The role of iconic gestures in predictive language processing: Evidence from corpus analyses and anticipatory eye-movements

Junfei Hu

Centre for Language Studies, Radboud University Nijmegen

Research Master in Linguistics and Communication Sciences
MA Thesis

July 14, 2020

Supervised by: Prof. Dr. Asli Özyürek (Centre for Language Studies; Donders Institute for Brain, Cognition and Behavior; Max Planck Institute for Psycholinguistics) and Prof. Dr. Falk Huettig (Max Planck Institute for Psycholinguistics; Centre for Language Studies)

Abstract

The multimodal nature of face-to-face communication has the potential for interlocutors to take advantage of visual (e.g., gesture) and verbal (e.g., speech meaning) cues to predict the upcoming speech. This pre-registered study is dedicated to investigate the extent and way in which gestures coordinate with speech to contribute to predictive language processing in Chinese by combining multimodal corpus analysis with visual world eye-tracking experiment in lab settings. First, in a multimodal natural Chinese conversation corpus, we annotated iconic gestures (e.g., piano-playing gesture) that cooccurred with subject-verb-object sentences to depict transitive events, and associated gestures with the part of speech (i.e., lexical affiliate) which was semantically-related to them. We explored whether iconic gestures temporally anticipated their lexical affiliates. We found that gestures as a whole as well as their strokes started before the lexical affiliates, such as before related verbs and their noun arguments. Based on this finding, we further asked to what extent can iconic gesture predict the upcoming nominal word in the sentence independently from the predictive power of the linguistic input. To this end, participants' eye movements are recorded as they look at a visual display showing an actor who would perform gestures (e.g., play the piano), a target object (e.g., piano) and three distractor objects. Participants will experience four conditions whilst viewing the display: hearing "*I today played the whole afternoon piano*" in which an object in the display is predictable based on the verb's selectional restrictions (i.e., target-speech condition) or "*I today moved the whole afternoon piano*" in which all objects in the display are predictable based on the verb's selectional restrictions (i.e., neutral-speech condition) or the target sentence with a piano-playing gesture accompanied (i.e., gesture+speech condition) or "*I today hmmm.. the whole afternoon piano*" in which the verb is replaced with a schwa-like filler sound with piano-playing gesture (i.e., gesture-only condition). We expect participants cannot anticipate the target picture only in the *neutral-speech condition*. Meanwhile, in the rest three conditions,

participants can anticipate the target to different extent. Crucially, we further predict that *gesture+speech condition* should attract the most predictive looks to the target object by the time of the target object is itself heard followed by target-speech condition and *gesture-only condition*. This study will reveal the nature of gesture-speech coordination in time in natural conversation and advance our understanding about gesture-speech interaction in production. Also, it is expected to uncover the mechanism of predictive gesture-speech integration during cascaded visual and linguistic processing.

Introduction

The phenomenon that people anticipate upcoming information before encountering it is known as *prediction* (Bar, 2003; Clark, 2013; Friston, 2005). Recent psycholinguistic research assumes that anticipatory mechanisms play a crucial role during language processing (e.g., Altmann & Mirković, 2009; Dell & Chang, 2014; Federmeier, 2007; Ferreira & Chantavarin, 2018; Gibson et al., 2013; Hale, 2001; Hickok, 2012; Huettig 2015; Kuperberg & Jaeger, 2016; Levy, 2008; Norris et al., 2016; Pickering & Gambi, 2018; Pickering & Garrod, 2013; Van Petten & Luka, 2012). Many different terms (e.g. prediction, anticipation, expectation, context effects, top-down processing) have been proposed for essentially the same phenomena. Researchers have also defined prediction in language in different ways. Here we do not draw any distinction between priming, 'expectation' for anticipated semantic content (Van Petten & Luka, 2012) and 'more global forecasting', etc. We avoid arbitrary decisions about what constitutes prediction and what not and define prediction in language processing as any pre-activation of upcoming linguistic (and associated non-linguistic) representations.

So far, most studies on predictive language processing have focused on how the spoken, written and visual (pictorial) input is used for prediction (Huettig et al., 2011, for overview). However, the fact that people gesture when they talk in real world communicative settings, especially in face-to-face interactions, suggests that human communication, and thus prediction, is intrinsically multimodal not only when integrating speech with visual referents (such as common objects) but also when integrating speech and gestures. In fact, there is mounting behavioral and neural evidence that interlocutors actively and mandatorily integrate the information encoded in gestures with speech to achieve mutual understanding (Beattie & Shovelton, 1999; Drijvers & Özyürek, 2017; Kelly et al. 1999; Kelly et al., 2010; Özyürek et al., 2007; Willems et al., 2007; Özyürek, 2014, for overview). However so far very little is known about the extent and way in which gestures coordinate with speech to contribute to

predictive language that appears to be such an important part of language processing. By combining corpus-based approaches analyzing multimodal natural Chinese conversation data with visual-world eye-tracking experiment in lab settings, the current study is dedicated to fill this gap in our knowledge about multimodal prediction at the lexical level where the idiosyncrasy of multimodal human communication is obviously observed. Specifically, we ask 1) do gesture possesses the potential to be used by language users to predict the upcoming linguistic input? and 2) to what extent can gestures be used by language users to predict the upcoming word?

Cues used for prediction

A large amount of research has investigated what kind of cues are used for prediction in language processing. The regularities presented in speech (e.g., syntactic, phonological, and semantic information) are unquestionably important predictive recourses for language users (Rothermich & Kotz, 2013). Staub and Clifton (2006) used reading eye-tracking to demonstrate that participants read follow-up syntactic elements occurring immediately after "or" more quickly when they encountered "either" in the preceding context than when they did not. This outcome indicated that participants were able to predict upcoming linguistic input on the basis of syntactic knowledge. Besides syntactic structures, research on word recognition has revealed that lexical candidates sharing identical word-initial phonemes would compete radically for recognition (Norris et al., 1995). For example, on hearing the spoken sequence /bi../ embedded in a sentence such as '*Pick up the* /bi/…', all words that start with these sounds, such as beaker and beetle, are parallelly activated (see Allopenna et al., 1998, for details). Thus, unfolding phonological information can pre-activate lexical representation before the whole word is heard. Apart from these, "Selectional restriction", the semantic constraint that a predicate places on its argument (Katz & Fodor, 1963), is one type of critical semantic knowledge that considered can be used for prediction (Altmann & Kamide, 1999; Hintz et al., 2017). Altmann and Kamide

(1999) deployed a visual world eye-tracking design and presented participants with a display containing four objects (cake, toy car, ball, and toy train) along with statements such as *The boy will eat the cake* or *The boy will move the cake*. Only one of the display objects (cake) could be eaten but all could be moved. Participants tended to gaze at the target object (cake) before hearing the target word in trials in which the spoken input included a verb that required an edible patient argument (i.e., *The boy will eat the cake*). Conversely, in trials in which the verb did not have this selectional restriction (i.e., *The boy will move the cake*), saccades to the cake were launched after the word *cake* was heard. The authors thus contended that the difference in the saccadic latency between two conditions reflected, to some extent, the online influence placed on prediction by selectional restriction.

In principle, prediction is a comprehensive reflection of linguistic knowledge and of the particular visual context in which language is used, especially within the visual world (see Vulchanova et al., 2019, for discussion). Huettig and McQueen (2007) revealed that the predictive eye-movement considered to be an index of pre-activation of a certain linguistic unit was indeed mediated by the combination of relevant phonological, semantic and visual information about that particular linguistic unit. Therefore, the visual information should also be contributive to the narrowing down of the contents of prediction. It is known that visual input can sequentially activate relevant linguistic representations at varying levels (Huettig & McQueen, 2007). Hintz et al.'s (2020) recent eye tracking study uncovered that the target object and the object sharing a similar shape with the target object could both attract significantly more looks than other distractors in a visual scene before the phonological representation of the target word was activated. Their result confirms that visual information may be used to implement predictive language processing. Further, Knoeferle and Crocker's visual world eye-tracking study (2006) demonstrated that when interpreting unfolding speech inputs, participants accorded the occurring event priority over stereotypical thematic knowledge when the verb

allowed both characters in a visual scene to be the semantically-possible agents of an action. Thus, in comparison to the speech input, the visual input can not only be recruited as a recourse for language comprehension but also plays an even more influential role in, and even beyond, the predictive thematic role assignment.

Contents of prediction

Many studies have also investigated representations (the "contents") that are predicted in language. Electrophysiological studies have provided some evidence that comprehenders can pre-activate the phonological form (DeLong et al., 2005; cf. Nieuwland et al., 2018) as well as the morphosyntactic features (Van Berkum et al., 2005; Wicha et al., 2003; Wicha et al., 2004). Meanwhile, it is also not surprizing that semantic information can be predicted. After all, understanding the meaning of an utterance is the critical part of language comprehension. For instance, as mentioned in Altmann and Kamide's (1999) study, the semantically relevant knowledge of the term is activated after hearing the verb *eat*, including its appropriate patient arguments.

In addition to those traditional linguistic elements, researchers turn their eyes to the exploration of whether specific visual information such as the shape of a word's referents can be activated by language users. In an eye-tracking study, Rommers et al. (2013) presented participants with spoken sentences that were predictive of a particular target word (e.g., 'moon' in *In 1969 Neil Armstrong was the first man to set foot on the moon*) half a second before the target word was itself heard. As they listened, the participants looked at visual displays containing three distractors and one target object, which was either the target object (i.e., moon), a shape-related object (i.e., tomato) or an irrelevant control object (e.g., rice). They found that within a time window in which they could not retrieve shape information from the spoken target word, listeners were subject to fixate the target as well as the shape-related object more often

than they fixated the irrelevant control object, indicating that they had already predictively activated the shape of the upcoming word's referent.

The importance of gestures for natural communication

Above mentioned findings convince that within the visual world, both linguistic knowledge can predict visual information as well as the other way around. Thus, in predictive language processing, linguistic as well as visual information are possibly be activated. However, as the previous section indicates, much of previous research has focused solely on speech or written comprehension and the interplay between static visual information and speech input. In fact, speakers in face-to-face communication also use gestures that carry semantic information relevant to what they are saying. This includes iconic gestures, which visually represent the physical, kinematic, or spatial characteristics of a referent (McNeill, 1992), such as mimicking piano-playing motions when saying '*I like to play the piano*'. Such gestures provide extra cues for comprehension. For example, if piano-playing gesture begins before and overlaps with the verb *play* and its associated noun argument *piano*, it can already activate the semantic representation of piano before hearing the word *piano*. Since gesture, in nature, is a kind of visual information depictive of semantic content, it is legitimate to expect that it can influence predictive language processing as the abovementioned pictorial cures do. For functioning so, two prerequisites need to be satisfied: on the one hand, people must be able to extract information from gesture and integrate it with the cooccurring speech; and secondly gesture should be temporally realized earlier than the semantically-relevant parts of speech for ecologically fulfilling the timing requirement of prediction.

For the first prerequisite, in the past two decades, the field of gesture inquiry has accumulated evidence indicating that gesture and speech form an integrated system of communication (Kelly et al., 2010; McNeill, 2015). Interlocutors, especially in face-to-face contexts, extract information from both gestural and verbal channels and incorporate them in

comprehension. For example, semantic information from iconic gestures can influence speech comprehension (Kelly et al., 1999; Kelly et al., 2010; McNeill, et al., 1994; Holler et al., 2009; Goldin-Meadow & Sandhofer, 1999; Singer & Goldin-Meadow, 2005). Goldin-Meadow and Sandhofer (1999) reported that adults had a better understanding of children's narration if the children supplemented their speech with iconic gestures. Beattie and Shovelton (1999) showed participants pre-recorded videos that either contained speech only or had both gesture and speech presented together. After that, they let participants answer questions about the size and position of the objects occurred in the videos. Participants remembered the size and position more accurately when gestures were presented in the stimuli and conveyed additional information to the speech. By adding face-to-face-talking condition to Beattie-and-Shovelton's (1999) design, Holler et al. (2009) stepped further to observe that even in face-to-face context where the gesture usually received less attention relative to watching the pre-recorded-video on the small screen (28[••]), participants were still capable to grasp the additional information conveyed by gestures, and answered the size and position information of the objects even more precisely relative to speech/gesture-only conditions. Kelly et al. (1999) further showed explicitly that listeners were able to incorporate information conveyed through iconic gesture with speech to understand an utterance's intended meaning. They showed participants videos in which gestures conveyed more information than the content of the speech (e.g., pantomiming playing basketball by performing shooting gesture while speaking the sentence '*My brother went to the gym*'). When the participants were asked to write exactly what they had heard, 23% of them wrote like or similar with, 'My brother went to play basketball'. Not only under the above-mentioned ideal listening contexts, but even in the adverse communication situation, iconic gestures positively contribute to comprehension. By manipulating whether participants could see the gestures or not when they heard the speech as well as manipulating the noise-level of the speech, Drijvers and Özyürek (2017) showed that in the same noise condition,

participants comprehended the speech more precisely if they could see the gesture versus not. Obviously, these accumulated evidence indicate that during the course of language processing, listeners integrate the messages from both channels.

Not only did the behavioral studies reveal that interlocutors were capable to extract meaning from gestures and integrate them with the concurrent speech, but the neuroscientific research uncovered that such extraction and integration had a neural basis. Neurophysiological research has demonstrated that brain areas involved in speech meaning processing were also activated when individuals comprehend gestures, allowing for greater ease of speech comprehension and lexical access (Willems et al., 2007; 2009; Straube et al., 2011; Green et al., 2009; Dick et al., 2014; Demir-Lira et al., 2018; Drijvers et al., 2018; see Özyürek, 2014, for overview; cf. Holle et al., 2008; Dick et al., 2009). Drawing on functional magnetic resonance imaging (fMRI), Willems et al. (2007) found that the condition in which an iconic gesture was not semantically in line (i.e., mismatching) with the preceding context elicited activity in the left inferior frontal gyrus (left IFG), which is considered crucial for the integration of semantic information into a previous context (Hagoort, 2003, 2005; Hagoort et al., 2004; Lau et al., 2008). When gesture did not contradict with the context but added extra information to the speech, more brain regions were also involved in processing gesture and speech (e.g., the left inferior frontal gyrus triangular, opercular portions, and left posterior middle temporal gyrus; see Dick et al., 2014, for details). Drijvers et al. (2018) , by manipulating the auditory conditions (i.e., clear vs. degraded speech input), also showed that gesture's disambiguation of noisy speech engages areas involved in language comprehension.

Taken together, these results demonstrate that human brain allows the processing of integrating speech and gesture information in comprehension, and gesture may play a role in (pre-)activating information in a predictive manner. Yet even though it has become clear that gesture and speech constitute an integrated system of language communication, which is

supported by overlapped neural systems, it remains unclear whether the semantic information obtained from iconic gestures plays a role in the predictive processing of speech.

