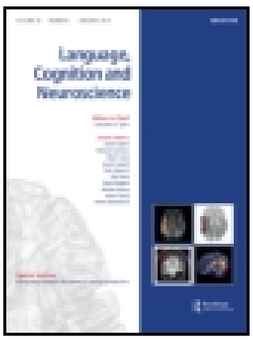

Routledge
Taylor & Francis Group

Language, Cognition and Neuroscience

ISSN: 2327-3798 (Print) 2327-3801 (Online) Journal homepage: https://www.tandfonline.com/loi/plcp21

Evidence for children's online integration of simultaneous information from speech and iconic gestures: an ERP study

Kazuki Sekine, Christina Schoechl, Kimberley Mulder, Judith Holler, Spencer Kelly, Reyhan Furman & Asli Özyürek

To cite this article: Kazuki Sekine, Christina Schoechl, Kimberley Mulder, Judith Holler, Spencer Kelly, Reyhan Furman & Asli Özyürek (2020): Evidence for children's online integration of simultaneous information from speech and iconic gestures: an ERP study, Language, Cognition and Neuroscience, DOI: 10.1080/23273798.2020.1737719

To link to this article: https://doi.org/10.1080/23273798.2020.1737719

LANGUAGE, COGNITION AND NEUROSCIENCE
https://doi.org/10.1080/23273798.2020.1737719

Routledge
Taylor & Francis Group

REGULAR ARTICLE

Evidence for children's online integration of simultaneous information from speech and iconic gestures: an ERP study

Kazuki Sekine [a], Christina Schoechl[b], Kimberley Mulder[c,d], Judith Holler[d,g], Spencer Kelly[e], Reyhan Furman[f] and Asli Özyürek[c,d,g]

[a]Waseda University, School of Human Sciences, Tokorozawa, Japan; [b]University of Massachusetts Boston, Boston, USA; [c]Centre for Language Studies, Radboud University, Nijmegen, The Netherlands; [d]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands; [e]Department of Psychological and Brain Science, Colgate University, Hamilton, USA; [f]School of Psychology, University of Central Lancashire, Preston, United Kingdom; [g]Donders Institute for Brain, Cognition & Behaviour, Radboud University, Nijmegen, The Netherlands

ABSTRACT

Children perceive iconic gestures, along with speech they hear. Previous studies have shown that children integrate information from both modalities. Yet it is not known whether children can integrate both types of information simultaneously as soon as they are available (as adults do) or whether they initially process them separately and integrate them later. Using electrophysiological measures, we examined the online neurocognitive processing of gesture-speech integration in 6- to 7-year-old children. We focused on the N400 event-related potential component which is modulated by semantic integration load. Children watched video clips of matching or mismatching gesture-speech combinations, which varied the semantic integration load. The ERPs showed that the amplitude of the N400 was larger in the mismatching condition than in the matching condition. This finding provides the first neural evidence that by the ages of 6 or 7, children integrate multimodal semantic information in an online fashion comparable to that of adults.

ARTICLE HISTORY
Received 15 June 2019
Accepted 12 February 2020

KEYWORDS
Multimodal integration; co-speech gestures; children; ERPs; N400

1. Background

1.1. Introduction

Face to face language use is multimodal in nature. Speakers often produce gestures while speaking. These gestures, so-called co-speech gestures, are meaningful hand movements related to what is expressed in the accompanying speech and are frequently used together with speech (Chui, 2005; Kendon, 2004; McNeill, 1992; Nobe, 2000). A subset of these gestures, i.e. iconic gestures, represent rich semantic information such as action, movement, shape, or size of a referent (McNeill, 1992). For example, an adult can use a drinking gesture (e.g. tilting a c-shaped hand towards mouth as if drinking) while saying "do you want something to drink?". As they grow up, children are exposed not only to speech but also at the same time to such gestures in their conversations with adults (e.g. Campisi & Özyürek, 2013; Gutmann & Turnure, 1979; O¨zçaliskan & Goldin-Meadow, 2011). It has been shown that children can understand information from such gestures along with speech from three years onwards (Demir-Lira et al., 2018; Sekine et al., 2015; Stanfield et al., 2013), but we still know little about the nature of speech and gesture integration. One important question concerns the online integration of simultaneous speech and gesture and its neural underpinnings in the child brain.

