production as they could function as a communicative tool alongside the verbal channel (Alibali et al., 2001; Bavelas, Gerwing, Sutton & Prevost, 2008). Indeed, gesture production is unaffected by visibility, also varies depending on the communicative situations (e.g. Alibali et al., 2001; Bavelas et al., 2008). It was also found that speakers often produced better speech quality when gestures are also executed (e.g. Rauscher, Krauss & Chen, 1996; Finlayson et al. 2003; Morrell-Samuels & Krauss, 2004). These evidence strengthen the

argument that gesture is also a communicative device which is able to deliver meaningful messages.

In short, gesture and speech are related in meanings, as evidenced by the fact that gesture-speech integration could facilitate both language production and comprehension in conversational situations. Gestures convey their own meaning alongside verbal language, thereby adding further information which helps the interlocutors to better understand the underlying messages (Bergmann et al., 2006; McNeill, 1992; Morsella & Krauss, 2004). This means that gestures can be employed as a separate communicative device by speakers to facilitate the language production and comprehension process (Alibali et al., 2001; Bavelas et al., 1995; Goldin-Meadow et al., 2001).

1.3.2.2. The temporal coordination between speech and gestures

The facilitative effect of speech-gesture coordination in language production and comprehension has lead to the assumption that these two modalities must be temporally aligned (e.g. Bergmann et al., 2006; Holler & Levinson, 2019). Efforts have been made to investigate the temporal coordination between gesture and speech, which pointed to two kinds of gesture-speech relationship: synchronization and asynchrony (e.g. Bergmann et al., 2006; Chui, 2005; De Ruiter, 2000; Ferre, 2010; Kendon, 1993; Levelt et al., 1985; Leonard & Cummins, 2009; Morrel-Samuels & Krauss, 1992; Schegloff, 1984; Ter Bekke et al., 2020). Gesture-speech synchronization refers to when the gestures are produced simultaneously with speech, whereas asynchrony happens when gestures precede the onset of their corresponding speech (Morrel-Samuels & Krauss, 1992). The earliest research attempts involved observations of gesture-speech synchronization in consideration of gesture types (Schegloff, 1984); semantic familiarity (Morrel-Samuels & Krauss, 1992); gesture-speech temporal parameters (Levelt et al., 1985) and stress location in speech (De Ruiter, 2000). Notably, most of these studies targeted the timing between the gesture strokes in representational gestures (iconic, deictic and beat) and their lexical affiliates, which are the words or phrases closest in meanings to these gestures (Bergmann et al., 2006; Chui, 2005; Ferre, 2010; Leonard & Cummins, 2009; Schegloff, 1984). This research focus was attributed to a close semantic affiliation between representational gestures and their lexical affiliates, as compared to the non-representational gestures (Bergmann et al., 2006; Chui, 2005; Ferre, 2010; Leonard & Cummins, 2009).

Observations of natural and spontaneous English conversations showed that deictic and iconic gestures are often produced prior to the onset of their lexical affiliates, while beat gestures tend to synchronize with their emphatic words (Schegloff, 1984). Levelt et al. (1985)

investigated the manifestation of gesture-speech coordination in the process of motor planning and execution, by examining the temporal alignment of deictice gestures and speech under the influence of different temporal parameters. His study involved four experiments, in which the speakers employed deictic gestures to indicate an array of referent lights. Each experimental condition was designed to control one specific influence factor, that is: Experiment 1 required participants to perform hand gestures in ipsilateral and contralateral visual fields; Experiment 2 compared the speech-gesture intergrated messages with speech-only and gesture-only constructed information; Experiment 3 varied the the number of referents to be indicated by the speakers; Experiment 4 involved manipulation of deictic gesture execution phase and examination of its influence upon the voicing latencies (Levelt et al., 1985). The final results from this study revealed that speech and gesture are temporally correlated, and pointing gestures often synchronize with their related verbal information. Furthermore, gestures preceding speech only occurred when speech is absence or the same verbal expression is used for different referential targets (Levelt et al., 1985). Speech-gesture temporal synchronization was again confirmed by De Ruiter's (2000) observatory study into deictic gestures. His study showed that this timing relationship could be influenced by the location of contrastive stress within the conversations (De Ruiter, 2000).

It can be noticed that the afore-mentioned studies mostly involved subjective observations of either face-to-face conversations or descriptive narrations to verify the speech-gesture temporal relation. Ferre (2010) later criticised this approach to be insufficient for investigation of speech-gesture timing, mostly for its failure to establish systematic gesture-speech annotation and to obtain precise statistics regarding the temporal coordination. As a result, subsequent studies attempted to address this gap by using the quantitative approach, which involves adopting annotation scheme for gestures and speech identification (e.g. Chui, 2005; Ferre, 2010; Morrel-Samuels & Krauss, 1992; Leonard & Cummins, 2009). Morrell-Samuels and Krauss (1992) were the first to quantitatively investigate gesture-speech temporal alignment, whose study annotated the iconic hand gestures and their lexical affiliates produced in 17 English narrations. They then found that gestures and speech can be either temporally aligned or misaligned, depending on the level of word familiarity (Morrell-Samuels & Krauss, 1992). Different languages were also examined to expand the existing evidence and increase generalizabilty of speech-gesture asynchrony (e.g. Bergmann et al., 2011; Chui, 2005; Ferre, 2010; Rochet-Capellan, 2011; Ter Bekke et al., 2020). In Chinese, it was found that iconic gestures would co-occur with their lexical affiliates rather than preceding them, as evidenced by 60% of gesture-speech synchronization as opposed to 36% of preceding iconic gestures in

speech (Chui, 2005). On the other hand, analysis of the English and French corpus demonstrated gesture-speech asynchrony, as the representational gestures are produced in anticipation of their lexical affiliates in face-to-face conversations (Ferre, 2010; Leonard & Cummins, 2009). Leonard and Cummins (2009) identified further evidence for gestures and gesture strokes preceding speech, particularly for iconic gestures in a small corpus of English conversations, as the onset of the gesture strokes started before their lexical affiliate onset. The speech-gesture asynchrony was also detected in French and Portugese conversation (Rochet-Capellan et al., 2008). Specifically, Rochet-Capellan et al. (2008) found that deictic gestures started before their lexical affiliates, also this speech-gesture temporal alignment would rely heavily on the number of syllables within the associated speech. Ferre (2010) also analyzed a French corpus, which consisted of six face-to-face conversations and confirmed the gesture-speech asynchrony for iconic gestures. Specifically, 95% of the iconic gestures were executed prior to their lexical affiliates, by an average of 0.82 seconds (Ferre, 2010). Similarly, the majority of gesture strokes (72%) started by approximately 0.45 seconds before their lexical affiliate onset (Ferre, 2010). To account for the gesture-speech temporal misalignment, McNeill (1992) proposed that gesture production requires no "complex grammatical encoding" like speech. Therefore, gestures require less time for execution and can begin earlier than the corresponding speech (as cited in Seyfeddinipur, p. 85).

So far, studies on gesture-speech temporal coordination primarily targeted representational gestures (iconic and deictic), yet little or no effort has been allocated towards the non-representational group (e.g. Bergmann et al., 2011; Ferre, 2010; Kendon, 1993; Levelt et al., 1985; Morrel-Samuals & Krauss, 1992; Schegloff, 1984). Speech and gesture are semantically related as they are systematically organized to express the same underlying message, yet not necessarily representing "identical aspects of it" (McNeill & Duncan, 2000, as cited in Bergmann et al., 2011, p. 1). Therefore, gestures can be produced to either resemble the speech or depict aspects that are not verbally expressed (Bergmann et al., 2011). It is then assumed that non-representational gestures, similar to the representational ones, could be produced in anticipation of their their corresponding speech in face-to-face conversations.

It is also noted that there seems to be a lack of a unified coding system for gestures and especially for lexical affiliates among the conducted studies on gesture-speech asynchrony (e.g. Bergmann et al., 2011; Chui, 2005; Ferre, 2010; Leonard & Cummins, 2009; Morrell-Samuel & Krauss, 1992). When investigating gesture-speech asynchrony in the Chinese corpus, Chui (2005) segmented the gestures into several gesture phases based on McNeill (1992) and Kendon (2004) definition, yet provided no descriptions on how to identify the lexical affiliates.

Ferre (2010), in her study, outlined general principles for hand gesture annotation, which involved identifying the gestural configurations then segmenting them into smaller units based on the frame-by-frame marking method. As for lexical affiliate annotation; however, no specific coding rules were provided except for the term definition (Ferre, 2010). Bergmann et al. (2011) adopted a similar coding scheme for hand gestures annotation, also provided a more detailed description of lexical affiliate annotation rules. The fact that there has been no universal annotation systems for hand gestures and lexical affiliates could influence the representativeness of the evidence produced for speech-gesture asynchrony. In order to achieve a certain level of universality in gesture-speech annotation, as well as to generate precise results regarding speech-gesture temporal alignment, a clear and universal annotation system for both gestures and lexical affiliates is required (Ter Bekke et al., 2020).

Holler and Levinson (2019), when discussing underlying mechanism in multimodal communication, suggested that speech-gesture temporal relation could be a prerequisite for predictive language processing in communication. This means that gestures preceding speech could allow the speakers to grasp the intended message and plan for timely responses while the upcoming is in still process. However, so far there have been few attempts to investigate the facilitative effect between speech-gesture asynchrony and predictive language processing. Most of the mentioned studies only targeted the gesture-speech timing in different language corpus (e.g. Bergmann et al., 2006; Chui, 2005; Ferre, 2010; Leonard & Cummins, 2009). This research then sets out to bridge this untouched research area, as well as the afore-mentioned gaps.