The temporal relation between gesture and speech

In order to investigate whether gesture plays a predictive role in language processing, one needs to find out that gesture can precede their semantically relevant part of speech (i.e., lexical affiliate; Schegloff, 1984). Observational studies (Streeck, 2009a; Schegloff, 1984; Kendon, 1980; see Wanger et al., 2014, for overview) have indeed demonstrated that a majority of the meaningful parts of gestures tend to be realized earlier than their lexical affiliate. The temporal asynchrony of gesture-speech coordination at the lexical level on the one hand provides listeners a chance to predict the incoming verbal input based on the message extracted from the speaker's gesture. Furthermore, it endows gestures with a potential to facilitate language processing, which is called "predictive potential" (ter Bekke et al., 2020). For instance, when an addressee hears '*I like to play the …*' seeing the speaker performs a piano-playing gesture preceding or cooccurring with the verb *play*, the addressee can probably guess, regardless of other linguistic cues such as prosodic features and syntactic structure, the potential follow-up argument will be *piano* or at least something that can be strummed by fingers.

One piece of evidence for the predictive potential of gesture comes from the exploration of joint turn construction (Lerner, 2002; see Hayashi, 2013, for overview). That is, in conversation, the addressee sometimes will join in the construction of the addresser's utterance by speaking part of it, either alone or together with the addresser. Hayashi (2005) noticed that in a natural Japanese conversation about how an individual should dress up, when the addresser completed the gesture of tying a bowtie without yet starting to pronounce the word *bowtie,* the addressee said *bowtie* immediately after seeing the gesture, although the addressor still held the right to the turn at that time. The addressee's utterance grammatically fit the in-progress utterance of the addresser. In this case, the addressee understood the information from the

gesture and successfully guessed that *bowtie* would be the upcoming word. Then, he joined in the turn construction by saying *bowtie* for the addresser. However, for the limitation of the study objectivity (i.e., revealing the resources of joint turn construction) and qualitative approach, Hayashi did not report further about the temporal relation between gesture and speech in such cases. Unfortunately, many other studies that were specifically designed to investigate temporal gesture-speech synchrony also did not clearly investigate the way in which gesture temporally coordinate with speech in terms of different phases of gesture (Butterworth & Beattie, 1978; Hadar & Butterworth, 1997; Morrel-Samuels & Krauss, 1992).

As McNeill (1992) pointed out, gesture can be roughly divided into three phases: preparation, stroke, and retraction. The preparation phase refers to "the limb mov[ing] away from its rest position to a position in gesture space where the stroke begins"; the retraction phase is the "return of the hand to a rest position"; and in-between is the stroke phase, during which "the meaning of the gesture is expressed" (McNeill, 1992: 83). In the excerpt shown in Figure 1, the manual movement starts from the third word *rang* 'let', as the both hands of the speaker face upward and move to the stomach-level from the thigh, preparing for the next phase. In the stroke phase, the meaning of the gesture is expressed. It refers to the verbal referent *jiehe* 'combine', as the two hands move towards each other and then move apart. The speaker repeats this movement twice. Finally, in the retraction phase, the speaker moves both hands back to the thigh at the moment of uttering *gengjin* 'more tightly'. These phases have different functions in communication. The stroke is the most informative stage in meaning expression. Retraction has been proposed to be useful in timing the turn-taking system (Holler et al., 2018). Thus, a fine-grained description of the temporal relationship between phases of gestures and lexical affiliate is the foundation on which we further discuss that whether gesture can predict linguistic information.

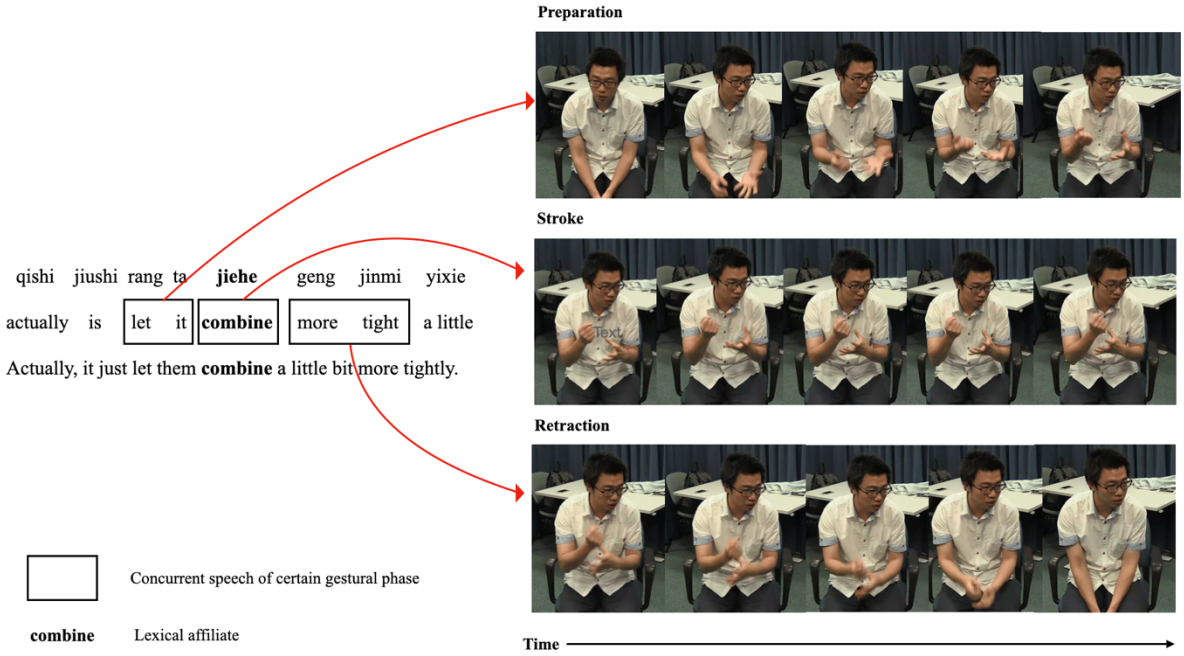

Fig. 1. Illustration of the gestural phases.

As far as we know, there is one relevant prior study that investigated speech and gesture synchrony in natural Chinese conversation. It was found that 60% of iconic gesture strokes are synchronized with the lexical affiliate, 36% preceded it, and 4% followed it (Chui, 2005). However, the author did not provide a clear description of the way in which the gesture stage was coded. More critically, in this the way of identifying lexical affiliate hasn't been clearly illustrated. Thus, it was unclear whether the long-held belief that gestures slightly precede their lexical affiliates means whether the whole stroke completes before the lexical affiliate begins, that the stroke starts first but overlaps with the lexical affiliate, or that the preparation phase initiates before the lexical affiliate.

Until very recently, Ter Bekke et al. (2020) elaborately examined temporal gesture-speech coordination in terms of the timing relation between each gesture stage and the lexical affiliate finding that not only gesture onsets (as a whole including preparation phrase (96%)), but also the stroke phase of the gesture (62%) typically start before their corresponding lexical affiliate. Specifically, strokes start around on average 215 ms before their lexical affiliate. However, it is worth noting that this conclusion was based on all kind of so-called

representative gestures including iconic and deictic gestures and specifically those that occurred in the interrogative utterances in natural Dutch conversation. Interrogative encoding is quite different from declarative encoding. As a consequence, whether ter Bekke and colleagues' finding can be generalized to other verbal expressions (e.g., declarative utterance) and languages is still unknown.

Altogether, this limits our knowledge of to what extent gesture precedes relevant speech segment, and subsequently to what extent gesture comprehension can take place in order to have a predictive potential. Therefore, drawing on those studies, we can hardly know the extent to which people can predict upcoming linguistic input based on gesture information.

The current study

The aims of the present study is first to explore whether and how in spontaneous natural conversations gestures precede speech (corpus study). Based on this foundation, we further investigate the predictive power of gesture in language comprehension (experimental study). We conduct our study in Chinese and thus we first collected a multimodal corpus and analyzed speech and gesture relations in the context of transitive event descriptions and then set up a visual world paradigm using similar sentence-gesture pairs. Our focus is to see whether gestures about transitive actions could predict information about the nominal arguments associated with the verbs.

In the corpus study, we examined first how iconic gestures in natural Chinese conversation temporally coordinate with the corresponding verb and its nominal argument. We tested whether iconic gestures accompanying speech that depicts a transitive event are realized slightly earlier than the verb and/or the nominal argument. Therefore, when gesture holistically depicts an event, which is concurrently described by a verb phrase in speech, people may obtain some information about the nominal argument (i.e., the noun in the verb phrase) before encountering it in speech, such that the gesture could potentially be used to facilitate predictive

language processing. In a multimodal Chinese corpus of unscripted triadic conversations, we annotated iconic hand gestures. For each gesture, we coded which word(s) in the speech corresponded most closely to the concept depicted by the gesture and compared the timing of the word(s) to the timing of the gesture. What is different from previous research identifying " lexical affiliates" here is that given the specific context we allow that the lexical affiliate can be a verb or a whole verb phrase including the nominal arguments of verb (see Kita & Özyürek, 2003, for arguing planning unit for iconic gesture to be verbal clause but not a single word) . The corpus study confirmed the tendency that iconic gestures temporally precede their lexical affiliate in natural Chinese conversation. The follow-up experiment aims to investigate the predictive power of iconic gesture in spoken language comprehension. The experiment uses a typical visual world paradigm containing four experimental conditions. Subjects are presented in preview time with four pictures of objects sufficient for participants to activate semantic and episodic representations corresponding to the objects in the visual display by the time the target linguistic expressions and accompanying gesture input are encountered. We measure the eye movement to the target object in different conditions and determine the fixation proportion to the target as the speech and or gesture input unfolded. For example, the participants view the target object *piano* in this case, in the context of "*I played the piano*" with or without a piano-playing gesture, with three distractors (refrigerator, mattress, trash bin). The speech and gesture pairs are created to form four conditions with which we can identify the predictive power of gesture independent of and contributing to that of speech (see Table 1). In the **neutral** condition, the target displays are each paired with a neutral sentence such that the verb cannot induce any bias to look towards any particular picture. In the **speech-only biasing** condition, the target displays are each paired with a sentence that contains a verb of which selectional restriction can bias eye gaze towards a particular object**.** In the **speech + gesture biasing** condition, the target displays are each paired with not only a sentence that can

bias eye gaze to a particular object but also an iconic gesture associated with the verb phrase in the sentence. The iconic gesture is specifically designed to give away some information about the target object. Finally, in the **gesture-only biasing** condition, the verb in the paired sentence is replaced with "ennn [ən:]". Meanwhile, the iconic gesture remains intact.

A) **Neutral**: the verb can modify both target and distractors e.g., I today *move* the whole afternoon <u>piano</u>
B) **Speech-only biasing**: the verb can modify only the target e.g., I today *play* the whole afternoon <u>piano</u>
C) **Speech + Gesture biasing**: 1) the verb can modify only the target; 2) the gesture is semantically associated with the verb phrase. e.g., I [today *play* the whole afternoon] <u>piano</u> [play the piano]
D) **Gesture-only biasing**: 1) the verb is replaced with "ennn"; 2) the gesture remains intact providing information about the nominal argument. e.g., I [today *ennn* the whole afternoon] <u>piano</u> [play the piano]

Table 1. An example of the four experimental conditions. The verb and its nominal argument are indicated in *italic* and <u>underlined</u>, respectively. A verbal description of the iconic gesture is presented in brackets [ ]. Gestural strokes are time-locked to the onset of the temporal noun and finishing before the onset of the nominal argument (duration is demoted by brackets [ ]).

Experimental hypotheses

We predict that:

- In the **neutral** condition:

  H1. Participants fixates the target objects more than the other three distractors (averaged fixation proportion) by the time that the signifier (i.e., the pronunciation of the noun referring to the target object's name) of the target object is heard.

H0. Participants will not fixate the target objects more than other three distractors (averaged fixation proportion) by the time that the signifier of the target object is heard.

- In the **speech-only biasing** condition:

  H1. Participants will fixate the target picture more relative to the onset of the noun for the target object in the speech-only biasing condition than that in the neutral condition does.

  H0. Participants will not fixate the target picture more relative to the onset of the signifier of the target object in the speech-only biasing condition than that in the neutral condition does.

- In the **speech + gesture biasing** condition

  H1. Participants will fixate more toward the target object up to the moment that the noun for the target is heard in the speech + gesture biasing condition compared to the speech-only biasing condition and the neutral condition.

  H0. Participants will not fixate more toward the target object up to the moment that he signifier of the target is heard in the speech + gesture biasing condition compared to the speech-only biasing condition.

- In the **gesture-only biasing** condition

  H1. Participants will fixate more to the target object in the gesture-only condition than that in the neutral condition relative to the onset of the signifier of the target object. But the target object will attract more fixations in the speech-only and speech + gesture conditions than that in the gesture-only condition ?

  H0. Participants will not fixate more to the target object in the gesture-only condition than that in the neutral condition relative to the onset of the signifier of the target object.

Corpus study: The temporal relations between gesture and speech

Methods

Corpus and Apparatus

Our data contained three triadic, no-task, natural conversations among friends lasting approximately one hour each. They were recorded in the Gesture Lab at the Max Planck Institute for Psycholinguistics (Fig. 2). The participants were Radboud University students with no knowledge of linguistics and who were native Chinese speakers who were living in the Netherlands not more than 3.5 years (M_{year} = 1.19, *SD* = 1.24 , ranging from .08 to 3.5). They were not informed about the particular focus of the study. After filming, they all reported knowing nothing about the research objectives. They were filmed in a full-body shot, with four visible CANON XF205 HD cameras, set to 1280 × 720 50p. Each camera was fitted with a Sennheiser ME-64 to record directional audio. Camera 1 generated a time-code signal, such that everything was synchronized. To ensure that the conversation was as natural as possible, only the middle 40 minutes of each conversation was analyzed. And we randomly picked up one participant from each triadic conversation (2 females and 1 male, M_{age} = 29.0, *SD* = .82).