1.2. Integration of iconic gestures with speech in adults

Research with adults has shown that iconic gestures and speech form an integrated system not only in production (e.g. Kita & Özyürek, 2003; McNeill, 1992), but also during comprehension (e.g. Kelly et al., 2010). Previous research has shown that semantic information from iconic gestures is indeed processed by listeners and that iconic gestures can affect language comprehension at behavioral and neural levels, in both clear as well as adverse listening situations, such as in noise (e.g. Beattie & Shovelton, 1999a, 1999b; Drijvers & Özyürek, 2017; Holle & Gunter, 2007; Holler et al., 2014; Holler et al., 2009; Kelly et al., 1999; Kelly et al., 2010; Obermeier et al., 2011; for a review, see Özyürek, 2014). These studies have provided firm evidence that speech and gesture are integrated in adults.

CONTACT Kazuki Sekine ksekine@waseda.jp
Supplemental data for this article can be accessed https://doi.org/10.1080/23273798.2020.1737719

Studies investigating online integration of semantic information from iconic gestures and speech using EEG measurements have focused on event-related potentials (ERP) and the N400 component. ERPs are a method that records electrical brain activity time-locked to some external or internal event. N400 is characterised by a negative deflection measured between 300 and 500 ms after stimulus onset, and it is a good measure to investigate neurocognitive processing of semantic information from multiple modalities such as pictures, words or sentences during language comprehension (Kutas & Hillyard, 1984; Kutas & Federmeiyer, 2000, 2014). The amplitude of the N400 varies as a function of the semantic fit between the meaning of a word and its context, and indexes the ease of semantic processing in language (Hagoort & Van Berkum, 2007). The N400 amplitude is larger in response to semantically mismatching information compared to matching information, and this difference is called the N400 effect.

Previous studies have found N400 effects depending on the degree to which iconic gestures were semantically matched to the previous sentence context (Özyürek et al., 2007) as well as to a single word or sentence accompanying the gesture (e.g. Drijvers & Özyürek, 2018; Habets et al., 2011; Kelly et al., 2004; Özyürek et al., 2007; Wu & Coulson, 2007; for review see Özyürek, 2014). In Drijvers and Özyürek's (2018) study, most relevant to our current study, adults watched a video clip of an actor saying "to drink" while she produced either a matching gesture ("to drink") or a mismatching one ("to type").[1] The co-occurring mismatching information from the visual modality elicited a stronger N400 effect than when adults saw a matching speech-gesture combination. This indicates an increased cognitive load of semantic integration during processing of both speech and gesture and is similar to what has been found in unimodal (auditory) semantic integration studies of words (or gestures) in relation to preceding sentence context (Hagoort & Van Berkum, 2007; Özyürek et al., 2007). We do not know how this process occurs in the child brain.

1.3. Iconic gesture comprehension in children

Recent behavioral studies have shown that children gradually develop their ability to combine gestures with speech they hear as they get older (e.g. Broaders & Goldin-Meadow, 2010; Kelly & Church, 1998; McNeil et al., 2000; Sekine et al., 2015; Stanfield et al., 2013). For example, in Stanfield et al.'s (2013) study with children aged 2–4 years, an experimenter, sitting across from a child, said "I am eating" while simultaneously producing an iconic gesture depicting an action on an object (e.g. moving the hands to the mouth as if eating a sandwich). Later the child was given two different pictures (e.g. sandwich vs. bowl of cereal), one of which always matched the object depicted by the iconic gesture, and was asked to select the picture that best matched what the experimenter had communicated. 3- and 4-year-old children, but not 2-year-olds, were able to reliably select the correct picture. By using a similar experimental procedure to Stanfield et al., Sekine et al. (2015) examined whether children aged 3–5 years can pick the correct picture. To do so, children needed to combine both information from gesture and speech in video clips that were shown prior to the response pictures. Results showed that the proportion of trials with a correct choice in 5-years-olds was significantly higher than in 3-year-olds. Thus, these behavioral studies (e.g. Sekine et al., 2015; Stanfield et al., 2013) suggest that children between the ages three to five are gradually learning to comprehend information conveyed through gesture related to the co-occurring speech.

Findings from these behavioral studies were obtained by using offline behavioral measures such as a forced-choice task with pictures. However, the offline measures make it difficult to conclude whether children integrate gesture and speech in an online manner as shown for adults. This is because, for example, in a forced-choice picture task, children could first narrow down a target action based on information in speech, and then choose the correct picture by matching the gesture in the video with the action in the pictures, rather than combining information from speech and gesture *at the same time* as they perceive the information from the two modalities. It is also possible that the 3-year-olds' integration difficulty could be due to the sequential nature of this task as it requires information to be maintained in working memory between seeing the speech-gesture combination and selecting the picture. Online measures provide more direct ways to assess integration, especially with children.