1.4. The present study

Considering the studies conducted upon gesture-speech temporal coordination, most of these targeted representational gestures such as iconic and deictic gestures, and the timing between gesture strokes and their lexical affiliates (e.g. Bergmann et al., 2011; Chui, 2005; Ferre, 2010; Leonard & Cummins, 2009; Morrel-Samuels & Krauss, 1992; Rochet-Capellan et al., 2008, Schegloff, 1984). Previous researchers chose to study representational gestures due to the close semantic affiliation, or "explicit affiliation" between this gesture type and their corresponding speech (Bergmann et al., 2006; Ferre, 2010; Morrel-Samuels & Krauss, 1992; Leonard & Cummins, 2009). Representational gestures are also much faster to be initiated; therefore, they would start before the corresponding speech and provide clues for speakers to grasp the intended message (Ter Bekke et al., 2020). However, this study proposed that non-

representational gestures could also precede their affiliated speech. According to McNeill and Duncan (2000), gestures can either closely resemble the speech or represent aspects that are not verbally expressed, yet both gestures and speech still synchronize to deliver the same underlying message. Gesture initiation also involves no syntactic rules as speech production, which means that performing a hand movement is much faster than speaking a sentence (McNeill, 1992). Given speech-gesture synchrony and ease in production, it is therefore believed that non-representational gestures, similar to representational gestures, can also be produced in anticipation of the associated speech. As mentioned previously, no effort has been made to investigate the role of speech-gesture temporal misalignment in predictive language processing. This will be another focus of the paper. Specifically, the researcher hypothesized that gestures would often precede their corresponding speech in question-answer pairs, and that gesture-speech asynchrony could have an influence upon the speakers' response time in these turn-taking sequences. This is because gestures preceding speech could provide potential clues for the speakers to anticipate the intended message and plan for their responses while the upcoming turn is still in progress (Ter Bekke et al., 2020). It is therefore assumed that the earlier a gesture appear before its corresponding speech in a turn, the faster a response is initiated. Question-response sequence was thereby chosen as the subject of this research due to its prevalence across languages and representativeness of a turn sequence (a response is obligatory once a question is given) (Geiger, 2019; Holler et al., 2018; Stivers et al., 2019). It is also noted that there seems to be a lack of systematic and unified annotation scheme for gestures and lexical affiliates in the previous researches, which could be a challenge for researchers to generate universal and representative results concerning speech-gesture temporal alignment (e.g. Bergmann et al., 2011; Chui, 2005; Ferre, 2010; Morrel-Samuels & Krauss, 1992; Rochet-Cappellan, 2008).

Considering the above-mentioned research gap, this study then aimed to investigate the speech-gesture timing without restriction to the gesture types in English, face-to-face conversations. Specifically, the study set out to verify the previous findings that representational gestures are produced in anticipation of their lexical affiliates, also to unravel a similar temporal misalignment for non-representational gestures. Via studying the speech-gesture timing through a quantitative approach, this study determined to generate evidence for the potential role of gestural components in predictive language processing. In order to achieve these objectives, the study examined whether representational and non-representational hand gestures would temporally precede their corresponding speech in an English corpus of face-to-

face conversations. Specifically, the following subjects were annotated: manual gestures associated with the question-answer sequences produced within the corpus, and their lexical affiliates or the words/ phrases that were gesturally conveyed. Notably, the lexical affiliates are annotated only for representational gestures, since non-representational gestures often convey pragmatic rather than semantic information of the speech (Kendon, 2004; McNeill, 1992). Moreover, the stroke conveys the meaningful part of the verbal utterance, it should be closely coordinated with their co-expressive speech (McNeill, 2005; Seyfeddinipur, 2006). Such semantic relation between speech and gesture suggests that there should be a negotiation between the two modalities concerning their time course of execution (Seyfeddinipur, 2006). It is, therefore, logical to assume a temporal relation between the stroke phase of representational gestures and their lexical affiliates (McNeill, 2005; Seyfeddinipur, 2006). The researcher then calculated the temporal duration from the gesture onset to the lexical affiliate onset. Furthermore, to seek evidence for the predictive potential of gesture-speech temporal alignment, speakers' response speed in QR sequences were targeted. Gestures preceding speech would assumably generate faster response to questions, or shorter temporal gap between a question and its answer. The researcher thus compared gesture-speech timing with the temporal gap in between the question-answer turns.

If there is indeed gesture-speech asynchrony in the English corpus, a high frequency of manual gestures preceding their corresponding speech should be detected in question-answer sequences of English conversations (e.g. Bergmann et al., 2011; Ferre, 2010; Leonard & Cummins, 2009; Rochet-Capellan et al., 2008). For representational hand gestures, the gesture onset and their stroke onset are expected to temporally precede their lexical affiliates onset in the question-answer sequences. The gesture-speech timing relationship should also be detected for non-representational gestures, which is likely to differ from the representational ones, since the former is easier to execute as compared to the latter (Loehr, 2004).

If gesture-speech asynchrony plays a role in predictive language processing, a significant correlation can be detected between gesture-speech asynchrony in QR sequences and the speakers' response rate. This means that the earlier a question/ response is preceded by hand gestures, the shorter the time gaps in the question-answer pairs would be detected. In short, findings from this research are expected to contribute to the existing evidence of temporal coordination between gesture and speech, thus providing evidence for the potential role of visual signals in predictive language processing. Besides, this study aspires to establish and implement a systematic quantitative corpus study upon speech-gesture temporal alignment in

the English corpus, which involves a detailed gesture-lexical affiliates annotation scheme. The hypothesis established for this study then allowed the researcher to formulate the following research questions:

1) *Do representational gestures precede their lexical affiliates in question-response turns in face-to-face English conversation?*
2) *Do the stroke phases of representational gestures precede their lexical affiliates in question-response turns in English conversation?*
3) *Do the non-representational gestures precede their corresponding speech in question-response turns in face-to-face English conversation?*
4) *Does the gesture-speech asynchrony influence the time gap in question-response turns in face-to-face English conversations?*

2 METHODOLOGY

This section describes how the study was conducted, which involves the subject of analysis, the coding procedure and the statistical analysis process.

2.1. The corpus

The study analyzed the Eye-Tracking in Multimodal Interaction Corpus (EMIC) established by Holler and Kendrick (2015). This corpus consists of 10 groups of participants engaging in both spontaneous dyadic and triadic conversations in English, with each lasting for 20 minutes. In total, the corpus involves 10 triadic and 10 dyadic conversations, which amounts to 400 minutes of conversational speech. The EMIC was established and recorded at the Max Planck Institute for Psycholinguistics in Nijmegen, the Netherlands (Holler & Kendrick, 2015). All participants were English native speakers living or studying in Nijmegen and were acquainted with each other prior to the experiment. Their ages ranged between 19-68 years, with a mean of 30 (Holler & Kendrick, 2015).

2.1.1.Experiment set-up and apparatus

The conversation and recordings took place in a soundproof room equipped with professional lighting suitable for high-quality audio and video recording. Participants were situated in standard height chairs with armrests, arranged in a triangle with the chair equidistantly placed from one another. Each participant was required to wear a pair of eye-tracking glasses, with a headphone to record their voices. Three high-definition video cameras (Canon Legria HFG10, 25 fps) were utilized to record each of the participants' visual behaviour from a frontal view. Check the image below for an illustration of the experiment layout. Also, for further details of the experiment apparatus and the audiovisual information, check Holler and Kendrick (2015, p. 26).

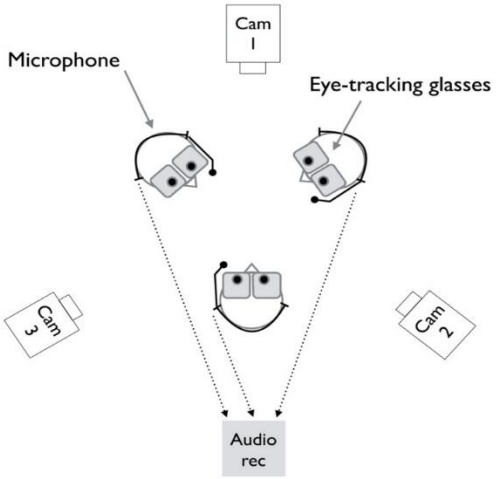

Figure 1. Illustration of the laboratory set-up used in the study (Holler & Kendrick, 2015, p. 98)

2.1.2. Procedure

The participants (in a group of three) were engaged in a 60-minute casual, unscripted conversation consisting of several steps. Initially, the researcher first welcomed the participants and briefly introduced the purpose of the study and the overall procedure. Next, the participants were provided with a study pack after greetings from the researcher. This study pack included information about the study and procedure of the session; language background forms, screening questionnaires ruling out motor and speech impairments, consent form and a questionnaire about handedness. After completing the study pack, the experimenters instructed the participants to put on the glasses and microphones. Each triad then engaged in a 40-minute conversation, with 20 minutes of trialogue followed by 20 minutes of dialogue, in which one participant would leave the group. The participants were allowed to talk about any topics, as long as they maintained the conversation within the given time constraints (Holler & Kendrick, 2015). During the recording process, the experimenters left the room and only returned once the session was over to confirm participants' research participation consent and deliver the financial compensations.

2.2. The coding process

To seek answer for the research questions, 10 dyadic conversations of the corpus were analyzed. In detail, the researcher targeted the hand gestures produced in association with the question-response (QR) sequences.

The corpus had undergone both auditory and visual coding conducted by different coders for a variety of experimental studies (Holler & Kendrick, 2015; Holler et al., 2018; Kendrick & Holler, 2017). For both the dyads and triads, the entire duration of each conversation was analysed and annotated according to different multimodal signals including the question-response pairs, speech gap, vocalisations, gesture types and such for previous research purposes (e.g. Geiger, 2019; Holler et al., 2018). In this study, the researcher proceeded to pinpoint the manual gestures produced in relation to the QR sequences within the 10 dyadic conversations, as well as to annotate the gesture strokes and their lexical affiliates in these speaking turns. The coding process, which consists of gesture coding and lexical affiliate annotation, will be described in detail in the following section.