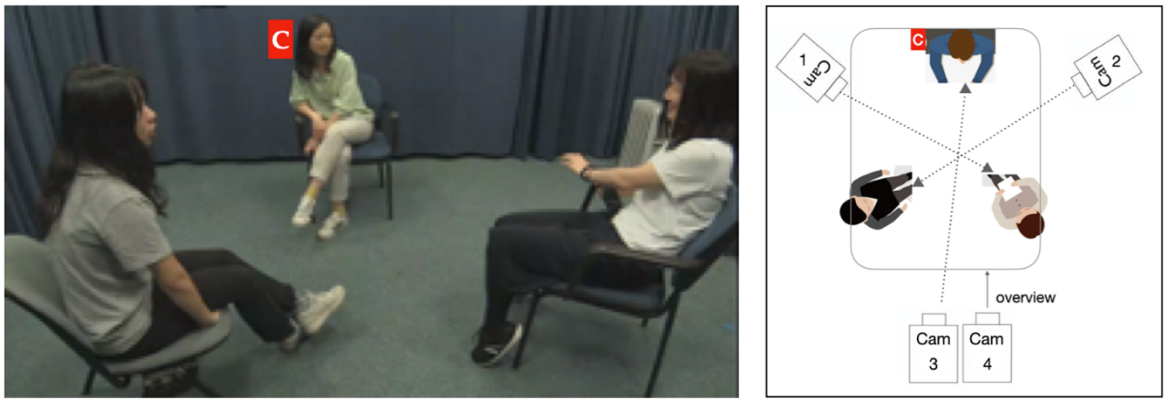

Fig. 2. Illustration of the laboratory set-up of the conversation-filming.

Coding

Study one focused on the timing relations between iconic gestures and their lexical affiliates in the context of transitive-event description. To this end, we first identify the iconic gestures,

followed by their corresponding lexical affiliates, and finally the gesture phases. Gesture annotations and speech-timing segmentation were made in ELAN 5.3 (Lausberg & Sloetjes, 2009) and Adobe Audition CC 12.1.4.5 (Adobe, 2019), respectively.

Iconic gesture coding

In our investigation, only the co-speech iconic gestures were coded. Deictic, metaphoric, beat (McNeill, 1992), and pragmatic gestures, such as "palm-up" (Müller, 2003), "listing" (Tao, 2019), "hand-closing" (Cuffari & Streeck, 2017), and "shrug" (Streeck, 2009b), were excluded from analysis. In addition, gestures produced in unnatural pause (i.e., the obvious unnatural interval within the speech of one speaker, Heldner & Edlund, 2010) were also ruled out given that the temporal relation between gesture and speech may be underpinned by a mechanism that is different from that in fluent speech (Butterworth & Hadar, 1989). Apart from these, self-adaptors, such as scratching the leg or smoothing the hair, were excluded because of the absence of the semantic relation with the speech.

The gestures were coded twice. The first time, iconic gestures were identified based on their form while the audio was muted. The second time, these gestures were checked to see if they were iconic gestures and what they meant based on the audio. In the final analysis only those that were meaningful in speech context were included. As a result, from the 115-minute speech of the 3 participants, we obtained 225 iconic gestures.

Coding lexical affiliates

Regarding sentences that cooccurred with the iconic gestures, we first selected the sentences in which the syntactic unit was realized in the complete or subject-(elliptical subject)-verb-object structure (i.e., SVO or VO structure), and where the nominal object was the patient argument of the verb. We only analyzed the gestures that depicted transitive actions that could have associated nouns and occurred within an SVO/VO-structure clause. We also conditioned that at least part of the gesture should overlap with part of/the whole clause. Because if words occur

a few sentences away from the relevant gesture, they are not considered as lexical affiliates with that gesture (Munhall et al., 1996). The clause was considered as the context in which we could understand the meaning of the gesture (Kita and Özyürek, 2003).

Then, we identified which part of speech was most semantically-close to a gesture in meaning (see Schegloff, 1984; ter Bekke et al., 2020, for a similar method). Gesture is considered an alternative channel, in addition to speech, of packaging human conception for production (Alibali et al., 2000; Hostetter et al., 2007; Kita, 2000). Hence, a gesture should be qualified to refer to more-than-one-word referents, and mapping it should be possible with two-or-more-word speech. Therefore, even if most of the studies have simply focused on one-word lexical affiliates (e.g., Chui, 2005), we did not limit our gesture-speech mapping to single word.

We first interpreted the meaning of the gesture based on the gesture features, especially its shape. Then, we rechecked the interpretation based on the sentential context in which the gesture was produced. Since gesture conveyed conception in a holistic way (McNeill, 2005), it was hardly possible to find a clear and clean corresponding relation between gesture and speech across all cases. Therefore, to keep the gesture–speech association as consistent and systematic as possible, we made the interpretation parsimoniously. We dealt with the action-description gesture in the following ways: 1) basically, we identified the corresponding action verb as the lexical affiliate of the gesture (ter Bekke et al., 2020). 2) If the semantic-related part of speech was a one-character verb, and the patient argument of the part of speech was also a one-character noun that realized right after it, we considered the whole verb phrase as the lexical affiliate (i.e., one-character verb + one-character noun). Because modern Chinese has been experiencing a bi-syllablization trend (Dong, 2011). The verb phrase which is realized by a one-character verb and one-character noun is subject to be considered as a verb rather than a verb phrase in daily use, even though this kind of “usage” still be categorized as verb phrase grammatically. This tendency is considered to be probably reshape the inner lexical knowledge

of Chinese speakers (Tao, 2003). That is, this type of verb phrase can probably processed as a unified word gradually. Thus, in our coding, for instance, when dealing with *kai (drive) che (car)*, we did not further distinguish whether the gesture specifically referred to drive or car, but identified *kai che (drive car)* as a whole. 3) In contrast, if the one-character verb and its patient argument was separated by no less than one syntactic unit (e.g., adjective, directional verb, measure word or auxiliary word, etc.), then we treated the verb itself as the lexical affiliate. 4) If the one-character verb was adjacently followed by a pronoun that served as the verb's patient argument, we only chose the verb as the lexical affiliate. Apart from these, when possible we excluded the affiliated adverbial and complementary elements from lexical affiliate selection.

This process yielded 37 cases (out of 225 iconic gesture cases). There are 32 lexical affiliates only contain a verb (e.g., *he* 'drink', *sha* 'kill'). And the rest 5 lexical affiliates are realized as a verb phrase (i.e., one-character verb + one-character noun; e.g., *xi* 'wash' *tou* 'hair', *pa* 'climb' *shan* 'mountain'). The total number of the valid case is not high. On the one hand, 29 iconic gestures that realized in one of the five following conditions were marked as invalid gestures: 1) lexical affiliate was uttered in English (7 cases); 2) concurrent speech was dysfluent (12 cases); 3) concurrent speech was hard to be recognized (1 case); 4) lexical affiliate was hard to be identified (1 case); and 5) no speech cooccurred with the gesture (8 cases). But the main reason was that it was required that the valid case to have a nominal patient argument. However, in natural Chinese conversation, interlocutors prefer to put more efforts on elaborating the results that the action leads to and the way in which the action is performed. As a consequence, speakers tend to omit the patient argument and add complementary and adverbial to the verb when depicting the action (Tao & Hu, 2019; see Thompson & Hopper, 2001, for a similar discussion based on English data). That is, most of the speech that

cooccurred with the iconic gesture did not have the nominal argument. As a result, gestures that cooccurred with such kind of speech were also considered as the invalid.

Gesture phases coding

For gesture phase coding, the gestures were first segmented into dynamic and static gesture phases using the frame-by-frame method described in Seyfeddinipur (2006). Next, the segmented phases were identified as preparation, stroke, or retraction. Sometimes, after arriving at the proper position, speakers will hold their hands for a while before initiating movement. That is the pre-stoke hold. As such, after the movement, the hand sometimes will be held for a while. That is the post-stoke hold. So, we also segmented pre/post-stroke hold. Only the stroke part was the mandatary constitute of a gesture. That is, some gestures in our coding did not have the other phases apart from stroke.

Overall, the first frame of a gesture was typically the first blurry frame of the preparation. The last frame of a gesture was the first frame in which the hands were still in their rest position. For identifing and distinguishing the boundary between stroke and the rest part of a gesture or between two successive strokes, we adopted four features: *Handedness* refers to which hand and how many hands are used to acknowledge the referent. *Orientation* refers to which direction the palm faces. *Motion* refers to hand movement and includes two aspects: *motion type* (e.g., circling or straight, still or rotating, or curved or straight-line tracing) and movement *direction* (e.g., inward or outward and upward or sideward). The last feature is *hand shape.* When one of the parameters changed, a new gesture started. When it was difficult to assess the boundary between two successive gestures, we analyzed the gesture pixel by pixel (1 msec./PX) based on the four parameters.

Reliability check

An independent coder, who was blind to the study objectives, identified gestural phases and the gesture–speech mapping that fulfilled the aforementioned criteria. Reliability was established

for 35% of the data (n = 86), which yielded a reliability of 72% and 95.3% for gesture identifiability and gesture–speech mapping identification, respectively, indicating a high degree of agreement.

Analysis

First, we asked whether gesture onsets and gesture strokes preceded lexical affiliate onset. We calculated the temporal difference between stroke onset time and lexical affiliate onset time for each gesture-affiliate pair. Meanwhile, the difference was calculated also between preparation onset time and lexical affiliate onset time.

Next, we asked whether gestural stroke was subject to be completed before the nominal argument. The difference was calculated between stroke offset and nominal argument onset. Also, the difference was calculated between retraction offset time and lexical affiliate onset time

We fitted linear mixed effects models using the lme4 package (version 1.1-21; Bates et al., 2015) in R (version 3.6.0; R Code Team, 2019), with *p*-values calculated using the package lmerTest (version 3.1-1; Kuznetsova et al., 2017).

Results

In general, the overwhelming majority of gestures (95%) started before their lexical affiliate, around 488 ms on average (Fig. 3). An intercept-only model with random intercept for triad for idiosyncratic variation that was due to individual and conversational context differences revealed that overall, gesture onsets significantly preceded lexical affiliate onsets (β = -487.54, *SE* = 55.45, t = -8.79, $p < .001$). The majority of strokes (81%) was realized earlier than their lexical affiliate, around 172 ms on average (Fig. 3). An intercept-only model with random intercept for triad revealed that stroke onset significantly precedes lexical affiliate onset (β = -172.00, *SE* = 57.55, t = -2.99, p = .005).

There were 5 cases that were different from the majority. Because the lexical affiliate of each of the 5 cases included not only a verb but also the nominal argument of the verb. When we exclude those 5 cases from analysis, the result did not show a significant change. The gesture onsets still significantly preceded lexical affiliate onsets (β = -484.66, SE = 58.68, t = -8.26, p < .001). And the stroke onset also significantly preceded lexical affiliate onset (β = -171.13, SE = 62.95, t = -2.71, p = .011).

Thus, not only gesture onsets as a whole, but also gesture strokes typically started before their corresponding information in speech (Fig. 3).

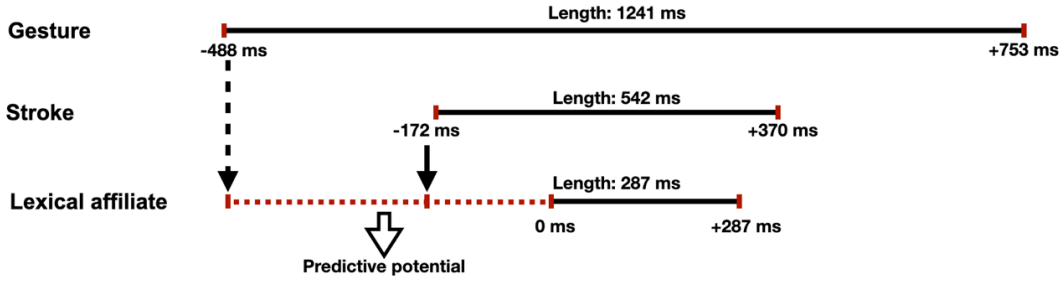

Fig. 3. Mean temporal relations between iconic gestures, their strokes and their lexical affiliates (FYI: Based on all 37 cases in corpus: the lexical affiliates of 32 cases were the verbs only; and the lexical affiliates of 5 cases were the verb phrases including a verb and a noun).

For the temporal relations between gestural offset and nominal argument onset, we first excluded the 5 cases from our analysis in which the nominal arguments were part of the lexical affiliates. An intercept-only model with random intercept for triad revealed that stroke offset was not significantly precedes nominal argument onset (β = -232.00, SE = 168.66, t = -1.38, p = .18). If we included all cases into account, the result did not show a significant change (β = -182.84, SE = 147.28, t = -1.24, p = .22). Within our coding framework, unlike the temporal relation between stroke onset and lexical affiliate onset, the timing relation between stroke offset and patient argument onset could be influenced by many predictable and unseen factors

which were out of the investigation scope of our present study. Such as, the information status of gesture (i.e., complementary or redundant gesture, Bergmann et al., 2011), the state of consciousness of gesture (i.e., foreground or background gesture, Cooperrider, 2017), and the pragmatic functions of gesture (e.g., expressing disagreement or making clarification, see Chui, 2014 for details), etc. However, our current corpus-based analysis showed the majority of strokes (73%) ended before the nominal argument onset, around 192 ms on average (Fig. 4).

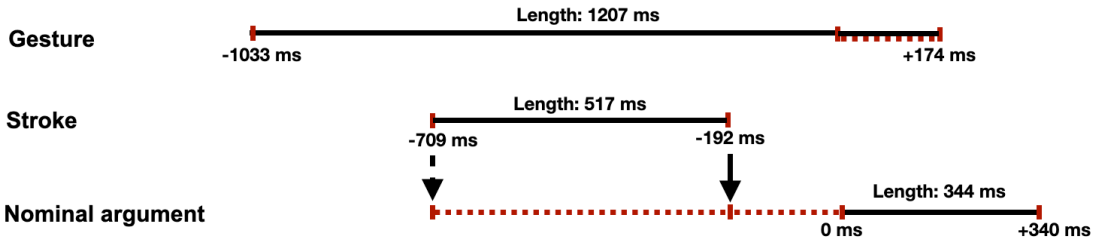

Fig. 4. Mean temporal relations between iconic gestures, their strokes and the nominal argument (FYI: Based on all 37 cases in corpus: the lexical affiliates of 32 cases were the verbs; the lexical affiliates of 5 cases were the verb phrases including a one-character verb and a one-character noun).

Interim summary

When people employ both gesture and speech to describe a transitive event in face-to-face natural conversation, gestures as a whole as well as their stroke parts, start before the corresponding semantically-related part in speech. Our results are basically converging with previous works in Dutch (Ter Bekke et al., 2020). Altogether with abovementioned studies, it is convincing that iconic gestures fulfil the two prerequisites for language prediction based on gestures to be possible: 1) interlocutors are able to grasp the shared semantic information of gesture and speech during language comprehension, and 2) gestures precede their shared semantic information in speech. Thus, co-speech iconic gestures indeed appear to legitimately have predictive potential that interlocutors can exploit to predict the upcoming linguistic input.

Experimental Study: Visual world eye-tracking

Study 1 shows that gesture tends to be realized earlier than its lexical affiliate. It paves the foundation on which we can further explore that to what extent can iconic gesture predict the upcoming nominal word independently from the predictive power of the linguistic input?