1.4. Neural measures of speech and gesture integration in children

In spite of the abundance of studies investigating neural processes underlying speech and gesture integration in adults, studies with children are rare. There are two neuroscientific studies focusing on gesture-speech integration in children aged 8–11 (Demir-Lira et al., 2018; Dick et al., 2012). Both studies used functional magnetic resonance imaging (fMRI). Dick et al. (2012) examined the difference between the neural networks that are used to process meaningful co-speech gestures and meaningless self-adaptor movements (e.g. touching one's hair or adjusting one's glasses) in children (8- to

11-year-olds) and adults. They found that compared with adults, children displayed more activity in inferior frontal gyrus, pars triangularis (IFGTr) and posterior middle temporal gyrus (MTGp) for gestures than self-adaptors. The authors interpreted the heightened activation in the sensory-semantic network (IFGTr and MTGp) in children as evidence for children's greater effort in retrieving semantic information from long-term memory to process gesture compared to adults. Demir-Lira et al. (2018) examined the relationship between gesture-speech integration ability and brain activation by presenting children aged 8–11 years with video stimuli that consisted of iconic gestures and speech in the fMRI scanner. They found that when gestures provided complementary information that was not presented in the speech (e.g. saying "pet" while flapping palms representing a bird), brain activity in the inferior frontal gyri (IFG), the right middle temporal gyrus (MTG), and the left superior temporal gyrus (STG) increased, compared to when gesture provided redundant information (e.g. saying "bird" while flapping palms). Importantly, this differential activation across the two conditions was found only in those children who were able to successfully integrate gesture and speech behaviourally as indicated by their performance on a post-test on story comprehension. Furthermore, the brain activation patterns for gesture-speech integration found in those children overlapped with adults, but the activated brain areas in children were broader than those in adults. This shows that children need to recruit a broader set of brain areas during gesture-speech integration than adults (Demir-Lira et al., 2018).

These two brain-imaging studies revealed which brain regions are involved in gesture comprehension and integration in children compared to adults, but they do not provide evidence that children can simultaneously integrate gestures and speech in an online manner, as fMRI's temporal resolution is limited by hemodynamic response time. This is also true for previously mentioned behavioral studies with children, which have used offline measures of gesture-speech integration, such as picture matching. Furthermore, the fact that not all children showed semantic integration behaviourally and neurally in Demir-Lira et al.'s (2018) study suggests there is a developmental trajectory of speech and gesture integration which merits investigating this process in earlier ages.

Even though there are no studies investigating online integration of simultaneous speech and gesture, similar studies have investigated the integration of information from a picture and a word that are presented simultaneously. These studies found an N400 effect that can be used as an index for the multimodal integration when children were presented with a picture that semantically mismatched a word they heard (Friedrich & Friederici, 2010 for 12 months old; Friedrich & Friederici, 2004 for 19 months old; Henderson et al., 2011 for 8–10 years old). They found that children's brain activity during the task showed similar pattern to that in adults. From their findings, it is clear that children can semantically and neurally integrate a picture with a word from infancy in an adult-like manner. However, we do not know yet to what extent children can integrate a word and a co-speech gesture that represent complex and rich semantic information.

Pictures are similar to iconic gestures in the sense that both are in the visual modality, but they are also different from iconic gestures. Pictures are conventionalised ways of representing referents, and they can bear full representational power on their own. In contrast, iconic gestures are non-conventionalised, and more "sketchy" compared to pictures, as they symbolically represent fewer aspects of the referent. For instance, a drawing would have to include a cat in its entirety or at the very least the cat's face or body to be identified as a picture of a cat. In addition, one does not need to be presented with the word "cat" when one sees a cat drawing to understand what it depicts. When a speaker uses an iconic gesture representing a cat, on the other hand, they might depict only the ears (e.g. tracing two triangle shapes at the top of one's head) or the whiskers of the cat (e.g. tracing lines at one's mouth) or they could just depict the action of petting (e.g. back and forth motion of the hand). If we just see one of those iconic gestures (such as petting for example) without hearing the co-occurring speech, we would not necessarily understand that it depicted a cat. Thus, the meaning of an iconic gesture is disambiguated with the meaning of concurrent speech (McNeill, 1992). In fact, it was found that adults find it quite difficult to understand what iconic gestures represent in the absence of speech (Krauss et al., 1991). Furthermore, gestures present information through movement dynamics rather than in a static manner. Given the difference between pictures and gestures, it might be expected that it is harder for children to integrate online semantic information in word-gesture pairs than in word-picture pairs. As mentioned above, behavioral studies using offline tasks have shown that by 5 years old, children can integrate a word and iconic gesture (Sekine et al., 2015) by assessing their responses to pictures, but we know nothing about the online integration of this process. Thus, the current study examined whether children's brains show evidence for online integration of information from both modalities simultaneously presented with an iconic gesture and a word.