2.2.1. Gesture coding

Both the QR sequences and gestures had been annotated by experienced annotators for previous studies (Geiger, 2019; Holler & Kendrick, 2015; Holler et al., 2018; Kendrick & Holler, 2017). Annotation of these two multimodal signals were conducted using the ELAN software (Version 5.2; Wittenburg, Brugman, Russel, Klassmann, & Sloetjes, 2006). Specifically, a total of 322 question-response pairs, along with 675 gestures were identified by M.Geiger and K.Kendrick (with reliability coding done by other coders). The QR sequences annotation adhered to the coding scheme suggested by Stivers et al. (2009), which employed both formal and functional criteria to identify the questions. As for gesture identification, communicative head or hand movements were annotated and categorized into specific gesture types namely: iconic, deictic, metaphoric, and pragmatic/interactive gestures (Holler et al., 2018). Furthermore, the precise onset and offset time of each gesture was identified (Geiger, 2019).

2.2.1.1. Identifying hand gestures associated with speech

As mentioned previously, head or hand gestures that appeared to convey meanings were identified and categorized into either representational (iconic, deictic, metaphoric gestures) or

non-representational gestures (interactive, pragmatic, beat gestures) by experienced coders in previous studies (e.g. Geiger, 2019; Holler et al., 2018). For this research, a frame-by-frame marking method was adopted to determine the onset and offset of each gesture. Specifically, the gesture onset starts with the first frame in which the hand(s) departs from the rest position while the gesture offset is signified by the first frame in which the hand moving back to the rest position (Ferre, 2010). The coding relliability for gestures in 10 dyads was verified by 76.7% of agreement between two independent annotators (Geiger, 2019).

In this study, the researcher focused on the manual gestures that were produced in association with the QR sequences then divided them into several gesture phases namely "preparation, pre-stroke hold, stroke/ stroke hold, post-stroke hold and retraction" based on Kendon (1980) and Kita et al. (1998) definition, using the ELAN annotation software (Version 5.5, Lausberg & Sloetjes, 2009). It should be noted that the phase segmentation process was applied for representational gestures only. For the hand gestures to be confirmed as part of the QR turns, they need to manifest the following critieria:

i. temporally precede or overlap with the questions or responses;
ii. be semantically or pragmatically related to their corresponding questions or responses;
iii. gestures not involving clear hand movements are omitted (e.g. hands on the lap with only fingers movement)

Based on these criteria, a total of 111 hand gestures associated with questions/ answers were identified and categorized into either representational or non-represenational gestures (Kendon, 2004). The representational gestures involve gestures that resemble the semantic information in speech namely iconic, deictic and metaphoric gestures; while the non-representational gestures refer to gestures that possess pragmatic or emphatic functions like beat and interactive gestures (e.g. Alibali et al., 2001; Bavelas et al., 1995; Kendon, 2004; McNeill, 1992). Characteristics of each gesture type are described as follows:

a) Iconic gestures illustrate concrete representations that bear a certain resemblance to the objects, entities, events, or actions (McNeill, 2005).
b) Metaphoric gestures, unlike iconic ones, represent abstract concepts under the form of an occupied space. An example of metaphoric gestures is the "conduit gesture", with the speaker's palms facing up as if he/she is holding something (McNeill,

2005). This gesture often represents an idea, or message, rather than a concrete object.

c) Deictic gestures, which are often represented by the pointing hand movement, or a hand with "the extended index finger", indicate the object direction or its physical location. The pointing gestures can be used to locate either physically present or abstract objects/ locations; thereby being categorized into concrete and abstract deictic (McNeill, 2005). The latter type is regarded as metaphoric gestures (McNeill, 2005).

d) Beat or baton gestures involve speakers' hands flicking up and down, back and forth in rhythmic synchrony with the speech (Efron, 1941; McNeill, 2005). These gestures also signal the temporal locus of the discourse or the important parts which speakers want to emphasize in their speech (McNeill, 2005). It was also found that increased beat gestures frequency during speech indicate increased importance of the delivered messaged (Zappavigna et al., p. 229).

e) Interactive, or pragmatic gestures are considered to be topic-independent, which means they reveal no information about the topic of the discourse (Bavelas et al., 1995). According to Bavelas et al. (1995), interactive gestures are represented by direct orientation at the addressees of the fingers and open palms, or hand movements with reference to the addressees in conversations. Furthermore, interactive gestures serve four main functions namely (1) information delivery; (2) citing other's contribution; (3) seeking a response and (4) turn coordination (Bavelas et al., 1995, p. 397).

2.2.1.2. Gesture segmentation rules

The gesture segmentation scheme for the gesture phases in representational gestures were adapted from Kita et al. (1998) and Seyfeddinipur (2006). In general, a frame-by-frame marking procedure was adopted to annotate the beginning and ending time codes (onset and offset time) of each gesture phase (Kita et al., 1998; Seyfeddinipur, 2006). This coding system aims to generate accurate and unambiguous categorization to annotate "consistent and frame-accurate timing" of the gesture phases (Seyfeddinipur, 2006, p. 104). According to Kita et al. (1998), a gesture or gesture unit is signified by the hand's departure and return to its resting position. McNeill (2005) later proposed the term gesture phrase, referring to the smaller components of a gesture unit. A gesture phrase, or a gestural movement would start when the

hand leaves the rest position and end when the hand returns to rest position (Kita et al., 1998). One gesture phrase can be segmented into several gesture phases, namely "preparation, pre-stroke hold, stroke, stroke hold, post-stroke hold and retraction" (Kita et al., 1998; McNeill, 2005; Seyfeddinipur, 2006). Characteristics of each gesture phase is as follows:

i. *The preparation phase* is signified by the hand moving from the resting position to a point where the stroke, the expressive part of the gesture is about to be deployed (Kita et al., 1998). The resting position refers to physical locations such as the lap, table, and such. For example, as a speaker said "it is not literally on the motorway", his hand left the rest position – his lap to move towards his chest before forming a straight line to depict the motorway. The moment his hand depart from the lap till reaching his chest would be considered the preparation phase.

ii. This phase is followed by a *stroke*, the obligatory component of the gesture phase which conveys the core meaning of the gesture (Kendon, 2004; Kita et al., 1998; McNeill, 2005; Seyfeddinipur, 2006). For a gesture phase to be categorized as a stroke, it would often demonstrate "well-defined hand configuration and well-articulated movement" (Seyfeddinipur, 2006, p. 83). For example, the hand moving in a circular shape to depict the information about a circle in a conversation could be categorized as a stroke phase.

iii. The stroke phase can be either preceded or followed by a static phase, which is often titled as *pre-stroke hold* and post-stroke hold (Kita et al., 1998; McNeill, 2005). The *pre-stroke hold* happens when the speaker's hands remain motionless "in the preparation-final and stroke-initial position", and the *post-stroke hold* is signified by a significant stop after the stroke and before the retraction (Seyfiddinipur, 2006, p. 83). A *hold* can be considered a *stroke* since it can express an entire meaning, yet in a motionless manner (McNeill, 2005, p.32). Accordingly, the most meaningful phase of a gesture phrase can be either a stroke or a stroke hold (Kita et al., 1998; McNeill, 2005). For example, in a pointing gesture, the moment the finger stops and indicates the referent is considered a stroke hold. Since both the *stroke* and the *stroke hold* are able to convey the full meaning, the most important phase of a gesture phrase can be either a stroke or a stroke hold (Kita et al., 1998; McNeill, 2005).

iv. The gesture phrase then ends with the *retraction phase*, when the hand returns to its resting position (Kendon, 1980; Kita et al., 1998). Seyfeddinipur (2006) later

proposed the concept of "partial retraction", which refers to the hand movement that illustrates "increasing relaxation" and "potential direction towards the resting position" (p. 109). In other words, partial retraction involves the hand returning to its starting point but halfway shifting to the preparation of another stroke (Seyfeddinipur, 2006).

The general coding rule applied for gesture segmentation in this study involves pinpointing the transition from dynamic to static phase and vice versa, based on the image quality of each frame (blur or clear) (Seyfedinnipur, 2006). In detail, the blurred image indicates the hand in motion while image with clarity represents a moment when the hand remains static (Seyfedinnipur, 2006). This coding rule, following Seyfedinnipur (2006), can be described as follows:

(a) ***Transition from a dynamic to a static phase*** (e.g. from stroke to poststroke hold) is marked by the first frame in which the hand configuration became clear. The subsequent frame is considered to be the starting point of a static phase.
(b) ***Transition from a static to a dynamic phase*** (e.g. from prestroke hold to stroke) begins as soon as the hand shows signs of movement, or the frame turns blurred. This frame will be coded as the first frame of a new dynamic phase.
(c) ***Transition from a dynamic to a dynamic phase*** (e.g. preparation to stroke) depends on changes in direction or speed of a hand movement. That is, the first frame in which the hand either alters its direction or increases in speed is coded as the end point of the dynamic phase in progress. The susbsequent phase; therefore, is signified by the second frame with changes in hand movements.