Methods

Participants

180 university students will take part in the main eye-tracking study. They are native speakers of Chinese. They report any history of learning or reading disabilities or neurological or psychiatric disorders. They have either normal or corrected-to-normal vision.

Materials

The materials of the eye-tracking study are 60 visual displays comprising 48 target displays and 12 fillers. Every display contains one gesture and four digital photos of one object each. The photos are isometrically located around the centre. In the centre of the display, there is a video interface, the same size as the photos, showing an actor uttering a verb phrase in Mandarin Chinese (e.g. tan 'play' gangqin 'the piano') along with a gesture. The target-displays are devised each consisting of two accompanying sentences, and each filler is created as having one sentence accompanied only (see Fig. 5).

For each target display, one of the two corresponding sentences contains a verb of which selectional restriction allows only a single object in the visual display to be the semantically associated object of the verb; whereas the other sentence contains a verb which permits all of the visual objects, including the target object, to be referred to postverbally. For instance, for the target display shown in Figure 5, two sentences are recorded: *Wo jintian **tan** le yi xiawu de gangqin* '(lit.) *I today **play** the whole afternoon piano*' and '*Wo jintian **ban** le yi xiawu de*

gangqin '*I today* ***move*** *the whole afternoon piano'*. The four objects are a refrigerator (bingxiang), a piano (gangqin), a mattress (chuangdian), and a trash bin (lajitong). Among them, only the piano (gangqin) can be played (tan) in principle. However, all the objects can be semantically modified by 'move (ban)'. In the video, the actor performs a strumming-type gesture semantically associated with 'play the piano'. Given that the actor can probably be imagined as the initiator of the action, we decide to use the first person 'I' as the agent (cf. Milburn et al., 2016). To allow sufficient time for the viewer's eyes to reflect language processing, we separate the verb and its nominal argument by a general measure phrase, which in Mandarin Chinese can be used to denote an instance of an event or to indicate the volume, weight or length of an object, etc. (Li & Thompson, 1981). The general measure phrase does not give away the semantic information of the nominal argument. In the above example, the general measure phrase is 'the whole afternoon'. This phrase can indicate the temporal duration of an event without telling what the event is. Unlike the target display, the filler display has only one corresponding sentence with a verb of which selectional restrictions allow every object in the scene to be the possible referent.

Fig. 5. Example scene used in the eye-tacking experiment. Participants hear 'Wo jintian *tan* le yi xiawu de gangqin [Literal English translation: I today *play* the whole afternoon piano]' or 'Wo jintian *ban* le yi xiawu de gangqin [Literal English translation: I today *move* the whole afternoon piano]' whilst viewing this scene. The actor performs a piano-playing gesture which is semantically associated with *tan (play) gangqin (piano)* in the former sentence. When hearing the latter sentence, the actor makes no movements but stands with both hands down naturally.

Gestural display: Iconicity ratings and gesture selection

In order to prepare appropriate stimuli of our main eye-tracking study in terms of gesture informativeness, we took an iconicity rating test (Ortega et al., 2017) to determine whether the iconic gestures we used were informative about their meanings of their associations with the objects in the visual world paradigm even without a speech context. We recorded another set of action gestures to be coupled with speech. To ensure that the iconic gestures to be used in the main experiment intelligibly depict the specified transitive events, we conducted a pre-test examining whether the gestures made by the actor in the video transparently depicted the verb–noun pairs (VPs) we associated them with in our audio files. Twenty native Chinese speakers (11 females and 9 males, M_{age} = 23.2, SD = 4.0) with no motor, visual, auditory or language impairments, and who eventually would not participate in the main experiment, participated in the test. They were students of Tilburg University, had no knowledge of linguistics and psychology and had spent no more than two years living outside of the mainland of China (M_{year} =.97, SD =.58).

The test participants were presented with 110 mute video stimuli ($M_{video\text{-}duration}$ = 2914 ms, SD = 455 ms) that contained a mouth-mosaicked actor performing a gesture. The 110 stimuli contained 75 VP types. Fifteen of those types were designed to function as the potential fillers

in the main experiment in which the gesture had no transparent semantic connections with the given VPs and thus were illegible to depict the particular transitive actions. Each potential filler type contained only one token. The remaining 60 types were designed as the potential target VPs. That is, only the gestures who could get the rating score higher than 4.0 (1-7 scale) would be finally selected as the genuine target gestural stimuli. The number of tokens of each target type varied from 1 to 5. Some of the VPs were easily depicted by various gestures from various aspects; in contrast, the others could hardly be depicted by several gestures from multiple perspectives. For example, there was only one gesture of riding motorcycle but four gestures of fishing. Tokens within one type varied from each on shape or (and) motion or (and) handiness, etc. We finally got 110 tokens. All stimuli were presented on a computer screen by using PsychoPy3 (Peirce et al., 2019) in a different, randomised order for each participant. Moreover, tokens of the same type did not adjacently occur.

The video stimuli were filmed with an upper-half-body shot using a visible CANON XF 205 HD camera set to 1280 x 720 50p and edited and analysed in Final Cut Pro X (Apple, 2019) and ELAN 5.3 (Lausberg & Sloetjes, 2009), respectively. To ensure the gesture was performed as naturally as possible, the actor was asked to utter the pre-designed semantically relevant verb phrase when making the gesture. For the fillers, the actor either randomly moved his hand(s) or made a superimposed beat that was semantically irrelevant to the verb phrase that he simultaneously uttered. Even though these gestures did not have semantic associations with the spoken phrases, they more or less possessed functional meanings. For example, a superimposed beat tended to be realised concurrently with the prosodic prominence for highlighting the gist of the speech (Leonard & Cummins, 2011). The random hand movements used as the fillers in our study were usually produced by speakers who had difficulty with verbalisation in daily conversation (Chui, 2014). That is, these gestures neither added

supplementary or redundant semantic information to the speech nor provided information that was semantically contradicted by the speech.

In the first section, participants were presented with a fixation cross for 1000 ms, after which the video stimulus was played. After the video onset, participants were asked to type one verb phrase (one verb + one noun) in Chinese (e.g. tan 'play' + gangqin 'the piano') that they associated with the movements in the video. They were allowed to answer by saying 'I do not know' if they could not understand the meaning conveyed by the video. After finishing the 110 stimuli, they were given a mandatory 10-minute break before starting the second section. In the second section, they were again exposed to the 110 stimuli but in a sequence different from that in the first section. We displayed a fixation cross for 1000 ms to the participants after which the video stimulus occurred followed by the noun we had originally matched to the gesture. Furthermore, we asked the participants to indicate on a scale from 1 (apparently non-transparent) to 7 (apparently transparent) the extent to which the gesture transparently depicted the certain action with the object embedded in and referred by the noun that was presented on the screen (see Fig.6).

All participants were expected to complete the task in approximately 50 minutes. After the experiment, no participant reported knowing the genuine purpose of the test. We first assessed the rating score in the second section. For the 60 target types, the 12 types that did not score more than 4 points on the 7-point scale were discarded. Among the remaining 48 types, some contained two or more tokens. The token with the highest score of each type was selected. If the scores of several tokens were the same, the token with the minimum standard deviation was selected. The mean score of iconicity over the 48 finally-selected videos was 5.58 (*SD* =.90) and ranged from 4.1 to 7.0. The typed answers to the question in the first section ('Which verb phrase do you associate with this manual movement?') were used to determine which VP had to be modified to a possibly more frequently occurring synonymous VP, or which gesture

should be re-associated with a completely different new VP. The modification was aiming to find the best-fit VP to match to the gesture. Thus, we do not think our modification would decrease the rating scores. We coded the answers as either 'intended' when the same or synonymous verb phrase was given or 'unintended' when the input was a completely unrelated verb phrase or a verb phrase constituted by the same and/or synonymous verb and a semantically unrelated noun. The results revealed a mean recognition rate of 47% for all the gesture videos. This intelligibility of the gestures seemed low; however, the result was unsurprising. Most of the gestures were not pantomimes, which are usually produced without speech and by simulating genuine behaviours (Otegar & Özyürek, 2020). The majority of gestures in our study were designed to depict a partial image of a transitive event. Additionally, interlocutors commonly encounter the ambiguity of a gesture when it is unaccompanied by speech in daily conversations (Krauss et al., 1991; Habets et al., 2011). By contrast, all gestures will be presented along with speech and pictures in the main experiment. Besides, during a small chat after the test, all the participants indicated that when they saw the noun in the second section they often found that that noun fit the gesture in the video as well, despite that it was not in line with their answer. This indicates that the mean recognition rate may be negatively biased which is reflected in the rating score: although participants may have answered a different VP in the first task, yet they highly scored on the transparency of the videos. Therefore, we conclude that the negative bias will not jeopardise the answers to our research questions and hypotheses.

Twelve out of fifteen fillers were selected according to the ascended rating score. The mean score on iconicity over the 12 fillers was 1.33 (SD = .16) ranging from 1.05 to 1.65. The typed answers revealed a mean recognition rate of 0% over the 12 fillers. That is, unlike the target gestures, the so-called "filler gesture" had no transparent association with the speech on semantic level. An independent one-tail t-test ensured that the fillers and targets were

sufficiently distinguishable that could dutifully implement their own functions in the main experiment ($t(55.73) = 30.75$, $p < .001$, $r = .97$).

The finally-selected 60 gestures ($M_{\text{target-stroke}} = 1094$ ms; $SD = 407$ ms) will be used in the eye-tracking study. We mute the actor's voice and play audio from another speaker in the eye-tracking study. Since the actor's mouth is blocked out by a grey mosaic, the problem of audio–video synchrony can be eliminated. The mosaic also preventes participants from receiving phonological cues about critical words from the actor's lip movements (Ross et al., 2007; Sumby & Pollock, 1954). When the speech signal is clear, blurring lip movement (or not) does not influence language comprehension (Drijvers & Özyürek, 2017). Altogether, we do not think the block-mouth will bias our findings. Operationally, we asked the actor to wear a dark green shirt that matched well with the dark blue background and to allow his forearms to be visible so that viewers could easily discern his gestures.

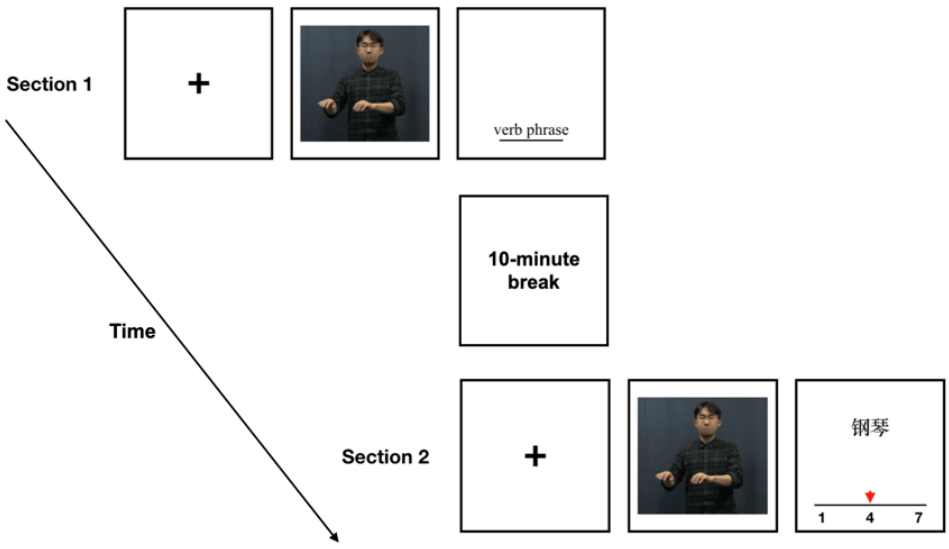

Fig. 6. Illustration Procedure of the iconicity rating test (e.g., comprehension of the strumming gesture; rating the degree of association between the strumming gesture and the noun, "piano 钢琴", in terms of the meaning-transparency).

Pictorial display

To create the pictorial scenes in the visual displays, the photographs of objects are drawn from the Bank of Standardized Stimuli (BOSS, Brodeur, Guérard, & Bouras, 2014; Brodeur et al., 2010) and the stimulus set developed by de Groot et al. (2016*),* which contain words and photographs of common objects matched for visual and semantic similarity. As Huettig and McQueen (2007) pointed out, in the visual world paradigm, eye movement considered as a reflection of the course of online language comprehension is guided by the phonological, semantic and shape information of objects. Therefore, in each visual display, the four objects differ from each other in terms of both their initial sounds and sematic categories (for a detailed discussion of semantic categories, see Frank et al., in prep). The analyses of the frequency of the verbs and objects are carried out by using the SUBTLEX-CH database (Cai & Brysbaert, 2010), which is developed based on film subtitles, believed to maximally represent language-use in all genres (Hu & Tao, 2017; Tao & Liu, 2010 a, b). Raw frequencies are transformed to Zipf values, as suggested by Van Heuven et al. (2014). In the constrained sentences, the mean Zipf-transformed frequency of the verbs is 4.47 (*SD* = .80). In the neutral sentences, the mean Zipf-transformed frequency of the verbs is 4.82 (*SD* = .72). The fact that the constrained verb is less frequent than the neutral verb is probably attributable to the constrained verbs' more specific selectional restrictions (Hintz et al., 2017). As we predicted facilitation effects for constrained rather than neutral items, this difference does not weaken our conclusions. The objects used in the 60 displays are sorted into four sets: one target-set (*M* = 3.77; *SD* = .90) and three distractor-sets ($M_{\text{distractor-one}}$ = 3.55, *SD* = .75; $M_{\text{distractor-two}}$ = 3.55, *SD* = .75; $M_{\text{distractor-three}}$ = 3.53, *SD* = .77*).* The averaged frequency of each set have no statistical difference as determined by one-way ANOVA ($F(3, 188) = .96$, $p = .412$).

Sentential stimuli

The sentences are spoken with neutral intonation at a normal pace by a young male native speaker of Mandarin Chinese. Recordings are made in a sound-damped booth, sampling at 44

kHz (mono, 16 bit sampling resolution) and stored directly on computer. The mean sentence duration is 3900ms (*SD* = 224 ms). Onsets and offsets of all words are marked using Audition CC (V 12.1.4, Adobe Systems, 2019).

Design

There are four experimental conditions as mentioned before. In the **neutral (or the baseline)** condition, the scene is paired with a neutral sentence such that the verb does not induce bias to look at any particular picture, for example, *Wo jintian* ***ban(move)*** *le yi xiawu de gangqin (piano)*. (lit.) *I today* ***move*** *the whole afternoon piano*. Every object within this display can be moved; consequently, none of the objects are assumed to be able to predominantly attract the eye gaze before the word piano is heard.

In the **speech-only biasing** condition, participants listen to this sentence: *Wo jintian* ***tan (play)*** *le yi xiawu de gangqin (piano)*. (lit.) *I today* ***play*** *the whole afternoon piano*. The selectional restriction of the verb *play* in Chinese (different than in English) particularly requires the patient argument to be a set of musical instruments which are basically manipulated by fingers, such as piano in our stimulus. Due to the limitation of the situated visual scene, participants are expected to interpret the linguistic input within the context of the visual display (Vulchanova et al., 2019). Other studies have observed that the (pre-)activated semantic information guides the eye movement (Allopenna et al., 1998; Dahan et al., 2001; Huettig & Altmann, 2005; Yee & Sedivy, 2006), and therefore, the picture of piano probably attracts more participants' eye gaze relative to the onset of piano than the neutral condition does.

In the **speech + gesture biasing** condition, the display is paired with the same sentence as in neutral speech sentence and with an iconic gesture overlapping with the verb phrase *play the piano* in the sentence. Because in addition to the selectional restriction provided by speech, gestures can represent semantic information relating to an underlying conception, which is

either contained (i.e., complementary gesture) or not (i.e., redundant gesture) in the accompanying speech (Cooperrider, 2017; Kita et al., 2017) and helps to disambiguate verbal information (Drijvers et al., 2019). The piano-playing gesture as well as the verb play can activate both the episodic and the semantic knowledge of pianos, as a result, participants will be more confident that the upcoming noun will be piano. Apart from this, since gesture tends to be realized earlier than the semantic-relearnt part of speech, we predict that there will be more looks toward the piano even before the word *piano* is heard compared to the speech-only biasing condition.

Finally, in the **gesture-only biasing** condition, the verb in the paired sentence is replaced with *ennn* [ən:], a meaningless syllable often produced by speakers who suddenly cannot retrieve a word. The iconic gesture remains the same. We predict that the fixation proportion to the piano in this condition will be higher than that in the neutral condition but not as high as that in the speech-only biasing condition because even if gesture is believed to originate from a common conceptual level, as speech does (Kita & Özyürek, 2003; de Ruiter, 2000; see Özyürek & Woll, 2019 for discussion), it cannot be fully interpreted independently from speech (Krauss et al. 1991; Habets et al., 2011). But comparing with *move* in the neutral condition, the iconic gesture can anyway convey part of the conceptual aspect of the follow-up nominal argument. In this condition, the gesture functions more like a so-called "silent gesture" which is produced without the semantically-relevant part of speech. As Ortega and Özyürek (2020) pointed out that when such kind of iconic gestures were designed to convey the conception of transitive event, they are highly intelligible.

It is a within-subject design. Participants are evenly divided into four groups. Each participant will be presented with 48 target trials together with 12 filler items. On each trial, participants are exposed to four objects and audio-video input. 60% of the 60 trials includes a gesture-video in which the actor makes iconic gesture or gesture that only has functional

meaning. There is no obvious connection between gesture-availability and the content of the speech. On the trials without gesture, the actor is still occurring with arms down naturally. Hence across trials participants may not build up an expectation that which trial will have a gesture-video.

Materials are counterbalanced across the experimental trials for four groups of participants. Each participant receives 12 trials in the neutral, speech-only biasing, speech + gesture biasing and gesture-only biasing conditions. The same 12 fillers are used for all of the four groups. Trials are presented in the same random order to each participant.

Procedures

The participants are tested individually in a sound-shielded booth. Eye movements are recorded by using an EyeLink 1000 tracker sampling at 1,000 Hz. Participants placed their heads in a chinrest approximately 75 cm from the computer screen. The experimental stimuli are displayed on a 23-inch computer screen. Participants are instructed to listen to the speech carefully; additionally, they are allowed to look at whatever they wanted, but during the experiment, they are supposed to look only at the screen. That is, their task is to look and listen (Altmann & Kamide, 1999; Huettig et al., 2011, for discussion). Meanwhile, They are allowed to blink only between each trail. After calibration, the participants are randomly assigned to one group. The speech is presented through headphones.

A trial starts with the presentation of a central fixation dot for 1500 ms. The dot disappears, and the playback of the sentence starts. The onset of the display is timed to 2000 ms before the occurrence of the verb in the speech signal. The gesture preparation starts 200 milliseconds after speech onset. The gestural stroke starts on average 777 milliseconds before verb onset and ends 300 milliseconds postverbally. The duration of the whole gesture is on average 600 milliseconds before the onset of the target noun. The time between the onset of

the verb and the onset of the target noun is on average 2000 milliseconds (see Fig.7 for the trial structure). The four objects and the actors remain in view for the remainder of the trial. The positions of the pictures are randomized across four fixed positions of a (virtual) 2 x 2 grid. All objects are the same distance from the center, with a direct visual angle of approximately 12°. The positions of the four objects are randomized. The colour of the background is set as 94-94-94 (GRB). Each participant is presented with all 60 trials of one list. The order of trials is randomized automatically before the experiment. The duration of the eye-tracking experiment, including the background investigation and calibration, is approximately 20 minutes. Regions of interests (250 x 250 pixels) are defined around each object. The data from a participant's left or right eye (depending on the quality of the calibration) is analyzed in terms of fixations, saccades, and blinks, using the algorithm provided in the EyeLink 1000 software. Fixations are coded as directed to the target, to one of the three unrelated distractors, or elsewhere.

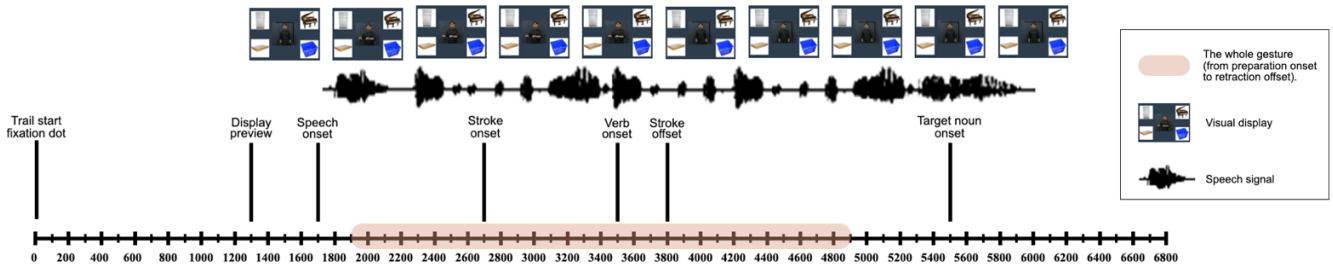

Fig. 7. Timeline of event in the trail of the eye-tracking study.

Sampling plan

Sample size

Based on the recommended sample size per condition provided by Lakens and Evers (2014, p. 280), to achieve 80% statistical power to observe the effect with an alpha of .05, for an estimated effect size (r = .3), we aim to recruit 180 participants (45 for each condition). As indicated by Lakens and Evers (2014, p. 280), if we consider the point of stability for the

correlation magnitudes, after which the sample estimates do not deviate from a "corridor of stability" around the true population value, the corresponding effect size of 45/condition falls between .3 to .4, based on a wider corridor of $w = .2$ with an 80% stability confidence.

We noticed that based on the Monte Carlo simulations of correlational analyses, Schnbrodt and Perugini (2013) provided a general recommendation of n = 250 per condition when examining effects of $r = .21$ (the average effect size in psychology based on Richard et al., 2003) if researchers want to reach a small ($w = .1$) corridor of stability. However, we think this does not diminish the contribution of our study, but future replication is required.

We recruit a community sample through advertising (i.e., posters, flyers, WeChat) in Nijmegen, Tilburg, Utrecht and Wageningen, the Netherlands. Participants are paid €6 for participating in the study (one eye-tracking study of 20 minutes conducted in an experiment room at the MPI for Psycholinguistics, Nijmegen, with the participants' written consent to use their data). Data collection will be terminated when 180 participants complete the testing.

Data exclusion

Remove the trails with blinks.

Missing data

Individuals with missing data will be kept in the analysis.

Analysis Plan

Fixations proportion over time (from 500 ms before the acoustic onset of the verb to 500 ms after the acoustic onset of the target word (time zero)) to target and to the averaged distractor objects in the four conditions is calculated. For making the comparison clearly, we first plot the fixation proportions to the target object and to the averaged distractor objects for the gesture-only biasing and the neutral conditions. We compute by-participant confidence intervals (95%)

for each line at every sampling step (2 ms). The area between the lower and the upper bounds is shaded. We log-transforme the fixation proportions and subtract fixations to the three distractor objects from fixations to the target objects in the gesture-only biasing and the neutral conditions (cf. Arai, van Gompel, & Scheepers, 2007). A difference of zero means that target and averaged distractors are fixated equally often, and a difference greater than zero means that more fixations were made to the target object. By-participant confidence intervals is calculated for each sampling step, based on the mean of the difference between target and distractors. We also present the variability in anticipating the target object in the gesture-only biasing condition and in the neutral condition. For doing so, we calculate each participant's mean difference between looks to the target and looks to the averaged distractors during the critical time window in the respective conditions. Standard deviation error bars indicate within-participant variation. A mean difference of zero indicates equal looks to target and distractors; a positive mean difference implies a bias for the target. The same way is applied to plot the comparison between the fixation proportions to the target object and to the averaged distractor objects for the gesture+speech biasing condition and gesture-only biasing condition; the speech-only biasing and the gesture+speech biasing conditions; the speech-only biasing and the neutral conditions.

To calculate the dependent variable, we divide each participant's proportion of looks to the target during the onset-verb-onset-target period (FYI: 200 ms is added to both verb and target onset to adjust for the time it takes to program and launch a saccadic eye movement) on a given trial by that participant's proportion of looks to the averaged distractors during the same time window. The resulting values will be log-transformed. Prior to the division and log-transformation fixation proportions of 0 or 1 will be replaced with 0.01 and 0.99, respectively (cf. Macmillan & Creelman, 1991). The data is aggregated by participant and by item yielding average scores for each participant and for each item.

We use R (R Core Team, 2019) and lme4 (Bates et al.,2015) to perform a linear mixed effects analysis of the relationship between fixation proportion and gesture-availability and predictability (high constrained vs. low constrained). The dependent variable (log-transformed fixation ratios) was calculated as described above. As fixed effects, we entered gestural-availability (available vs unavailable) as well as the speech condition (high constrained vs. low constrained) into the model. Participants and Items were included as random factors, each with random intercepts and slopes. Interactions between condition were added. The nonpredictable condition was put on the intercept. *P*-values were obtained by using the package lmerTest (version 3.1-1; Kuznetsova et al., 2017)

Blinding

Data collection and analysis will not be performed blind to the conditions of the experiments.

Pilot data

There are no pilot data.

References

Alibali, M. W., Yeo, A., Hostetter, A., & Kita, S. (2017). Representational gestures help speakers package information for speaking. In R. B. Church, M. W. Alibali, & S. D. Kelly (Eds.), Why gesture? How the Hands Function in Speaking, Thinking and Communicating (pp. 15-37). John Benjamins Publishing Company.

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*, 419-439. https://psycnet.apa.org/doi/10.1006/jmla.1997.2558

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition*, *73*, 247-264. https://doi.org/10.1016/S0010-0277(99)00059-1

Altmann, G. T. M., & Mirković, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science 33*, 583-609. https://dx.doi.org/10.1111%2Fj.1551-6709.2009.01022.x

Apple. (2019). Final Cut Pro X (Version 10.4.8). https://www.apple.com/nl/final-cut-pro/

Arai, M., van Gompel, R. P., & Scheepers, C. (2007). Priming ditransitive structures in comprehension. *Cognitive Psychology*, *54*, 218-250. https://doi.org/10.1016/j.cogpsych.2006.07.001

Adobe. (2019). Adobe Audition CC (Version 12.1.4.5). https://www.adobe.com/be_en/

Backus, A. M. (2014). Towards a usage-based account of language change: Implications of contact linguistic theory. In R. Nicolai (Ed.), Questioning Language Contact: Limits of Contact, Contact at its Limits (pp. 91-118). Brill.

Bar, M. (2003). A cortical mechanism for triggering top-down facilitation in visual object recognition. *Journal of Cognitive Neuroscience*, *15*, 600-609. https://doi.org/10.1162/089892903321662976

Bates, D.,Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 48. http://dx.doi.org/10.18637/jss.v067.i01

Beattie, G., & Shovelton, H. (1999). Do iconic hand gestures really contribute anything to the semantic information conveyed by speech? An experimental investigation. *Semiotica*, *123*, 1-30. https://doi.org/10.1515/semi.1999.123.1-2.1

Bergmann, K. Aksu, V., & Kopp, S. (2011). The relation of speech and gestures: Temporal synchrony follows semantic synchrony. *Proceedings of the 2nd Workshop Gesture and Speech in Interaction*, Bielefeld, Germany.

Brodeur, M.B., Dionne-Dostie, E., Montreuil, T., & Lepage, M. (2010). The Bank of standardized stimuli (BOSS), a new set of 480 normative photos of objects to be used as visual stimuli in cognitive research. *PLoS ONE*, *5*, e10773. https://doi.org/10.1371/journal.pone.0010773

Brodeur, M.B., Guérard, K., & Bouras, M. (2014). Bank of standardized stimuli (BOSS) phase II: 930 new normative photos. *PLoS ONE*, *9*, e106953. https://dx.doi.org/10.1371%2Fjournal.pone.0106953

Butterworth, B., & Beattie, G. W. (1978). Gesture and silence as indicators of planning in speech. In R. N. Campbell & P. T. Smith (Eds.), *Recent Advances in the Psychology of Language* (pp. 347-260). Springer.

Butterworth, B., & Hadar, U. (1989). Gesture, speech, and computational stages: A reply to McNeill. *Psychological Review*, *96*, 168-174. https://doi.org/10.1037/0033-295x.96.1.168

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS ONE*, *5,* e10729. https://doi.org/10.1371/journal.pone.0010729

Chui, K. (2005). Temporal patterning of speech and iconic gestures in conversational discourse. *Journal of Pragmatics*, *37*, 871-887. https://doi.org/10.1016/j.pragma.2004.10.016

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

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*, 181-204. https://doi.org/10.1017/S0140525X12000477

Cooperrider, K. (2017). Foreground gesture, background gesture. *Gesture*, *16*, 176-202. https://doi.org/10.1075/gest.16.2.02coo.

Cuffari, E., & Streeck, J. (2017). Taking the world by hand: How (some) gestures mean. In C, Meyer, J. Streeck, & J. S. Jordan (Eds.), Intercorporeality: Emerging Socialities in Interaction (pp. 172-202). Oxford University Press.

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive Psychology*, *42*, 317-367. https://doi.org/10.1006/cogp.2001.0750.