Besides the blurry-clear principle, the gesture annotation for this study applied a set of criteria in coding specific gesture phases as follows:

i. The *preparation phase* involves the hand motioning from the rest position to the location at which the stroke is about to be formed. For example, in one dyad, the speaker raises his hand from his knee – the rest position while forming a fist in the air. This is the point from which the stroke takes place.
ii. A *stroke* is the most meaningful part of the gesture, which can be multisegment, semi-multi-segment and repetitive. In other words, the *stroke phase* is the only phase that comprise multiple, continuous directional changes or repeated hand movements. Furthermore, a gesture phase to be coded as a *stroke* needs to

demonstrate “well-defined hand configuration and well-articulated movement”, along with symmetry and uniformity in trajectory, velocity and motion (Seyfedinnipur, 2006). Consider the hand movement of the speaker in one dyad as he said “you know those old rubbers, those black board rubbers” and attempts to demonstrates the “rubbers”. His hand, from the preparation phase, formed a fist then moved up and down continuously to mirrors the rubbing action. All the repetitive movements, which are constant in speed and direction, are annotated as one single stroke unit (e.g. Kita et al., 1998; Seyfedinnipur, 2006).

iii. It can be difficult to separate the end of *preparation phase* and the beginning of the *stroke phase* as both are dynamic. In this case, meaning and velocity (speed) are utilized to distinguish these two gesture phases. The stroke phase can start even before the complete form is established, as long as the meaning is present. In other words, the first frame in which the meaning of the gesture could be identified would signify the start of the stroke phase. Another way to distinguish the stroke from preparation is to look at the intensity of the exerted force on the gesture phase. When there seems to be more force exerted in one frame, include this frame as the onset of the stroke phase.

iv. A *stroke* can be preceded or followed by pre/poststroke hold, or itself can be a hold. To be categorized as a hold, the gesture phase should remain relatively static in a a non-rest position with slight drifting for more than two frames. The first frame with motionless hand movement is coded as the end of the dynamic phase, while the next still frame is coded as the start of the hold phase.

v. A gesture phrase is ideally ended by the *retraction phase,* during which the hand returns to rest position. In some cases, the hand wouldn’t move back to rest position immediately but instead engages in preparation phase for the next gesture. For this situation, the retraction would be coded as the preparation phase. For example, a speaker in one dyad demonstrated a typical example of this exceptional case. As soon as he finished the stroke phase depicting the “rubbers”, his hand quickly dropped but then went up again to initiate the next gesture. In addition, sometimes the hand would come to a halt before returning to its rest position, or it remains in a relaxing state. This movement is referred to as partial retraction (Seyfedinnipur, 2006).

For special gestures including beat/ baton and deictic gestures, the following coding rules were applied:

i. Beat and baton gesture are seggregrated into preparation and stroke if the accented movement and the preparatory movement were clearly distinguishable by one frame (40 ms) (Seyfeddinipur, 2006). In case there is no clear separation between frames, the whole beat gesture is coded as one stroke.
ii. With deictic gestures, particularly minimal pointing gestures, the single still frame between two dynamic phases is coded as the stroke phase. On the other hand, for gestures serving as temporal/spatial reference, both the dynamic and the subsequent still phase are coded as part of the stroke. This is because referential gestures are often expressed in the form of arching movement.

2.2.2.Lexical affiliate coding

The gesture phase coding stage was then followed by identifying the lexical affiliates in the question-response sequences associated with these gestures. According to Schegloff (1984), lexical affiliates refer to the word or phrase that is lexically associated with their preceding gestures. In other words, lexical affiliates can be understood as the words that are closest to a gesture in meanings (e.g. Bergmann et al., 2011; Ferre, 2010; Leonard & Cummins, 2009; Ter Bekke, 2020). These typically involve nouns or adjectives employed for narration or description, while prepositions, articles, demonstratives and numerical words are excluded (Bergmann et al., 2011). In this study, a detailed coding system was developed to identify the lexical affiliates, which adopted and expanded the rules outlined in Bergmann et al. (2011).

First of all, words and phrases in the speech should be considered as lexical affiliates as long as they conveyed the following characteristics:

i. Lexical affiliates can be a single word, a group of words or non-verbal sounds. In one dyad, the speaker asks “Don’t you have to have a degree in whatever you want to teach?”, while his hand makes an emphatic gesture then stretches across the space in front. “Whatever you want to teach” is coded as the lexical affiliate for this case;
ii. An entire compound word is counted as a lexical affiliate even though only a part of the compound is related to the gesture;
iii. The lexical affiliates would exclude prepositions (in, on, at), definite/ indefinite articles (a, an, the), demonstratives (this, that, these, those). For example, the

speaker said “Hamburg is like in the northern of Germany” with hands moving upwards to illustrate the north, “northern” was then decided as the lexical affiliate while “in” and “the” were omitted;

iv. The lexical affiliates also eliminate the numerical information, unless such information is depicted through the gestures. When a speaker said “they pay for two people” and showed two fingers, “two people” would be the lexical affiliate.

As for the coding rules, the primary principle to identify the lexical affiliates is to decide the type of information that is delivered through the gestures, such as: an action, a location, an entity and so forth. If the gesture represents an action, only the corresponding action verb is coded as the lexical affiliate. For example, one speaker said “maybe like lasers shooting out of the side”, with his hands producing a circular movement around the eyes. In this case, “lasers shooting” was considered the lexical affiliate. In detail, the lexical affiliate coding scheme can be described as follows:

i. For gestures depicting an entity namely a person, an object, or an animal, only the nouns describing this entity are coded as lexical affiliates. Other modifiers of these nouns are excluded unless the gestures represent them as well. Consider this example of a speaker in a dyad, she shaped her hand as long tube then moves up and down, while saying “that’s on a wooden pole”. “Pole” was decided as the lexical affiliate for this gesture. In some cases, a pointing gesture could be used to refer to the entity, choose the noun and omit the accompanying demonstratives (this, that, these, those). If the entity represented by gesture involve nouns and modifying lexicon like adjectives or adverbs, both the noun and the modifiers are chosen as lexical affiliate. However, the modifiers should be considered the lexical affiliate when the gesture refers to the modifying information (adjectives, adverbs) instead of the entity.

ii. For gestures that depict spatial content, the lexical components that convey the locational information would be coded as the lexical affilates. This would be similar to the speech and gestures that represent temporal information. In one dyad, the speaker said “it is like several years or whatever it is”, at the same time stretched both his hands widely to describe a duration of time. In this case, “several years” was coded as the lexical affiliate.

Once the lexical affiliates in the QR sequences had been identified based on the presented coding rules, their exact onset and offset time were adjusted and annotated using the Praat software (Version 6.1; Boersma & Weenink, 2019).

2.3. Data analysis

Once the annotation stage was finished, the researcher proceeded to conduct statistical analysis to generate evidence for the research questions.

2.3.1. Statistical measurements

The coding stage is then followed by statistical analysis, which was conducted using the Statistical Package for Social Science (SPSS) version 20.0 (IBM corp, 2010). For the purposes of this study, the following variables were calculated, namely gesture-speech timing, and response time for question/response with gestures. The gesture-speech timing involves calculating the temporal gap from the start of a gesture till the start of the corresponding verbal utterance, specifically: 1) the representational gesture onset till the lexical affiliate onset; 2) the gesture stroke (of the representational gesture) onset till the lexical affiliate onset; and 3) the non-representational gesture onset till the onset of their corresponding speech. As for the response speed, this measurement was represented by the gap/overlap between the question and response, which is the temporal interval between the question offset and their corresponding response onset. The turn transition timing measurement had been annotated and calculated by M. Geiger in her Masters thesis on the influence of co-speech gesture on turn-taking timing (refer to Geiger, 2019).

2.3.2. Statistical analysis methods

For the purpose of this study, different statistical measures were adopted in accordance with the corresponding research subject.

To investigate the gesture-speech asynchrony, the following descriptive statistics were calculated: 1) the frequency of the gestures preceding corresponding speech; and 2) the mean temporal gap of the gesture/stroke onset-lexical affiliate onset and non-representational gesture onset–speech onset. In addition, paired sample T-test was adopted to test the anticipative effect of hand gestures for their associated verbal utterances. According to Field (2014), paired sample T-test is used to compare two means coming from the same individual, or group. The purpose of this test is to examine the presence of a statistical difference between two observations of a particular subject (Field, 2014). This study thus employed paired sample T-

test to test whether the difference between gesture and speech onset is significant. However, one of the important assumptions to conduct paired T-test requires the data to be normal distributed (Field, 2014). Therefore, the average onset time of each variable (e.g. gesture onset, lexical affiliate onset) was calculated for each speaker to generate a set of data without outliers (Field, 2014).

To examine whether the response speed relies upon the gesture-speech asynchrony, the study employed the simple linear regression model (Field, 2014). Linear regression refers to a statistical method which is used to test the relationship between an independent and a dependent variable. In other words, we use linear model when we want to predict an outcome (dependent) variable from a predictor (independent) variable (Field, 2014). In this study, predictor variables included time gap for gesture/stroke onset-lexical affiliate onset, and gesture onset-speech onset, while speakers's response speed was considered the outcome variable.

3 Results

In this section, results from the statistical analysis are discussed in accordance with the established research questions, which are:

1) Do the representational gestures precede their lexical affiliates in question-response QR sequences?
2) Do the gesture strokes precede their lexical affiliates in QR sequences?
3) Do the non-representational gestures precede their corresponding speech in QR sequences?
4) Does the gesture-speech temporal alignment influence the turn transition gap in question-response turns?

3.1. Gesture-speech asynchrony

Out of the 321 question-response pairs annotated in the dyads, a total of 675 gestures were detected, with 335 representational gestures and 330 non-representational gestures. For this study, the researcher was able to identify overall 111 hand gestures (54 representational and 57 non-representational hand gestures) that are associated in meanings with either the questions or responses produced in the 10 dyadic conversations. Furthermore, in questions 32 representational and 33 non-representational manual gestures were identified. In this section, the results for data analysis will be presented in answer to each of the research questions.

3.1.1. Do representational gestures and their gesture strokes precede their lexical affiliates in the QR responses?

First of all, to investigate the assumption that representational gestures are often produced in anticipation of their corresponsing speech, the researcher calculated the frequency of gestures and gesture strokes preceding their lexical affiliates, as well as the time gap between gesture onset/ stroke onset and lexical affiliate onset. It was found that the majority (83%) of the representational hand gestures associated with question-response sequences started before their lexical affiliates, and over half (57%) of the gesture strokes preceded their lexical affiliates (as can be seen in Table 1).

	Gestures starting before lexical affiliates/speech (%)	Gestures starting after lexical affiliates/speech (%)
Representational gestures	83% (N=57)	17% (N=57)
Gesture strokes	57% (N=54)	43% (N=54)
Non-representational gestures	15% (N=57)	85%(N=57)

Table 1. Percentage of representational gestures and strokes starting before/ after their lexical affiliates

To examine speech-gesture asynchrony, the mean values for the time gap between gesture onset/ stroke onset and their lexical affiliate onset were calculated, with the value below 0 indicating gesture/ stroke onset preceding their lexical affiliates and vice versa. The generated descriptive statistics then revealed that the representational gestures were often produced in anticipation of their lexical affiliates (M=-1965.5, SD=9981.85). Not only did the representational gestures start before their lexical affiliates, but their gesture strokes typically preceded their corresponding speech (M=-1589.65, SD=10009.66). This means that gestures would start by an average of 1965 ms before their lexical affiliates, and for gesture strokes the gap would be approximately 1589 ms. The distribution of the gesture/stroke-lexical affiliate onset time gap can be seen in figure 2 and figure 3. A paired sample T-test was then conducted to compare the difference between representational gestures/ gesture strokes onset and their lexical affiliate onset in each dyad. The analysis revealed significant difference for both representational gesture-lexical affiliate onset ($t(9)=1.40$; $p=0.007$) and gesture onset-lexical affiliate ($t(9)=-1.85$; $p=0.009$). In general, representational hand gestures and their sroke phases would often start before their lexical affiliates in question-answer turns, based on the data from this study.

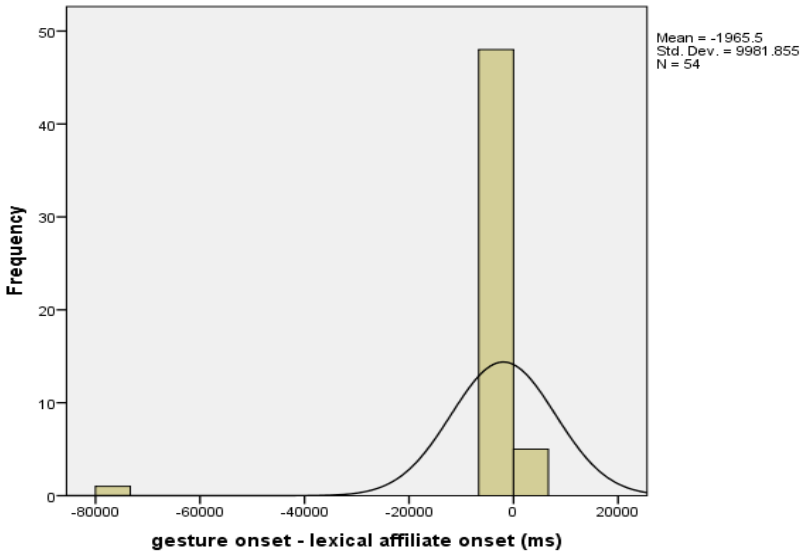

Figure 2. Distribution of gesture onset – lexical affiliate onset timing with mean and standard deviation

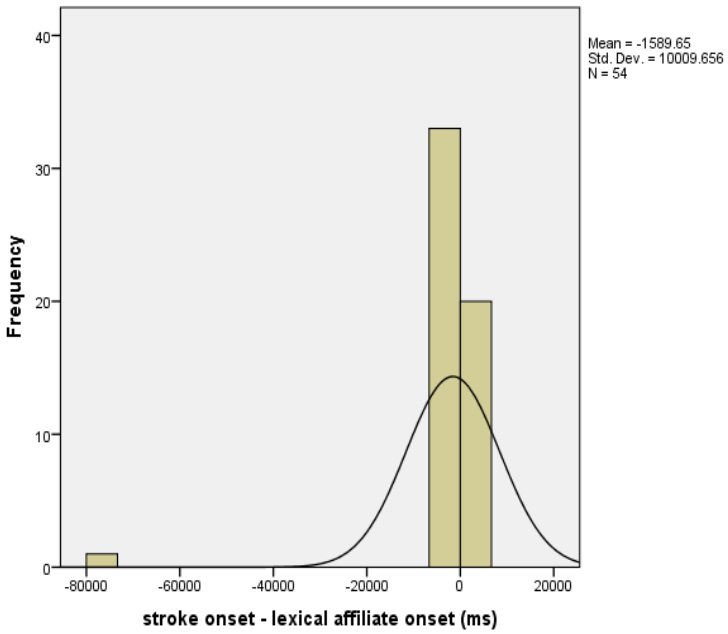

Figure 3. Distribution of stroke onset - lexical affiliate onset timing with mean and standard deviation

The interdependence of speech and gesture, or gesture-speech affiliation is often represented by the semantic relation between a gesture and the lexical component that corresponds to that gesture in meaning, or the lexical affiliates (Bergmann et al., 2011; Chui,

2005; Schegloff, 1984). However, it was argued that gesture-speech affiliation might go beyond the widely-accepted gesture-lexical affiliate relationship (Kirchhof, 2011). Indeed, both manual and bodily movements possess the capacity to convey meanings that are independent of the speech, as gestures could either resemble the verbal information (iconic, deictic) or convey abstract content not present in the speech (Efron, 1972; Ekman & Friesen, 1969; Kirchhof, 2011, Morrell-Samuels & Krauss, 1992). The semiotic relationship between gestures and lexical affiliates is thus claimed to be only a "subset" of the speech-gesture co-expressitivity (Kirchhof, 2011, p. 3). According to Kirchhof (2011), it is likely that gestures can be produced to deliver the meaning of a combination of speech signals, rather than a single lexical unit. As gestures also require no syntactical rules in their formation and execution, they should start before not only their lexical affiliates, but also the entire verbal utterance that corresponded with them (Bergmann et al., 2011; Kirchhof, 2011). Based on this assumption, another variable was examined for the purpose of this study, which was the time gap between gesture onset and their corresponding questions onset. Given the primary purpose of the study, which was to investigate the potential role of gesture-speech asynchrony in predictive processing, only questions and their associated hand gestures were targeted. Similar to the gesture-lexical affiliate temporal coordination, it was assumed that the representational hand gesture should be produced in anticipation of their corresponding question in a QR turn. A total of 32 representational hand gestures were identified to be associated with questions in the corpus; however, only 9 gestures were generated before the start of their associated questions. Similar to the analysis for gesture/ stroke – lexical affiliate timing, the mean value for gesture onset – question onset was calculated, which amounted to 1112.06 msec (M=1112.06; SD=1254.44). The distribution of the gesture onset – question onset timing can be observed in Figure 4. Paired sample T-test analysis revealed a significant difference regarding gesture onset and question onset ($t(9)=2.645$; $p=0.027$), confirming that the speakers tended to produce gestures after questions. In this study, even though gestures typically started before their lexical affiliates, gestures would follow the entire verbal utterance.

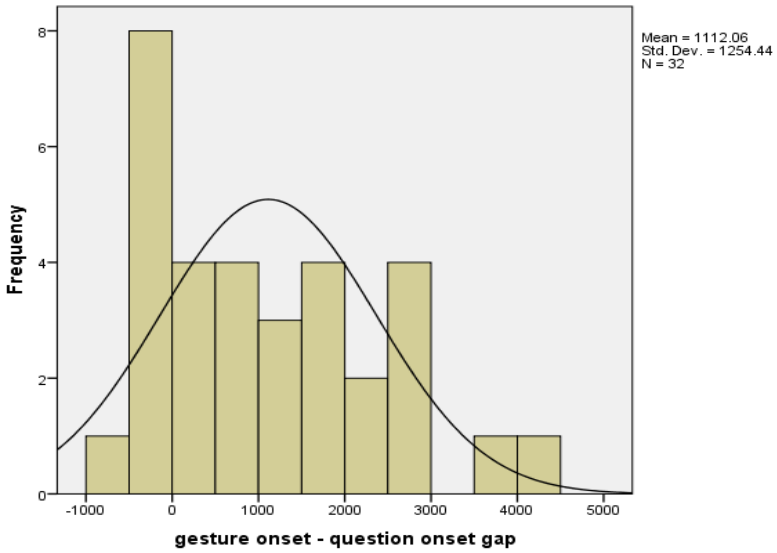

Figure 4. Distribution of representational hand gesture onset – question onset timing with mean and standard deviation

3.1.2. Do non-representational gesture precede their corresponding speech in the QR sequences?

This study also attempted to investigate the gesture-speech asynchrony for non-representational gestures to verify the predictive effect of co-speech gestures without restriction to gesture types. Similar descriptive statistical units were calculated including the frequency of non-representational gestures preceding speech and the mean time gap between gesture onset and corresponding speech onset (as can be seen in figure 5). 57 non-representational hand gestures were detected to be associated with the question-response sequences in the corpus. The descriptive analysis showed that only 15% of the non-representational gestures were initiated prior to their associated verbal utterances (as can be seen in Table 1). Furthermore, a mean value of 579.05 ms (M=579.05; SD=5293.67) was found for the average time gap between gesture onset and speech onset, suggesting that non-representational gesures would follow their corresponding speech. However, this gesture-speech asynchrony was insignificant based on the analysis from paired sample T-test (t(9)=1.882; p=0.092).