De Groot, F., Koelewijn, T., Huettig, F., & Olivers, C. N. L. (2016). A stimulus set of words and pictures matched for visual and semantic similarity. *Journal of Cognitive Psychology*, *28*, 1-15. https://doi.org/10.1080/20445911.2015.1101119

DeLong, K., Urbach, T., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*, 1117-1121. https://doi.org/10.1038/nn1504

Dell, G. S., & Chang, F. (2014). The P-chain: relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions Royal Society o*f London, Series B: Biological Sciences, *369*, 20120394. https://dx.doi.org/10.1098%2Frstb.2012.0394

Demir-Lira, O. E., Asaridou, S., Beharelle, A. R., Holt, A., Goldin-Meadow, S., & Small, S. (2018). Functional neuroanatomy of gesture-speech integration in children varies with individual differences in gesture processing. *Developmental Science*, *21*, e12648. https://doi.org/10.1111/desc.12648

de Ruiter, J. (2000). The production of gesture and speech. In D. McNeill (Ed.), Language and Gesture (pp. 248-311). Cambridge University Press.

Dick, A. S., Goldin-Meadow, S., Hasson, U., Skipper, J. I., & Small, S. L. (2009). Co-Speech gestures influence neural activity in brain regions associated with processing semantic information. *Human Brain Mapping*, *30*, 3509-3526. https://dx.doi.org/10.1002%2Fhbm.20774

Dick, A. S., Mok, E. H., Beharelle, A. R., Goldin-Meadow, S., & Small, S. L. (2014). Frontal and temporal contributions to understanding the iconic co-speech gestures that accompany speech. *Human Brain Mapping*, *35*, 900-917. https://dx.doi.org/10.1002%2Fhbm.22222

Dong, X. (2011). Lexicalization: The Origin and Evolution of Chinese Disyllabic Words. The Commercial 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*, 212-222. https://doi.org/10.1044/2016_JSLHR-H-16-0101

Drijvers, L., Özyürek, A., & Jensen, O. (2018). Hearing and seeing meaning in noise: Alpha, beta and gamma oscillations predict gestural enhancement of degraded speech comprehension. *Human Brain Mapping*, *39*, 2075-2087. https://doi.org/10.1002/hbm.23987

Drijvers, L., Vaitonyte, J., & Özyürek, A. (2019). Degree of language experience modulates visual attention to visible speech and iconic gestures during clear and degraded speech comprehension. *Cognitive Science*, *43*, e12789. https://doi.org/10.1111/cogs.12789

Eisner, F., & McQueen, J. M. (2018). Speech perception. In D. Thompson-Schill (Ed.), *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience* (pp. 1-46).Wiley.

Federmeier, K. D. (2007). Thinking ahead: the role and roots of prediction in language comprehension. *Psychophysiology*, *44*, 491-505. https://doi.org/10.1111/j.1469-8986.2007.00531.x

Ferré, G. (2010). Timing relationship between speech and co-verbal gestures in spontaneous French. *Language Research and Evaluation, Workshop on Multimodal Corpora*, 86-91.

Ferreira, F., & Chantavarin, S. (2018). Integration and prediction in language processing: a synthesis of old and new. *Current Directions in Psychological Science*, *27*, 443-448. https://doi.org/10.1177%2F0963721418794491

Frank, M. C., Braginsky, M., Marchman, V. A., and Yurovsky, D. (in press). Variability and Consistency in Early Language Learning: The Wordbank Project. MIT Press.

Frisson, S., Rayner, K., & Pickering, M. J. (2005). Effects of contextual predictability and transitional probability on eye movements during reading. *Journal of Experimental Psychology :Learning, Memory and Cognition*, *31*, 862-877. https://doi.org/10.1037/0278-7393.31.5.862

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences of the United States of America*, *110*, 8051-8056. https://doi.org/10.1073/pnas.1216438110

Green, A., Straube, B., Weis, S., Jansen, A., Willmes, K., Konrad, K., & Kircher, T. (2009). Neural integration of iconic and unrelated coverbal gestures: A functional MRI study. *Human Brain Mapping*, *30*, 3309-3324. https://doi.org/10.1002/hbm.20753

Grisoni, L., Miller, T. M., & Pulvermüller, F. (2017). Neural correlates of semantic prediction and resolution in sentence processing. *The Journal of Neuroscience*, *37*, 4848-4858. https://doi.org/10.1523/jneurosci.2800-16.2017

Goldin-Meadow, S.,& Sandhofer C. M. (1999). Gestures convey substantive information about a child's thoughts to ordinary listeners. *Developmental Science*, *2*, 67-74. https://doi.org/10.1111/1467-7687.00056

Habets, B., Kita, S., Shao, Z., Özyürek, A., & Hagoort, P. (2011). The role of synchrony and ambiguity in speech–gesture integration during comprehension. *Journal of Cognitive Neuroscience*, *23*, 1845-1854. https://doi.org/10.1162/jocn.2010.21462

Hadar, U., & Butterworth, B. (1997). Iconic, gestures, imagery, and word retrieval in speech. *Semiotica*, *115*, 147-172. https://doi.org/10.1515/semi.1997.115.1-2.147

Hagoort, P. (2003). How the brain solves the binding problem for language: A neurocomputational model of syntactic processing. *Neuroimage*, *20*, S18-S29. https://doi.org/10.1016/j.neuroimage.2003.09.013

Hagoort, P. (2005). On Broca, brain, and binding: A new framework. *Trends in Cognitive Sciences*, *9*, 416-423. https://doi.org/10.1016/j.tics.2005.07.004

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science*, *304*, 438-441. https://doi.org/10.1126/science.1095455

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies* (pp. 1-8). Association for Computational Linguistics).

Hayashi, M. (2013). Turn allocation and turn sharing. In J. Sidnell, & T. Stivers (Eds.), *The Handbook of Conversation Analysis* (pp. 167-190). Wiley-Blakwell.

Hayashi, M. (2005). Joint turn construction through language and the body: Notes on embodiment in coordination participation in situated activities. *Semiotica*, *156*, 21-53. https://doi.org/10.1515/semi.2005.2005.156.21

Heldner, M., & Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, *38*, 555-568. https://doi.org/10.1016/j.wocn.2010.08.002

Hickok, G. (2012). Computational neuroanatomy of speech production. *Nature Reviews Neuroscience*, *13*, 135-145. https://doi.org/10.1038/nrn3158

Hintz, F., Meyer, A. S., & Huettig, F. (2017). Predictors of verb-mediated anticipatory eye movements in the visual world. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*, 1352-1374. https://doi.org/10.1037/xlm0000388

Hintz, F., Meyer, A. S., & Huettig, F. (2020). Visual context constrains language-mediated anticipatory eye movements. *Quarterly Journal of Experimental Psychology*, *73*, 458-467. https://doi.org/10.1177%2F1747021819881615

Holle, H., Gunter, T. C., Rüschemeyer, S. A., Hennenlotter, A., & Iacoboni, M. (2008). Neural correlates of the processing of co-speech gestures. *Neuroimage*, *39*, 2010-2024. https://doi.org/10.1016/j.neuroimage.2007.10.055

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.

Holler, J., Shovelton, H., & Beattie, G. (2009). Do iconic gestures really contribute to the semantic information communicated in face-to-face interaction? *Journal of Nonverbal Behavior*, *33*, 73-88. https://doi-org.ru.idm.oclc.org/10.1007/s10919-008-0063-9

Hostetter, A. B., & Alibali, M. W., & Kita. S. (2007). I see it hands' eye: representational gestures reflect conceptual demands. *Language and cognitive processes*, *22*(3), 33-336. https://doi.org/10.1080/01690960600632812

Hu, J., & Tao, H. (2017). A corpus-based study of low transitivity features of the verb *nong* in Chinese. *Foreign Language Teaching and Research*, *49*, 50-58. https://ucla.app.box.com/s/vzpbxhg0inoeckpakcy1hdykhqexw4h0

Huettig, F. (2015). Four central questions about prediction in language processing. *Brain Research*, *1626*, 118-135. https://doi.org/10.1016/j.brainres.2015.02.014

Huettig, F., & Altmann, G. T. M. (2005). Word meaning and the control of eye fixation: Semantic competitor effects and the visual world paradigm. *Cognition*, *96*, B23-B32. https://doi.org/10.1016/j.cognition.2004.10.003

Huettig, F., & Mani, N. (2016). Is prediction necessary to understand language? Probably not. *Language, Cognition and Neuroscience*, *31*, 19-31. https://doi.org/10.1080/23273798.2015.1072223

Huettig, F., & McQueen, J. M. (2007). The tug of war between phonological, semantic and shape information in language-mediated visual search. *Journal of Memory and Language*, *57*, 460-482. https://doi.org/10.1016/j.jml.2007.02.001

Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica*, *137*, 151-171 https://doi.org/10.1016/j.actpsy.2010.11.003

Jackendoff, R. (2007). A parallel architecture perspective on language processing. *Brain Research*, *1146*, 2-22. https://doi.org/10.1016/j.brainres.2006.08.111

Katz, J., & Fodor, J. A. (1963). The structure of a Semantic Theory. *Language*, *39*, 170-210 https://www.jstor.org/stable/411200

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

Kelly, S. D., Özyürek, A., & Maris, E. (2010). Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension. *Psychological Science*, *21*, 260-267. https://doi.org/10.1177%2F0956797609357327

Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. *The relationship of verbal and nonverbal communication*, *25*, 207-227. https://www.researchgate.net/publication/243779963_Gesticulation_and_speech_Two_aspects_of_the_process_of_utterance_in_M

Kita, S. (2000). How representational gestures help speaking. In D. McNeill (Ed.), Language and Gesture (pp. 162–185). Cambridge University Press.

Kita, S., Alibali, M. W., & Chu, M. (2017). How do gesture influences influence thinking and speaking? The gesture-for-conceptualization hypothesis. *Psychological Review*, *124*, 245-266. https://psycnet.apa.org/doi/10.1037/rev0000059

Kita, S., & Özyürek, A. (2003). What does crosslinguistic variation in semantic coordination of speech and gesture reveal? Evidence for an interface representation of spatial

thinking and speaking. *Journal of Memory and Language*, *48*, 16-32. https://doi.org/10.1016/S0749-596X(02)00505-3

Knoeferle, P., & Crocker, M. W. (2006). The coordinated interplay of scene, utterance, and world knowledge: Evidence from eye tracking. *Cognitive Science*, *30*, 481-529 https://doi.org/10.1207/s15516709cog0000_65

Krauss, R. M., Morrel-Samuels, P., & Colsante, C. (1991). Do conversational hand gestures communicate? *Journal of Personality and Social Psychology*, *61*, 743-754 https://psycnet.apa.org/doi/10.1037/0022-3514.61.5.743

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31*, 32-59. https://doi.org/10.1080/23273798.2015.1102299

Kuzenstova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in Linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1-26. https://doi.org/10.18637/jss.v082.i13

Lakens, D., & Evers, E. R. K. (2014). Sailing from the seas of chaos into the corridor of stability: Practical recommendations to increase the informational value of studies. *Perspectives on Psychological Science*, *9*, 278-292. https://doi.org/10.1177/1745691614528520

Lau, E., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (De) constructing the N400. *Nature Reviews Neuroscience*, *9*, 920-933 https://doi.org/10.1038/nrn2532

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

Leonard, T., & Cummins, F. (2011). The temporal relation between beat gestures and speech. *Language and Cognitive Processes*, *26*, 1457-1471 https://doi.org/10.1080/01690965.2010.500218

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

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*, 1126-1177. https://doi.org/10.1016/j.cognition.2007.05.006

Li, C. N., & Thompson, S. A. (1981). Mandarin Chinese: A Functional Reference Grammar University of California Press.

Macmillan, N. A., & Creelman, C. D. (1991). Detection theory: A user's guide. Cambridge University Press.

Marslen-Wilson, W. & Warren, P. (1994). Levels of perceptual representation and process in lexical access: words, phonemes, and features. *Psychological Review*, 101, 653-675. https://doi.org/10.1037/0033-295x.101.4.653

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

McNeill, D. (2005). Gesture and Thought. The University of Chicago Press.

McNeill, D. (2015). Gesture in linguistic. In J. D. Wright (Ed.) International Encyclopedia of the Social and Behavioral Sciences (pp. 109-120). Elsevier.

McNeill, D., Cassell, J., & McCullough, K-E. (1994). Communicative effects of speech-mismatched gestures. *Research on Language and Social Interaction*, *27*, 223-238. https://doi.org/10.1207/s15327973rlsi2703_4

McQueen, J. M., & Huettig, F. (2014). Interference of spoken word recognition through phonological priming from visual objects and printed words. *Attention, Perception & Psychophysics*, *76*, 190-200. https://doi.org/10.3758/s13414-013-0560-8

McQueen, J. M., Cutler, A., & Norris, D. (2003). Flow of information in the spoken word recognition system. *Speech Communication*, 41, 257-270 https://doi.org/10.1016/S0167-6393(02)00108-5

Milburn, E., Warren, T., & Dickey, M. W. (2016). World knowledge affects prediction as quickly as selectional restrictions: Evidence from the visual world paradigm. *Language, Cognition and Neuroscience*, *31*, 536-548. https://doi.org/10.1080/23273798.2015.1117117

Morrel-Samuels, P., & Krauss, R. M. (1992). Word familiarity predicts temporal asynchrony of hand gestures and speech. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *18*(3), 615-622. https://psycnet.apa.org/doi/10.1037/0278-7393.18.3.615

Müller, C. On the gestural creation of narrative structure: A case study of a story told in a conversation. In I. Poggi., M. Rector & N. Trigo (Eds.), Gestures: Meaning and Use (pp. 259–265), Universidade Fernando Pessoa.

Munhall, K., Gribble, P., Sacco, L., and Ward, M. (1996). Temporal constraints on the McGurk effect. *Perception & Psychophysics*, *58*(3), 351-362. https://doi.org/10.3758/BF03206811

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., Von Grebmer Zu Wolfsthurn, S., Bartolozzi, F., Kogan, V., Ito, A., Mézière, D., Barr, D. J., Rousselet, G., Ferguson, H. J., Busch-Moreno, S., Fu, X., Tuomainen, J., Kulakova, E., Husband, E. M., Donaldson, D. I., Kohút, Z., Rueschemeyer, S.-A., & Huettig, F. (2018). Large-scale replication study reveals a limit on probabilistic

prediction in language comprehension. *eLife*, *7*, e33468. https://doi.org/10.7554/eLife.33468.001

Norris, D., McQueen, J. M., & Cutler, A. (1995). Competition and segmentation in spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 1209-1228 https://doi.org/10.1037//0278-7393.21.5.1209

Norris, D., McQueen, J. M., & Cutler, A. (2016). Prediction, Bayesian inference and feedback in speech recognition. *Language, Cognition and Neuroscience*, *31*, 4-18. https://dx.doi.org/10.1080%2F23273798.2015.1081703

Ortega, G., Schiefner, A., & Özyürek, A. (2017). Speakers' gestures predict the meaning and perception of iconicity in signs. In G. Gunzelmann, A. Howe, & T. Tenbrink (Eds.), Proceedings of the 39th Annual Conference of the Cognitive Science Society (CogSci 2017)(pp. 889-894). Austin, TX: Cognitive Science Society.