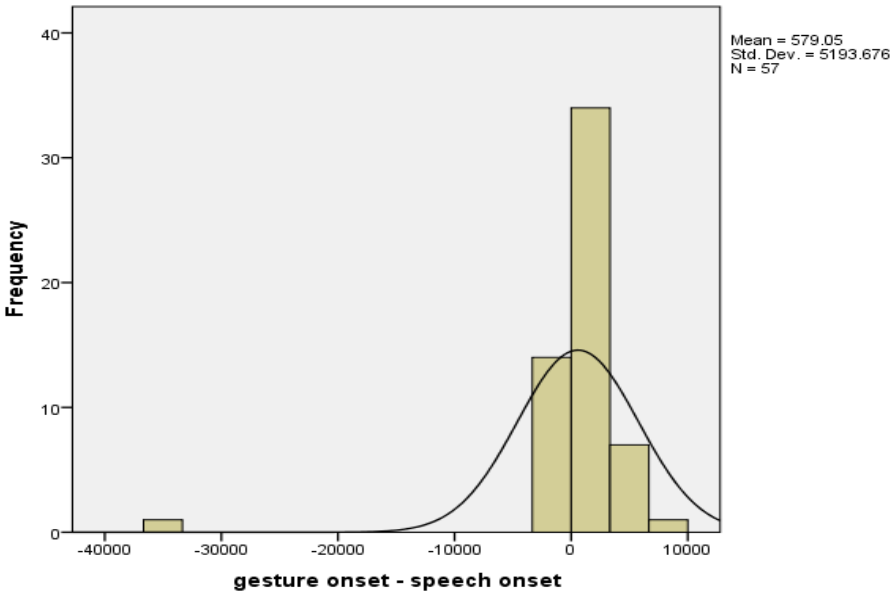

Figure 5. Distribution of gesture onset – speech onset timing with mean and standard deviation

3.2. Does gesture-speech temporal misalignment influence the turn transition gap in question-answer turns?

As mentioned previously, another primary objective of this study is to investigate and verify the potential role of speech-gesture temporal alignment in predictive language processing. That is, gestures preceding their corresponding speech could provide clues for speakers to guess the upcoming message and plan for their response while the turn is still in progress (Holler & Levinson, 2019). As a result, faster response is generated given the gesture-speech temporal misalignment. To test this assumption, the researcher conducted linear regression analysis to test the relationship between gesture-speech asynchrony and response time in turn-taking, which is represented by the gaps or overlaps in question-answer sequences. In total, speakers in 10 dyads produced 322 questions, with 321 questions having corresponding verbal responses. Among these, 54 questions/ responses were produced with representational gestures and 57 questions/ answers contained non-representational gestures. Results for representational and non-representational gestures are reported in the following sections.

3.2.1. Does the gesture-speech asynchrony of representational gestures influence response time in question-answer sequences?

For representational gestures, the correlation between gesture/ stroke – lexical affiliate asynchrony and the transition time wihin the each QR turn was analyzed. The analysis revealed an insignificant linear regression ($F(1;53)=0.37$; $p=0.545$), with correlation coefficient statistics $R^2=0.007$. It can be indicated that no significant correlation could be detected between gesture-lexical affiliate asynchrony and the response time in the QR sequences. As for stroke–lexical affiliate asynchrony, insignificant regression statistics was also produced ($F(1;53)=0.451$; $p=0.505$), along with the correlation coefficient $R^2=0.045$. The statistical results suggested that there was no significant relationship between the stroke-lexical affiliate asynchrony and turn transition time. In short, gesture-lexical affiliate timing and stroke-lexical affilicate timing did not predict response times in the data. A simple linear regression was also conducted to test the predictive effect of questions preceded by representational gestures upon their response rate. The generated results also revealed no significant correlation between question-gesture asynchrony and the transition gap in QR turns, as evidenced by insignificant linear regression statistics ($F(1;31)=1.989$; $p=0.169$, $R^2=0.062$). Again, the speech-gesture temporal misalignment for representational gestures did not predict the response rate in the QR sequences. In short, the assumption regarding the potential role of speech-gesture asynchrony in predictive language processing was not verified in this study.

3.2.2. Does speech-gesture asynchrony of non-representational gestures influence response time in question-answer sequence?

To prove that gestures preceding speech, regardless of their types, could initiate faster response in conversation, this study also examined the relationship between turn transition time in QR sequences and the temporal alignment of non-representational and their corresponding speech. A simple linear regression test was employed, which was similar to the analysis for the representational gestures. Again, insignificant regression effect was detected for these two variables ($F(1;55)=0.116$; $p=0.734$), indicating no significant relationship between the two tested variables in this model. Cooefficient statistics analysis ($R^2=0.002$) also showed that there was no correlation between non-representational gesture-speech timing and the transition time in QR turns. Therefore, gesture-speech asynchrony for non-representational gestures did not predict faster response in the question-answer sequences. Explanation for why no evidence had been found for the predictive effect between speech-gesture asynchrony of non-representational gestures on response rate will be discussed in the following section.

4 Discussion

Face-to-face conversation is multimodal as it involves the exploitation of different communicative modalities to formulate and deliver a coherent message (Holler & Levinson, 2019). Furthermore, natural/ spontaneous conversations often involves rapid turn-taking sequences, with the temporal transition between these turns lasting no longer than 200 ms (Levinson, 2016; Stivers et al., 2009). This transition gap is remarkably quick given the 600 ms duration allocated towards lexical processing and production (Levinson, 2016; Stivers et al., 2009). These time contrainsts indicate that direct communication requires speakers to multitask, that is: they must simultaneously predict the intended message and plan for their responses during the incoming turn (Holler & Levinson, 2019; Levinson, 2016). Holler and Levinson (2019) then proposed the predictive mechanism to explain the fast language processing, which emphasizes temporal combination of both visual and auditory modalities to help speakers comprehend and anticipate the communicative message. In line with this argument, it is suggested that co-speech gesture could potentialy enhance prediction in language processing (Holler et al., 2018; Holler & Levinson, 2019). Researches and studies have provided evidence for the facilitative effect of gesture-speech integration: questions accompanied by gestures could initiate significantly faster response from the speakers (e.g. Holler et al., 2018; Kelly et al., 2010; Nagels et al., 2015). The predictive potential of co-speech gestures is further supported by the speech-gesture asynchrony, as studies found that representational gestures (iconic, deictic) often start before their lexical affiliates in natural conversations of Dutch, French, English and Portugese (Bergmann et al., 2011; Chui, 2005; Ferre, 2010; Leonard & Cummins, 2009; Rochett-Capellan et al., 2008; Ter Bekke et al., 2020). These researches; however, only looked at the temporal coordination between gestures and their corresponding speech while not attempting to link the gesture-speech asynchrony with predictive language processing.

This study then aims to investigate the role of co-speech gestures in predictive language processing, by seeking answers to the following hypothesis:

1) Gestures, particularly representational hand gestures and their stroke phases tend to precede their lexical affiliates in question-response turns to help speakers predict the content of the incoming turn;
2) Not only representational gestures, but also non-representational gestures should be produced in anticipation of their corresponding speech in question-answer sequences;

3) When the gestures precede their corresponding speech, the speakers are provided with clues to process the message and plan for their response. Accordingly, there should be a correlation between gesture-speech temporal alignment and the response time. In other words, larger gesture onset-speech onset asynchrony would lead to shorter transition gap in a question-answer turn.

In order to prove these assumptions, the researcher analyzed 10 dyadic conversations in an English corpus (Holler & Levinson, 2015) to calculate the temporal gap between (representational) gesture/ stroke onset and their lexical affiliates, and to test their predictive effect upon the response time. This study also attempted to expand on previous researches upon speech-gesture asynchrony by including non-representational gestures, particularly calculating the time range from gesture onset to corresponding speech onset (Bergmann et al., 2006; Chui, 2005; Ferre, 2010; Leonard & Cummins, 2009; Rochet-Capellan, 2008; Ter Bekke et al., 2020).

4.1. Gesture preceding corresponding speech

First of all, the study aimed to verify the speech-gesture temporal misalignment in the English conversation as evidenced in the previous findings (Bergmann et al., 2006; Ferre, 2010; Rochet-Capellan et al., 2008; Ter Bekke et al., 2020). This was done by calculating the temporal gap between hand gesture onset and their corresponding speech onset in the question-answer sequences. It was then revealed from the analysis that representational gestures and their stroke phases would precede their lexical affiliates in question-response sequences by an average of up to 2000 ms. This speech-gesture asynchrony was proven to be significant for both gesture onset-lexical affiliate onset and stroke onset-lexical affiliate onset. The researcher also looked at the temporal relation between gesture onset and the associated question onset in the dyadic conversation, given the assumption that representational gestures can convey meanings beyond the lexical affiliates. The findings; on the contrary, revealed that representational hand gestures would typically start by 1200ms after the questions were initiated. This suggested that a hand gesture might precede the lexical component that it represents; however, it tend to follow the entire verbal utterance containing that lexical affiliates. Accordingly, the assumption that gestures encoding information other than their associated lexical components was not proven in this study.

In general, the findings of this study are partly in line with previous studies, which proved that representational hand gestures and their stroke phases are produced in anticipation

of their lexical affiliates (e.g. Bergmann et al., 2011; Ferre, 2010; Leonard & Cummins, 2009; Rochet-Capellan et al., 2008; Ter Bekke et al., 2020). Specifically, this study managed to produce further evidence to confirm gesture/stroke-lexical affiliate asynchrony found in English, French, Portugese, and Dutch face-to-face conversation (e.g. Bergmann et al., 2011; Ferre, 2010; Leonard & Cummins, 2009; Rochet-Capellan et al., 2008; Ter Bekke et al., 2020), thus expanding the research target to include temporal coordination between gestures and their associated questions. Considering why gestures are often produced in anticipation of their lexical affiliates, one possible explanation might be gesture formation requires no complex syntactical rules as for speech production, thus hand movement tend to be initiated more quickly than verbal expressions (Kendon, 2004; McNeill, 1992; Kita et al., 1998).

Analysis generated for non-representational gesture-speech temporal alignment; on the contrary, demonstrated the opposite situation. That is, only a few non-representational gestures started before their corresponding speech, while the majority would start at around 580 ms after the speech was initiated. This finding is contradictory to previous researches on speech-gesture asynchrony, which mostly targeted representational such as iconic and deictic gestures (e.g. Bergmann et al., 2011; Chui, 2005; Ferre, 2010; Rochet-Capellan et al., 2008; Ter Bekke et al., 2020). There are fewer studies conducted upon non-representational, yet their findings were inconclusive. For example, Leonard and Cummins (2009) discovered an anticipative effect of beat phases upon their lexical affiliates, yet in a restricted corpus. Loehr (2004) also attempted to include both representational (iconic, deictic, metaphoric) and non-representational (beat, interactive/ pragmatic) gestures while studying speech-gesture temporal coordination, and a high level of gesture-speech synchrony instead of asynchrony was then revealed. This study; however, detected gesture-speech misalignment with non-representational hand gestures following their related speech. Explanation for this outcome could be attributed to the functional aspects of these gesture types. Beat gestures are produced for emphatic and rhythmic purposes: i.e. emphasize important information in speech, or rhyme with speech progression, which means that they should be enacted after the speech initiation (Schegloff, 1984). This means that speakers would produce beat gestures while they are speaking to emphasize important points. Another non-representational gesture type, which is titled interactive/ pragmatic gestures, would often function as turn coordinating devices, i.e. to deliver new information, to acknowledge others' contribution, to seek responses or to transfer the speaking turn to another speaker (Bavelas et al., 1995). These functions; therefore, indicate that interactive/pragmatic gestures would potentially start after the speech onset since they require

speakers to already conduct the turns to be properly initiated. Furthermore, interactive/ pragmatic gestures are not semantically related to the speech, which make it unnecessary to be produced in anticipation of speech production. These explanations; however, are just speculations from the researcher and require further investigations from future studies.

4.2. Gesture-speech asynchrony and its potential role in predictive language processing

The second objective of this study is to prove the role of gesture-speech asynchrony in predictive language processing. It was hypothesized that gestures preceding their associated verbal utterances could potentially enhance speakers' ability to predict the upcoming message, which lead to faster response. In other words, ealier initiation of hand gestures before speech in QR turns should lead to faster response rate. Accordingly, the relationship between speech-gesture asynchrony and turn transition time (measured by the gap/overlap between a question-response pair) was tested. The results revealed no significant predictive effect for the two variables, indicating that speech-gesture asynchrony did not influence speakers' response rate in this study. Given the restricted number of hand gestures identified in this study (54 representational and 57 non-representational gestures), it could be difficult to generate reliable analysis to prove the predictive effect of speech-gesture asynchrony for response time. Furthermore, the response times are facilitated by different communicative factors other than gesture-speech temporal alignment, namely the length of questions and responses, syntactical structures of the language, or the question types (e.g. Hollet et al., 2018; Roberts, Torreira, & Levinson, 2015). Despite the evidence for questions with gestures initiating faster response (Geiger, 2019; Holler et al., 2018), it is inconclusive whether the timing arrangement between gesture and speech could enhance response time in turn-taking. As a result, further research with more controlled conditions is necessary to investigate the predictive effect of speech-gesture asynchrony in conversational turn-taking. The fact that no evidence was generated for the relationship between speech-gesture asynchrony and response time in question-answer pairs corroborates with analysis for Dutch conversation (Ter Bekke et al., 2020).

4.3. Limitations and future directions

In hindsight, this study managed to investigate the speech-gesture asynchrony in English corpus, via establishing and implementing a systematic coding scheme for gesture phase segmentation and lexical affiliate identification. Not only did the results confirm previous evidence of gesture/ stroke onset preceding lexical affiliate onset, but also extended the research scope to non-representational gestures. Besides, the research attempted to prove the role of gesture-speech timing in predictive language processing based on the link between gesture-speech asynchrony and response time, which has remained so far an ill-researched area.

However, there are several limitations to consider and improve for future studies on the same subject. For this research, only question-answer sequences were selected for analysis due to their controlled factor: questions often make response compulsory, which is representative of turn-taking in conversation (Geiger, 2019; Holler et al., 2018). Even though restricting the research subjects to question-answer sequences allows for more control from the researcher, it limited the analysis to only a specific type of turn-taking. As a result, it is uncertain whether the findings generated in this study can represent the speech-gesture asynchrony in conversational turn-taking in general. It is thus important that turn sequences other than question-answer pair be investigated in future studies to produce more evidence concerning gesture-speech temporal alignment.

In total, 111 hand gestures associated with question-response sequences (54 representational and 57 non-representational) were identified in the corpus, which could be rather limited due to the sheer focus on question-answer pairs. This small number of gestures, as a result, might make it challenging to generate generalizable findings. This limitation is also confirmed by a study into Dutch conversation, which detected a total of 74 representational hand gestures with lexical affiliates (ter Bekke et al., 2019). For future studies investigating gesture-speech asynchrony, again it is fundamental to target other turn-taking organizations in conversation such as turn with self-selection, backup translation, semi-interpreted talk, overlaps and such (Geiger, 2019; Holler et al., 2018; Li, 2015).

The current study did not find evidence for the predictive effect of gesture-speech asynchrony upon the response speed in QR sequences, potentially due to the intervention of other confounding factors like question/ answer duration, syntactical structure of the language and questions types (Holler et al., 2018; Ter Bekke et al., 2020). This limitation requires later research to establish a controlled experiment condition which rule out the influence of the aforementioned elements. Furthermore, it is suggested that future studies consider the

difference among gesture types (iconic, deictic, beat, interactive gestures) when investigating speech-gesture temporal alignment. Since gestures can vary in functions (Kendon, 2004; McNeill, 1992): some serve as semantic representation while some are used as turn coordinators, it would be interesting to look at how such difference affect their timing in association with speech.

4.4. Conclusion

The purpose of this study was to examine the speech-gesture temporal coordination and its potential role in predictive language processing. Based on quantitative analysis of 10 dyadic conversations in English, the study showed that:

i. For representational hand gestures, hand gestures and their stroke phases often preceded their lexical affiliates;
ii. Representational hand gestures would start after a question initiation;
iii. For non-representational gestures, manual gestures would go after their corresponding speech onset;
iv. Gesture onset-speech onset asynchrony did not demonstrate predictive ability in question-answer sequences.

The proven speech-gesture asynchrony for representational hand gestures could outline potential indication for the predictive effect of manual gesture in language processing in natural conversations. That is, representational gesture or stroke phases are situated before the corresponding information to help speakers anticipate the message about to be delivered during the upcoming turn (Holler and Levinson, 2019). Findings from this study, which are in line with previous studies on speech-gesture asynchrony, further confirm the facilitative role of gestures in language processing, as well as the multimodality of communication (e.g. Bergmann et al., 2011; Ferre, 2010; Hollet et al., 2018; Ter Bekke et al., 2020). In other words, in natural, face-to-face conversations, speakers rely on both visual and auditory signals to process the intended message, thus achieving the characteristic “fast and predictive” turn-taking of human communication (Holler & Levinson, 2019, p. 639). What this study added to the current research attempts upon the temporal coordination between speech and gesture was that it investigated both representational and non-representational hand gestures. For non-representational gestures, gesture onset following speech onset was found instead of the preceding effect. This contradicted the few studies that also examined non-representational

gesture, which discovered a high level of gesture-speech synchrony (e.g. Loehr, 2004). These findings; however, are inconclusive given the small sample size and controlled design of the study, which only focused on question - answer sequences. Another aspect that makes this study different from other researches of the same topic is that it tried to prove the facilitative capacity of gestures in predictive language based on the speech-gesture asynchrony. Last but not least, this study also established and implemented a systematic annotation system for representational hand gestures and their lexical affiliates, which was rarely observed in previous studies upon gesture-speech timing relationships (e.g. Bergmann et al., 2011; Chui, 2005; Ferre, 2010; Leonard & Cummins, 2009). Not only did the annotation scheme maximize the capacity of annotation software employed for this study like ELAN, Praat, it also allowed for precise results in terms of speech-gesture temporal alignment.

In short, this study has provided evidence to confirm the speech-gesture asynchrony for representational gestures in English conversation, which was demonstrated in previous studies on the similar topic (e.g. Bergmann et al., 2011; Ferre, 2010; Leonard & Cummins, 2009; Ter Bekke et al., 2020). What this study added to the existing researches was the inclusion of non-representational hand gestures alongside representational gestures, as well as the attempt for the role of gesture-speech asynchrony in predictive processing. Gesture onset preceding speech onset effect was found for representational gestures, but not for the non-representational ones. Furthermore, gesture-speech asynchrony did not predict the response time in question-response sequences according to the data. The achievements and limitations of this study, hopefully, could provide grounds for further empirical investigation into speech-gesture asynchrony and their role in predictive language processing. That is, further studies upon speech-gesture temporal coordination should include different types of turn-taking other than question-answer pairs. It is also essential to look at non-representational gestures to generate more conclusive findings about speech-gesture asynchrony. Finally, expanding the corpus to other languages is also suggested to seek evidence for the potential role of gestures in predictive language processing.

References

Abner, N., Cooperrider, K., & Goldin-Meadow, S. (2015). Gesture for linguists: A handy primer. Language and linguistics compass, 9(11), 437-451.

Alibali, M. W., Heath, D. C., & Myers, H. J. (2001). Effects of visibility between speaker and listener on gesture production: Some gestures are meant to be seen. Journal of Memory and Language, 44(2), 169-188.