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

Özyürek, A., Willems, R. M., Kita, S., & Hagoort, P. (2007). On-line integration of semantic information from speech and gesture: Insights from event-related brain potentials. *Journal of Cognitive Neuroscience*, *19*(4), 605-616. https://doi.org/10.1162/jocn.2007.19.4.605

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

Özyürek, A., & Woll, B. Language in the visual modality: Cospeech gesture and sign language. In P. Hagoort (Ed.), Human language: From genes and brain to behavior (pp. 67-83). MIT Press.

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods, 51,* 195-203**.** https://doi.org/10.3758/s13428-018-01193-y

Pickering, M.J., & Gambi, C. (2018). Predicting while comprehending language: A theory and review. *Psychological Bulletin*, *144*, 1002-1044. https://psycnet.apa.org/doi/10.1037/bul0000158

Pickering, M.J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*, 329-392. https://doi.org/10.1017/s0140525x12001495

Pulvermüller, F., Shtyrov, Y., & Ilmoniemi, R. (2005). Brain signatures of meaning access in action word recognition. *Journal of Cognitive Neuroscience*, *17*, 884-892. https://doi.org/10.1162/0898929054021111.

R Core Team. (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rommers, J., Meyer, A. S., Praamstra, P., & Huettig, F. (2013). The contents of predictions in sentence comprehension: Activation of the shape of objects before they are referred to. *Neuropsychologia*, *51*, 437-447. https://doi.org/10.1016/j.neuropsychologia.2012.12.002

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex*, *17*, 1147-1153. https://doi.org/10.1093/cercor/bhl024

Rothermich, K., & Kotz, S. A. (2013). Predictions in speech comprehension: fMRI evidence on the meter-semantic interface. *NeuroImage*, *70*, 89-100. https://doi.org/10.1016/j.neuroimage.2012.12.013

Schegloff, E. A. On some gestures 'relation to talk. In J. M. Atkinson & J. Heritage (Eds.) Structures of Social Action, (pp. 266-295). Cambridge University Press.

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*, 609-612. https://doi.org/10.1016/j.jrp.2013.05.009

Seyfeddinipur, M. (2006). Disfluency: Interrupting speech and gesture. (Doctor Dissertation).

Singer, M. A., & Goldin-Meadow, S. 92005). Children learn when their teacher's gestures and speech differ. *Psychological Science*, *16*, 85-89. https://doi.org/10.1111/j.0956-7976.2005.00786.x

Staub, A. & Clifton Jr., C. Syntactic prediction in language comprehension: evidence from either... or. *Journal of Experimental Psychology: Learning, Memory, and Cognition* ***32***, 425-436 https://dx.doi.org/10.1037%2F0278-7393.32.2.425 (2006)

Straube, B., Green, A., Bromberger, B., & Kircher, T. (2011). The differentiation of iconic and metaphoric gestures: Common and unique integration processes. *Human Brain Mapping*, *32*, 520-533. https://doi.org/10.1002/hbm.21041

Streeck, J. (2009a). Gesturecraft: The manufacture of meaning. John Benjamins Publishing Company.

Streeck, J. (2009). Forward-gesturing. *Discourse Processes*, *46*, 161-179. https://doi.org/10.1080/01638530902728793

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, *26*, 212. https://doi.org/10.1121/1.1907309

Tao, H. (2003). Toward an Emergent View of Lexical Semantics. *Language and Linguistics*, *4*, 837-856. http://ht37.bol.ucla.edu/publications.html

Tao, H. (2019). List gestures in Mandarin conversation and their implication for understanding multimodal interaction. In X. Li & T. Ono (Eds.), Multimodality in Chinese Interaction (pp. 65-98). Mouton De Gruyter.

Tao, H., & Hu, J. (2019). Structural, semantic, and pragmatic properties of *nong* constructions in Mandarin discourse: Evidence from corpora: Evidence from corpora. *International Journal of Chinese Linguistics*, *6*(1), 162-176. https://doi-org.ru.idm.oclc.org/10.1075/ijchl.18003.tao

Tao, H., & Liu, Y. (2010). From register difference to grammatical difference: grammatical constructions in natural speech and the media (part one). *Dangdai Xiucixue*, *157*, 37-44. https://ucla.app.box.com/s/52orxbk9eee8bhnysu5d1i4mkxyjwrk6.

Tao, H., & Liu, Y. (2010). From register difference to grammatical difference: grammatical constructions in natural speech and the media (part two). *Dangdai Xiucixue*, *158*, 22-27 https://ucla.app.box.com/s/52orxbk9eee8bhnysu5d1i4mkxyjwrk6

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *26*8, 1632-1634. https://doi.org/10.1126/science.7777863

ter Bekke, M., Drijvers, L., & Holler, J. (2020). Hand gestures have predictive potential during face-to-face conversation. https://doi.org/10.6084/m9.figshare.12415847.v3

Thompson, S., & Hopper, P. J. Transitivity, clause structure, and argument structure: evidence from conversation. In J. Bybee & P. J. Hopper (Eds.), Frequency and the Emergence of the Linguistic Structure (pp. 27–60). John Benjamins Publishing Company.

Van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 443-467. https://doi.org/10.1037/0278-7393.31.3.443

Van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *67*, 1176-1190. https://doi.org/10.1080/17470218.2013.850521

Van Petten, C., & Luka, B.J. (2012). Prediction during language comprehension: benefits, costs, and ERP components. *International Journal of Psychophysiology*, *83*, 176-190. https://doi.org/10.1016/j.ijpsycho.2011.09.015

Vulchanova, M., Vulchanov, V., Fritz, I., & Milburn, E. A. (2019). Language and perception: Introduction to the Special Issue ‘‘Speakers and Listeners in the Visual World’’. *Journal of Cultural Cognitive Science*, *3*, 103-112. https://doi.org/10.1007/s41809-019-00047-z.

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

Wicha, N. Y. Y., Bates, E. M., Moreno, E. M., & Kutas, M. (2003). Potato not Pope: Human brain potentials to gender expectation and agreement in Spanish spoken sentences. *Neuroscience letters*, *346*, 165-168. https://doi.org/10.1016/s0304-3940(03)00599-8

Wicha, N. Y. Y., Moreno, E. M., & Kutas, M. (2004). Anticipating words and their gender: An event-related brain potential study of semantic integration, gender expectancy, and gender agreement in Spanish sentence reading. *Journal of Cognitive Neuroscience*, *16*, 1272-1288. https://doi.org/10.1162/0898929041920487

Willems, R. M., Özyürek, A., & Hagoort, P. (2007). When language meets action: The neural integration of gesture and speech. *Cerebral Cortex*, *17*, 2322-2333. https://doi.org/10.1093/cercor/bhl141

Willems, R. M., Özyürek, A., & Hagoort, P. (2009). Differential roles for left inferior frontal and superior temporal cortex in multimodal integration of action and language. *Neuroimage*, *47*, 1992-2004. https://doi.org/10.1016/j.neuroimage.2009.05.066

Yee, E., & Sedivy, J. C. (2006). Eye movements to pictures reveal transient semantic activation during spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 1-14. https://doi.org/10.1037/0278-7393.32.1.1.

Acknowledgements

The authors received no specific funding for this work.

Competing interests

The authors have declared that no competing interests exist.

Table 2. Design Table

Research questions	Hypothesis	Sampling plan	Analysis plan	Interpretation
To what extent can iconic gesture predict the upcoming nominal word independently from the semantic restriction of the linguistic input?	1) Participants generate higher fixation proportion to the target object than to the distractors (averaged fixation proportion) by the time of hearing the signifier of the target in the gesture-only condition than that in the neutral condition. 2) Participants generate higher fixation proportion to the target object than to the distractors (averaged fixation proportion) by the time of hearing the signifier of the target in the gesture+speech biasing condition than that in the speech-only biasing condition. 3) Participants generate higher fixation proportion to the target object than to the distractors (averaged fixation proportion) by the time of hearing the signifier of the target in the gesture-only biasing condition than that in the speech-only biasing condition relative to the neutral condition.	Based on the recommended sample size per condition provided by Lakens and Evers (2014, p. 280), to achieve 80% statistical power to observe the effect with an alpha of .05, for an estimated effect size (r = .3), we aim to recruit 180 participants (45 for each condition). As indicated by Lakens and Evers (2014, p. 280), if we consider the point of stability for the correlation magnitudes, after which the sample estimates do not deviate from a "corridor of stability" around the true population value, the corresponding effect size of 45/condition falls between .3 to .4, based on a wider corridor of w = .2 with an 80% stability confidence.	We use R (R Core Team, 2019) and lme4 (Bates et al.,2015) to perform a linear mixed effects analysis of the relationship between fixation proportion and gesture-availability and predictability (high constrained vs. low constrained). The dependent variable is the log-transformed fixation ratios. As fixed effects, we enter gestural-availability (available vs unavailable) as well as the speech condition (high constrained vs. low constrained) into the model. Participants and items are included as random factors, each with random intercepts and slopes. Interactions between condition are added. The neutral condition is put on the intercept. *P*-values are obtained by using the package lmerTest (version 3.1-1; Kuznetsova et al., 2017).	1) Gesture-only vs. Neutral: By the time of encountering the target in speech, if the target attracts significantly more fixation than the distractors, we consider that participants exploit information extracted from gesture to make prediction. If the target does not attract significantly more fixation than any of the distractors, we consider that gesture does not play significantly influential role in prediction. 2) Gesture+speech biasing vs. Speech-only By the time of encountering the target in speech, if the target attracts significantly more fixation comparing with that in the speech-only condition, we consider gesture significantly facilitates predictive language processing. If the target does not attract significantly more fixation comparing with that in the speech-only condition, we consider that gesture does not significantly boost predictive language processing. 3) Gesture-only vs. Speech-only (relative to Neutral condition) Relative to the neutral condition, by the time of encountering the target in speech, if the target attracts significantly more fixation in the gesture-only biasing condition than that in the speech-only biasing condition, we consider that gesture plays significantly more influential role in predictive language processing than speech does. If the target does not attract significantly more fixation in the gesture-only biasing condition than that in the speech-only biasing condition, we consider that gesture does not play significantly more influential role in predictive language processing than speech does.

Appendix

Supplementary Information: Eye-tracking Study: Stimulus Material

Target object	Constrained verb	Neutral verb	Distractor 1	Distractor 2	Distractor 3
门 (door)	敲 (knock)	安 (install)	秋千(swing)	监控 (monitor)	马桶 (toilet)
钢琴 (piano)	弹 (play)	搬 (move)	冰箱 (refrigerator)	垃圾桶 (trash bin)	床垫 (mattress)
哑铃 (dumbbell)	练 (practice)	握 (grip)	叉子 (fork)	方向盘 (steering wheel)	挂钩 (coat hook)
飞镖 (dart)	扔 (throw)	观察 (observe)	蝴蝶 (butterfly)	巡洋舰 (cruiseship)	钟楼 (belltower)
箭 (arrow)	射 (shoot)	捆 (bundle)	包裹 (package)	衣架 (hanger)	电话线 (telephoneline)
香蕉(banana)	剥 (peel)	拿 (bring)	计算器 (calculator)	眼镜盒 (glasscase)	手电筒 (flashlight)
纸 (paper)	撕 (tear)	丢 (throw)	椅子(chair)	拖鞋 (slipper)	口红(lipstick)
烟 (cigar)	抽 (smoke)	找 (find)	气筒(pump)	帽子 (cap)	笔记本 (notebook)
车(taxi)	开 (drive)	试 (try)	降落伞 (parachute)	胶卷 (film)	泳镜 (swimingoggle)
果酱 (jam)	舀 (scoop)	费 (cost)	子弹 (bullet)	粉笔 (chalk)	肥皂 (soap)
鼠标 (mouse)	点 (click)	碰 (touch)	酒杯 (beer cup)	大炮 (cannon)	烤箱 (oven)
笛子 (flute)	吹 (play)	做 (make)	围裙 (apron)	风筝 (kite)	面包 (bread)
遥控器 (remotor)	按 (press)	查 (check)	消防栓 (firehydrant)	煤气灶 (gas burner)	捕鼠器 (mousetrap)
黄瓜 (cucumber)	切 (chop)	晾 (dry)	拖把 (mop)	毛巾 (towel)	救生衣 (life jacket)
色子 (dice)	掷 (throw)	研究 (study)	微波炉 (microwave)	城堡 (castle)	电梯 (escalator)
小提琴 (violin)	拉 (play)	背 (bear)	斧头 (axe)	救生圈 (life bouy)	望远镜 (telescope)
手机 (cellohone)	刷 (slide)	修 (fix)	电视 (TV)	剃须刀 (shaver)	暖气 (heater)
把手 (handle)	拽 (pull)	看 (search)	旱冰鞋 (skiboot)	音响 (speaker)	相机 (camera)
香水(perfume)	喷 (spray)	聊 (chat)	台球 (pool ball)	房子 (house)	蛋糕 (cake)
体温计 (thermometer)	甩 (toss)	换 (change)	水龙头 (tap)	火车 (train)	窗户 (window)
面条 (noddles)	擀 (roll)	煮 (boil)	苹果 (apple)	南瓜 (pumpkin)	螃蟹 (carb)
篮球 (basketball)	投 (play)	还 (return)	尺子 (ruler)	萨克斯管 (saxphone)	指甲剪 (nail clipper)
羽毛球 (badminton)	打 (play)	学 (learn)	三角铁 (triangle)	跳绳 (skipping rope)	自行车 (bicycle)
扣子 (button)	缝 (sew)	系 (tie)	头盔 (helmet)	垃圾袋 (trash bad)	腰带 (belt)
树枝 (branch)	掰 (break)	烧 (burn)	礼物 (gift)	帐篷 (tent)	气球 (balloon)
火柴 (match)	划 (strike)	摆 (put)	风扇 fan)	筹码 (porketchip)	吸管 (straw)
可乐 (cola)	喝 (drink)	买 (buy)	鞋带 (shoelace)	书 (book)	拐杖 (crutch)
雨伞 (umbrella)	撑 (unfold)	拎 (carry)	书包 (schoolbag)	袋子(bag)	煤气罐 (gas cylinder)
咖啡豆 (coffeebean)	磨 (grind)	挑 (select)	戒指 (ring)	日历 (calendar)	灯泡 (bulb)
土豆 (potato)	削 (peel)	抬 (lift)	草莓 (strawberry)	电子琴 (electronic organ)	向日葵 (sunflower)
柠檬 (lemon)	挤 (squeeze)	用 (consume)	电池 (battery)	回形针 (clip)	餐巾纸 (tissue)
饺子 (dumpling)	包 (make)	订 (order)	葡萄酒 (wine)	吹风机 (hairdryer)	台灯 (lamp)
纺车 (spinning wheel)	摇 (spin)	毁 (destroy)	开瓶器 (corkscrew)	螺丝刀 (screwdriver)	地球仪 (globe)
蒜 (garlic)	捣 (mash)	借 (borrow)	平底锅 (pan)	别针 (safety pin)	旱冰鞋 (skiboot)
木板 (woodboard)	锯 (saw)	送 (dispatch)	发卡 (hairband)	西红柿 (tomato)	水管 (water pipe)
桌子 (desk)	擦 (clean)	选 (choose)	领结 (bowtie)	毛线 (wool)	耳环 (earring)
碗 (bowl)	洗 (wash)	介绍 (introduce)	打印机 (printer)	竖琴 (harp)	高尔夫 (golf)
电源线 (power line)	拔 (pull)	缠 (twine)	围巾 (scarf)	窗帘 (curtain)	球拍 (pat)
衬衫 (shirt)	熨 (iron)	叠 (fold)	尿布 (dipper)	信封 (envelope)	纸船 (paper ship)
摩托车 (motorcycle)	骑 (drive)	卖 (sell)	面具 (mask)	牙刷 (toothbrush)	打火机 (lighter)
沙子 (snad)	筛 (sieve)	运 (transport)	西瓜 (watermelon)	柜子 (closet)	电脑 (laptop)
头发 (hair)	剪 (cut)	整理 (tidy)	指甲油 (nail polish)	夹子 (paperclip)	唱片 (disc)
牛仔裤 (jeans)	抖 (shake)	补 (patch)	鼓 (drum)	皮鞋 (shoes)	轮胎 (tire)
瓜子 (sunflower seed)	捧 (cradle)	带 (take)	木桶 (barrel)	显微镜 (microscope)	颜料 (watercolor)
鱼 (fish)	钓 (fish)	收拾 [2] (clean)	淋浴 (shower)	空调 (air conitioner)	除草剂 (lawnmower)
树 (tree)	砍 (cut)	晃 (shake)	储钱罐 (piggybank)	铃铛 (bell)	鱼缸 (aquarium)
花 (flower)	浇 (water)	插 (insert)	蜡烛 (kindle)	旗子 (flag)	三脚架 (tripod)
娃娃 (doll)	抓 (pick)	造 (produce)	台球桌 (poll table)	快艇 (motorboat)	清洁剂 (dish soap)