Bavelas, J. B., Chovil, N., Coates, L., & Roe, L. (1995). Gestures Specialized for Dialogue. *Personality and Social Psychology Bulletin*, *21*(4), 394-405. DOI:10.1177/0146167295214010

Bavelas, J. B., Chovil, N., Lawrie, D. A., & Wade, A. (1992). Interactive Gestures. *Discourse processes*, *15*(4), 469-489.

Bavelas, J., Gerwing, J., Sutton, C., & Prevost, D. (2008). Gesturing on the telephone: Independent effects of dialogue and visibility. Journal of Memory and Language, 58(2), 495-520.

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

Beattie, G. W., and Barnard, P. J. (1979). The temporal structure of natural telephone conversations (directory enquiry calls). Linguistics, 17(3-4), 213–229.

Benitez-Quiroz, C. F., Wilbur, R. B., & Martinez, A. M. (2016). The not face: A grammaticalization of facial expressions of emotion. *Cognition*, *150*, 77-84. doi:10.1016/j.cognition.2016.02.004

Bergmann, K., Aksu, V., & Kopp, S. (2006). *The Relation of Speech and Gestures: Temporal Synchrony Follows Semantic Synchrony*. Retrieved from https://pdfs.semanticscholar.org/a29b/407b252775c9b93d72808d8c37be41089b55.pdf

Boersma, Paul & Weenink, David (2019). Praat: doing phonetics by computer (Version 6.1) [Computer program]. Retrieved from http://www.praat.org/

Chovil, N. (1991). Discourse-oriented facial displays in conversation. *Research on Language & Social Interaction*, *25*(1-4), 163-194. doi:10.1080/08351819109389361

Chui, K. (2005). Temporal Patterning of Speech and Iconic Gestures in Conversational Discourse. Journal of Pragmatics, 37, pp. 871--887.

De Ruiter, J. P. (2000). The production of gesture and speech. In D. McNeill (Ed.), Language and gesture (pp. 284–311). Cambridge, UK: Cambridge University Press.

Drijvers, L., & Özyürek, A. (2017). Visual context enhanced: The joint contribution of iconic gestures and visible speech to degraded speech comprehension. Journal of Speech, Language, and Hearing Research, 60(1), 212-222.

Driskell, J. E., & Radtke, P. H. (2003). The effect of gesture on speech production and comprehension. Human Factors 45(3), 445-454.

Efrón, D. (1972). *Gesture, race and culture*. King's Crown Press.

Ekman, P., & Friesen, W. V. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, *1*(1). https://doi.org/10.1515/semi.1969.1.1.49

ELAN (Version 5.9) [Computer software]. (2020). Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Retrieved from https://archive.mpi.nl/tla/elan

Enfield, N., Stivers, T., & Levinson, S. C. (2010). Question–response sequences in conversation across ten languages: An introduction. *Journal of Pragmatics*, *42*(10), 2615-2619. doi:10.1016/j.pragma.2010.04.001

Ferre, G. (2010). *Timing Relationships between Speech and Co-Verbal Gestures in Spontaneous French*. Workshop on Multimodal Corpora.

Field, A. (2014). *Discovering statistics using IBM SPSS statistics*. SAGE.

Finlayson, S., Forrest, V., Lickley, R., & Beck, J. M. (2003). Effects of the restriction of hand gestures on disfluency. Proceedings of Diss, Gothenburg Papers in Theoretical Linguistics.

Geiger, M. (2019). *Investigating the influence of gestures on the timing of turn-taking: Implications for language processing* (Unpublished master's thesis). Radboud University , Nijmegen, The Netherlands.

Gullberg, M. (2006). Some reasons for studying gesture and second language acquisition (Hommage à Adam Kendon). IRAL-International Review of Applied Linguistics in Language Teaching, 44(2), 103-124.

Gullberg, Marianne. (2006). Some reasons for studying gesture and second language acquisition (Homage à Adam Kendon). IRAL - International Review of Applied Linguistics in Language Teaching. 44. 103-124. 10.1515/IRAL.2006.004.

Holler, J., & Levinson, S. C. (2019). Multimodal language processing in human communication. Trends in Cognitive Sciences, 23(8), 639-652. doi:10.1016/j.tics.2019.05.006.

Holler, J., & Kendrick, K. H. (2015). Unaddressed participants' gaze in multi-person interaction: optimizing recipiency. *Frontiers in Psychology*, *6*. doi:10.3389/fpsyg.2015.00098

Holler, J., Kendrick, K. H., & Levinson, S. C. (2018). Processing language in face-to-face conversation: Questions with gestures get faster responses. *Psychonomic Bulletin & Review*, *25*(5), 1900-1908. doi:10.3758/s13423-017-1363-z

Hömke, P., Holler, J., & Levinson, S. C. (2018). Eye blinks are perceived as communicative signals in human face-to-face interaction. *PLOS ONE*, *13*(12), e0208030. doi:10.1371/journal.pone.0208030

Hoskin, J., & Herman, R. (2001). The communication, speech and gesture of a group of hearing-impaired children. International journal of language & communication disorders, 36(S1), 206-209

IBM Corp. Released in 2011. IBM SPSS Statistics for Windows, Version 20.0. Armonk, NY: IBM Corp.

Iverson, J. M., & Goldin-Meadow, S. (1998). Why people gesture when they speak. Nature, 396(6708), 228.

Kelly, S. D., Özyürek, A., & Maris, E. (2010). Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension. Psychological Science, 21(2), 260- 267.

Kendon, A. (1993). Human Gesture. In K. R. Gibson & T. Ingold (Eds), *Tools, language, and cognition in human evolution* (pp. 43-62). Cambridge: Cambridge University Press.

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

Kirchhof, C. (2011, September). *So What's Your Affiliation With Gesture?* , Bielefeld, Germany.

Kita, S., Van Gijn, I., & Van der Hulst, H. (1998). Movement phases in signs and Co-speech gestures, and their transcription by human coders. *Gesture and Sign Language in Human-Computer Interaction*, 23-35. doi:10.1007/bfb0052986

Lausberg, H., & Sloetjes, H. (2009). Coding gestural behavior with the NEUROGES-ELAN system. Behavior Research Methods, Instruments, & Computers, 41(3), 841-849. doi:10.3758/BRM.41.3.841.

Leonard, T., and Cummins, F. (2009). Temporal Alignment of Gesture and Speech. In Proceedings of Gespin, Poznan, Pologne (24-26 September). [CDRom].

Levelt, W. J., Richardson, G., & La Heij, W. (1985). Pointing and voicing in deictic expressions. *Journal of Memory and Language*, *24*(2), 133-164. doi:10.1016/0749-596x(85)90021-x

Li, S. (2015). Nine types of turn-taking in interpreter-mediated GP consultations. *Applied Linguistics Review*, *6*(1), 73-96. doi:10.1515/applirev-2015-0004

Llanes-Coromina, J., Vilà-Giménez, I., Kushch, O., Borràs-Comes, J., & Prieto, P. (2018). Beat gestures help preschoolers recall and comprehend discourse information. *Journal of Experimental Child Psychology*, *172*, 168-188. doi:10.1016/j.jecp.2018.02.004

Loehr, D. (2004). Gesture and Intonation. Ph.D. Thesis. Georgetown University.

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

McNeill, D. (2005). Gesture and thought. Chicago: Chicago University Press.

Morrel-Samuels, P. & Krauss, R. (1992). Word Familiarity Predicts Temporal Asynchrony of Hand Gestures and Speech. Journal of Experimental Psychology: Learning, Memory, and Cognition. 18. 615-622. 10.1037/0278-7393.18.3.615.

Morsella, E., & Krauss, R. M. (2004). The role of gestures in spatial working memory and speech. American Journal of Psychology, 117(3), 411-424.

Pouw, W. & Hostetter, A. (2016). Gesture as a Predictive Action. Reti, saperi, linguaggi: Italian Journal of Cognitive Sciences.

R Core Team. (2019). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.Rproject.org/

Rauscher, F. H., Krauss, R. M., & Chen, Y. (1996). Gesture, speech, and lexical access: The role of lexical movements in speech production. Psychological Science, 7(4), 226-231.

Roberts, S. G., Torreira, F., & Levinson, S. C. (2015). The effects of processing and sequence organization on the timing of turn taking: A corpus study. *Frontiers in Psychology, 6*. doi:10.3389/fpsyg.2015.00509

Rochet-Capellan, A., Vilain, C., Dohen, M., Laboissière, R. & Schwartz, J.-L. (2008). Does the Number of Syllables Affect the Finger Pointing Movement in a Pointing-naming Task? 8th International Seminar on Speech Production (ISSP 2008). Strasbourg, pp. 257-- 260.

Schegloff, E. A. (1984). On Some Gestures' Relation to Talk. In J. M. Atkinson and J. Heritage (Eds.), Structures of Social Action. Cambridge: CUP, pp. 266- -298.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T. & Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. Proceedings of the National Academy of Sciences of the United States of America, 106, 10587–10592.

Straube, B., Green, A., Bromberger, B., & Kircher, T. (2010). The differentiation of iconic and metaphoric gestures: Common and unique integration processes. *Human Brain Mapping*, *32*(4), 520-533. doi:10.1002/hbm.21041

ter Bekke, M., Drijvers, L., & Holler, J. (2020). The predictive potential of hand gestures during conversation: An investigation of the timing of gestures in relation to speech. Retrieved from https://psyarxiv.com/b5zq7/

Wagner, P., Malisz, Z., & Kopp, S. (2014). Gesture and speech in interaction: An overview. *Speech Communication, 57,* 209-232. doi:10.1016/j.specom.2013.09.008

Wang, L., & Chu, M. (2013). The role of beat gesture and pitch accent in semantic processing: An ERP study. Neuropsychologia, 51(13), 2847-2855.

Wu, Y. & Coulson, S. (2007). Iconic gestures prime related concepts: An ERP study. Psychonomic bulletin & review. 14. 57-63. 10.3758/BF03194028.