Eye-tracking Study: Stimulus Material

1. 我今天**敲**/**安**了好长时间的门。

   I today **knocked** at / **settled** for a long time the door. (literal English translation)

   I today **knocked** at / **settled** the door for a long time. (grammatical English translation)

2. 我今天**弹**/**搬**了一下午的钢琴。

   I today **play** / **move** a whole afternoon piano. (literal English translation)

   Today I **played** / **moved** the piano for a whole afternoon. (grammatical English translation)

3. 我今天**练**/**握**了几下那个哑铃。

   I today **practise** / **grip** several times that dumbbell. (literal English translation)

   Today I **practised** with at / **griped** that dumbbell for several times. (grammatical English translation)

4. 我今天**扔**/**观察**了一下午的飞镖。

   I today **throw** / **observe** a whole afternoon dart. (literal English translation)

   Today I **threw** / **observed** the darts for a whole afternoon. (grammatical English translation)

5. 我今天**射**/**捆**了一下午的箭。

   I today **bundle** / **shoot** a whole afternoon arrow. (literal English translation)

   Today I **bundled** / **shot** the arrows for a whole afternoon. (grammatical English translation)

6. 我今天**剥**/**拿**了好多好多的香蕉。

   I today **peel** / **bring** lots of banana. (literal English translation)

   Today I **peeled** / **brought** lots of bananas. (grammatical English translation)

7. 我今天**撕**/**丢**了很多很多的纸。

I today **tear** / **throw** lots of paper. (literal English translation)

Today I **tore** / **threw** lots of papers. (grammatical English translation)

8. 我今天抽/找了一下午的烟。

I today **smoke** / **find** a whole afternoon cigarette. (literal English translation)

Today I **smoked** / **found** the cigarette for a whole afternoon. (grammatical English translation)

9. 我今天开/试了一下午的车。

I today **drive** / **try** MW car. (literal English translation)

Today I **drove** / **tried** the car for a bit. (grammatical English translation)

10. 我刚才舀/费了很多很多的的果酱。

I just now **scoop** / **consume** lots of jam. (literal English translation)

Just now I **scooped** / **consumed** lots of jam. (grammatical English translation)

11. 我刚才点/碰了两下那个鼠标

I just now **click** / **touch** several times that mouse. (literal English translation)

Just now I **clicked** / **touched** that mouse for several times. (grammatical English translation)

12. 我今天吹/做了一下午的笛子。

I today **play** / **make** a whole afternoon flute. (literal English translation)

Today I **played** / **made** the door for a whole afternoon. (grammatical English translation)

13. 我今天按/查了好长时间的遥控器。

I today **press** / **check** a long time the remoter. (literal English translation)

Today I **pressed** / **checked** the remoter for a long time. (grammatical English translation)

14. 我今天切/晾了一下午的黄瓜。

I today **chop** / **dry** a whole afternoon cucumber. (literal English translation)

Today I **chopped** / **dried** the cucumbers for a whole afternoon. (grammatical English translation)

15. 我刚才掷/研究了好长时间的色子。

I today **throw** / **probe** a long time dice. (literal English translation)

Today I **threw** / **probed** the dice for a long time. (grammatical English translation)

16. 我今天拉/背了一下午的小提琴。

I today **play** / **bear** a whole afternoon violin. (literal English translation)

Today I **played** / **bore** the violin for a whole afternoon. (grammatical English translation)

17. 我今天刷/修了一下午的手机。

I today **slide** / **fix** a whole afternoon cell phone. (literal English translation)

Today I **slid** / **fixed** the cell phone for a whole afternoon. (grammatical English translation)

18. 我刚才拽/看了两下那个把手。

I just now **pull** / **look** MW handle. (literal English translation)

Just now I **pulled** / **looked** at the handle for a bit. (grammatical English translation)

19. 我今天喷/聊了一下午的香水。

I today **spray** / **talk** a whole afternoon perfume. (literal English translation)

Today I **sprayed** / **talked** the perfume for a whole afternoon. (grammatical English translation)

20. 我今天甩/换了好几次体温计。

I today **toss** / **change** several times thermometer. (literal English translation)

Today I **tossed** / **changed** the thermometer for several times. (grammatical English translation)

21. 我今天擀/煮了一下午的面条。

I today **roll** / **boil** a whole afternoon noodles. (literal English translation)

Today I **rolled** / **boiled** the noodles for a whole afternoon. (grammatical English translation)

22. 我今天投/还了一下午的篮球。

I today **shoot** / **return** a whole afternoon basketball. (literal English translation)

Today I **shot** / **returned** the basketball for a whole afternoon. (grammatical English translation)

23. 我今天打/学了一下午的羽毛球

I today **play** / **learn** a whole afternoon badminton. (literal English translation)

Today I **played** / **learnt** the badminton for a whole afternoon. (grammatical English translation)

24. 我今天缝/系了好长时间的扣子。

I today **sew** / **tie** a long time the button. (literal English translation)

Today I **sewed** / **tied** the button for a long time. (grammatical English translation)

25. 我今天掰/烧了一下午的树枝。

I today **break** / **fire** a whole afternoon branch. (literal English translation)

Today I **broke** off / **fired** the branches for a whole afternoon. (grammatical English translation)

26. 我今天划/摆了好长时间火柴。

I today **strike** / **place** a long time match. (literal English translation)

Today I **stroke** / **placed** the match for a long time. (grammatical English translation)

27. 我今天喝/买了一下午的可乐。

I today **drink** / **buy** a whole afternoon cola. (literal English translation)

Today I **drank** / **brought** the cola for a whole afternoon. (grammatical English translation)

28. 我今天**撑**/**拎**了一下午的雨伞。

I today **put up** / **carry** a whole afternoon umbrella. (literal English translation)

Today I **put up** / **carried** the umbrella for a whole afternoon. (grammatical English translation)

29. 我今天**磨**/**挑**了一下午的咖啡豆。

I today **grind** / **pick out** a whole afternoon coffee bean. (literal English translation)

Today I **grinded** / **picked out** the coffee bean for a whole afternoon. (grammatical English translation)

30. 我今天**削**/**抬**了一下午的土豆。

I today **peel** / **carry** a whole afternoon potato. (literal English translation)

Today I **peeled** / **carried** the potato for a whole afternoon. (grammatical English translation)

31. 我今天**挤**/**用**了一下午的柠檬。

I today **squeeze** / **consume** a whole afternoon lemon. (literal English translation)

Today I **squeezed** / **consumed** the lemons for a whole afternoon. (grammatical English translation)

32. 我今天**包**/**订**了一下午的饺子。

I today **make** / **book** a whole afternoon dumpling. (literal English translation)

I today **made** / **book**ed the dumplings for a whole afternoon. (grammatical English translation)

33. 我今天**摇**/**毁**了好多好多的纺车。

I today **spin** / **destroy** lots of spinning wheel. (literal English translation)

Today I **spun** / **destroyed** lots of spinning wheels. (grammatical English translation)

34. 我今天捣/借了一下午的蒜。

I today **mash** / **borrow** a whole afternoon garlic. (literal English translation)

Today I **mashed** / **borrowed** the garlic for a whole afternoon. (grammatical English translation

35. 我今天锯/送了一下午的木板。

I today **saw** / **dispatch** a whole afternoon wood board. (literal English translation)

Today I **sawed** / **dispatched** the wood board for a whole afternoon. (grammatical English translation)

36. 我今天擦/选了一下午的桌子。

I today **clean** / **choose** a whole afternoon table. (literal English translation)

Today I **cleaned** at / **chose** the table for a whole afternoon. (grammatical English translation)

37. 我今天洗/介绍了一下午的碗。

I today **wash** / **introduce** a whole afternoon bowl. (literal English translation)

Today I **washed** / **introduced** the bowls for a whole afternoon. (grammatical English translation)

38. 我今天拔/缠了好几下那个电源线。

I today **pull** out / **twine** several times that power line. (literal English translation)

Today I tried to **pull** out / **twine** that power line for several times. (grammatical English translation)

39. 我今天熨/叠了一下午的衬衫。

I today **iron** / **fold** a whole afternoon shirt. (literal English translation)

Today I **ironed** / **folded** the shirt for a whole afternoon. (grammatical English translation)

40. 我今天**骑**/**卖**了一下午的摩托车。

I today **ride** / **sell** a whole afternoon motorcycle. (literal English translation)

I today **rode** / **sold** the motorcycle for a whole afternoon. (grammatical English translation)

41. 我今天**筛**/**运**了一下午的沙子。

I today **sieve** / **transport** a whole afternoon sand. (literal English translation)

Today I **sieved** / **transported** the sands for a whole afternoon. (grammatical English translation

42. 我今天**剪**/**整理**了好长时间的头发。

I today **cut** / **tidy** a long time hair. (literal English translation)

Today I **cut** / **tidied** the hair for a long time. (grammatical English translation)

43. 我今天**抖**/**补**了一下午的牛仔裤。

I today **shake** / **patch** a whole afternoon jeans. (literal English translation)

Today I **ahaked** / **patched** the jeans for a whole afternoon. (grammatical English translation)

44. 我今天**捧**/**带**了好多好多的瓜子。

I today **cradle** / **take** lots of sunflower seed. (literal English translation)

Today I **cradled** / **took** lots of sunflower seeds. (grammatical English translation)

45. 我今天**钓**/**收拾**了一下午的鱼。

I today **fish** / **prepare** a whole afternoon door. (literal English translation)

Today I **fished** / **prepared** the fish for a whole afternoon. (grammatical English translation)

46. 我今天**砍**/**晃**了两下那个树。

I today **cut** / **shake** several times tree. (literal English translation)

Today I **cut** / **shake** the trees for several times. (grammatical English translation)

47. 我今天**浇**/**插**了一下午的花。

I today **water** / **insert** a whole afternoon flower. (literal English translation)

Today I **watered** / **inserted** the flowers for a whole afternoon. (grammatical English translation)

48. 我今天**抓**/**造**了好几个娃娃。

I today **pick up** at / **produced** several doll. (literal English translation)

Today I **picked** up / **produced** several dolls. (grammatical English translation)

49. 我今天**抱**了一下午的搅拌机。

I today **hold** a whole afternoon mixer. (literal English translation)

Today I **held** the mixer for a whole afternoon. (grammatical English translation)

50. 我今天**装**了一下午的饼干。

I today **package** a whole afternoon cookie (literal English translation)

Today I **packaged** the cookie for a whole afternoon. (grammatical English translation)

51. 我今天**寄**了一下午的运动鞋。

I today **send** a long time sneaker. (literal English translation)

Today I **sent** the sneaker for a whole afternoon. (grammatical English translation)

52. 我今天**砸**了一下午的核桃。

I today **smash** a whole afternoon walnut. (literal English translation)

Today I **smashed** the walnut for a whole afternoon. (grammatical English translation)

53. 我今天**搜**了一下午的榨汁机。

I today **search** a whole afternoon blender. (literal English translation)

Today I **searched** the blender for a whole afternoon. (grammatical English translation)

54. 我今天**贴**了一下午的瓷砖。

I today **stick** a whole afternoon floor tile. (literal English translation)

Today I **struck** the floor tile for a whole afternoon. (grammatical English translation)

55. 我今天**捡**了一下午的蘑菇。

I today **pick up** a whole afternoon mushroom. (literal English translation)

Today I **picked up** the mushroom for a whole afternoon. (grammatical English translation)

56. 我今天**推**了一下午的货车。

I today **push** a whole afternoon truck. (literal English translation)

Today I **pushed** the truck for a whole afternoon. (grammatical English translation)

57. 我今天**泡**了一下午的洋葱。

I today **put** in water a whole afternoon onion. (literal English translation)

Today I **put** the onion in water for a whole afternoon. (grammatical English translation)

58. 我今天**等**了一下午的体恤衫。

I today **wait** a whole afternoon T-shirt. (literal English translation)

Today I **waited** the T-shirt for a whole afternoon. (grammatical English translation)

59. 我今天**取**了一下午的眼镜。

I today **collect** a whole afternoon glasses. (literal English translation)

Today I **collected** the glasses for a whole afternoon. (grammatical English translation)

60. 我今天**扶**了一下午的梯子。

I today **hold** a whole afternoon ladder. (literal English translation)

Today I **held** the ladder for a whole afternoon. (grammatical English translation