1.5. Present study

In the current study, we examined whether Dutch-speaking 6- and 7-year-olds can integrate simultaneously presented iconic gestures and action words. Even though gestures can be integrated with speech at the word, sentence and discourse levels as shown in behavioral and neural studies in adults, this is the first ERP study on speech-gesture integration in children. In order to determine whether comparable effects between children and adults could be found, we examined children's brain activity using ERPs, employing an N400 paradigm successfully used with adults to measure the integration of single words and iconic gestures (Drijvers & Özyürek, 2018).

We investigated children in the age range of 6–7 years because, first of all, we wanted to make sure we had enough action verbs and gestures that were comprehensible to children in order to have enough items in each condition. Secondly, it has been shown that a domain-general processing shift occurs during development (Ramscar & Gitcho, 2007), that is "a shift from behavioral responses driven by a single factor to those that integrate or select between multiple ones" beginning after age 3 (Ramscar & Gitcho, 2007, p. 274). In fact, behavioral studies found that by the time children become 5 years old, they can integrate gesture and speech in an offline task (e.g. Sekine et al., 2015; Stanfield et al., 2013). If this ability also generalises to online integration then, children aged 6–7 years should process gesture and speech in the same way as adults integrating both modalities simultaneously as soon as they are perceived (e.g. Drijvers & Özyürek, 2018). In this case we expect to observe a larger N400 to the mismatching as compared to the matching speech-gesture stimuli. This would indicate that this simultaneous integration, even though cognitively more effortful, is possible by children. In contrast, if children integrate gesture and speech only after they have processed the information separately in each modality, and simultaneous online integration ability continues to develop until 8–11 years old as found by some subjects in Demir-Lira et al.'s (2018) study, their brain activity would be different from adults, and we would expect not to see an N400 effect. Furthermore, if the integration recruits different brain areas in children vs. adults, a more distributed topography of the N400 effect would be expected, as found in previous fMRI studies.

2. Method

2.1. Participants

Twenty-three native Dutch-speaking children with a mean age of 87 months (7:03 year olds) (SD = 4.76, Range = 80–94 months, 11 female) participated in the study. All were right-handed and reported no developmental issues. We recruited participants by contacting local schools and libraries in Nijmegen, The Netherlands. The analyses reported below are based on a final sample of 15 children (see section 2.6 EEG data acquisition and analysis for details on participant exclusion).

2.2. Materials

2.2.1. Video stimuli

Verb list. The list of Dutch action verbs used in the present study was originally based on a list created by Drijvers and Özyürek (2017). We selected 170 out of 190 verbs, based on the criteria that 80% of 5- and 6-year-old Dutch children are familiar with these verbs (Schaerlaekens et al., 1999).

Gesture list and speech-gesture combinations. For each verb, a native female Dutch speaker produced a gesture with simultaneous speech in both the matching and the mismatching conditions. We instructed the actor, wearing neutral coloured clothing standing in front of a neutral-coloured background (see Figure 1) facing the camera placed in front of her, to create the gestures spontaneously. We made sure the gestures were iconic and representative of the action the verbs described (e.g. typing gesture resembling fingers typing on a keyboard for the verb "*to type*"). In the mismatch condition the actor combined a verb with a mismatching gesture. The videos displayed the actor from head to knees, her hands hanging casually to the side of her body.

To ensure that children (a) understood the gestures and (b) could relate them to the relevant verbs, we conducted a pre-test at two elementary schools in the Netherlands. We tested 104 children (M_{age} = 6.74, SD = 0.64) who did not participate in the subsequent ERP experiment. The details of the pre-test are described in the Supplementary Materials.

Based on the pre-test ratings, we created the final verb/gesture list for the EEG trials by selecting 126 items (120 for experimental and 6 for practice trials). The final set of videos was trimmed from the beginning by using Adobe Premier Pro and ELAN (Lausberg & Sloetjes, 2009) so that the times from the video onset until the onset of a gesture, and from the end of gesture until the end of video, were similar. No further editing in the whole gesture or speech segment was done. All videos were a total of 2300 ms long. The preparation of the gesture in the videos always started at 120 ms after video onset, and the gesture ended before the end of the video clip. For the matching condition, the average onset of gesture stroke (the meaningful part of the gesture) was 669 ms (SD = 97.6) and the

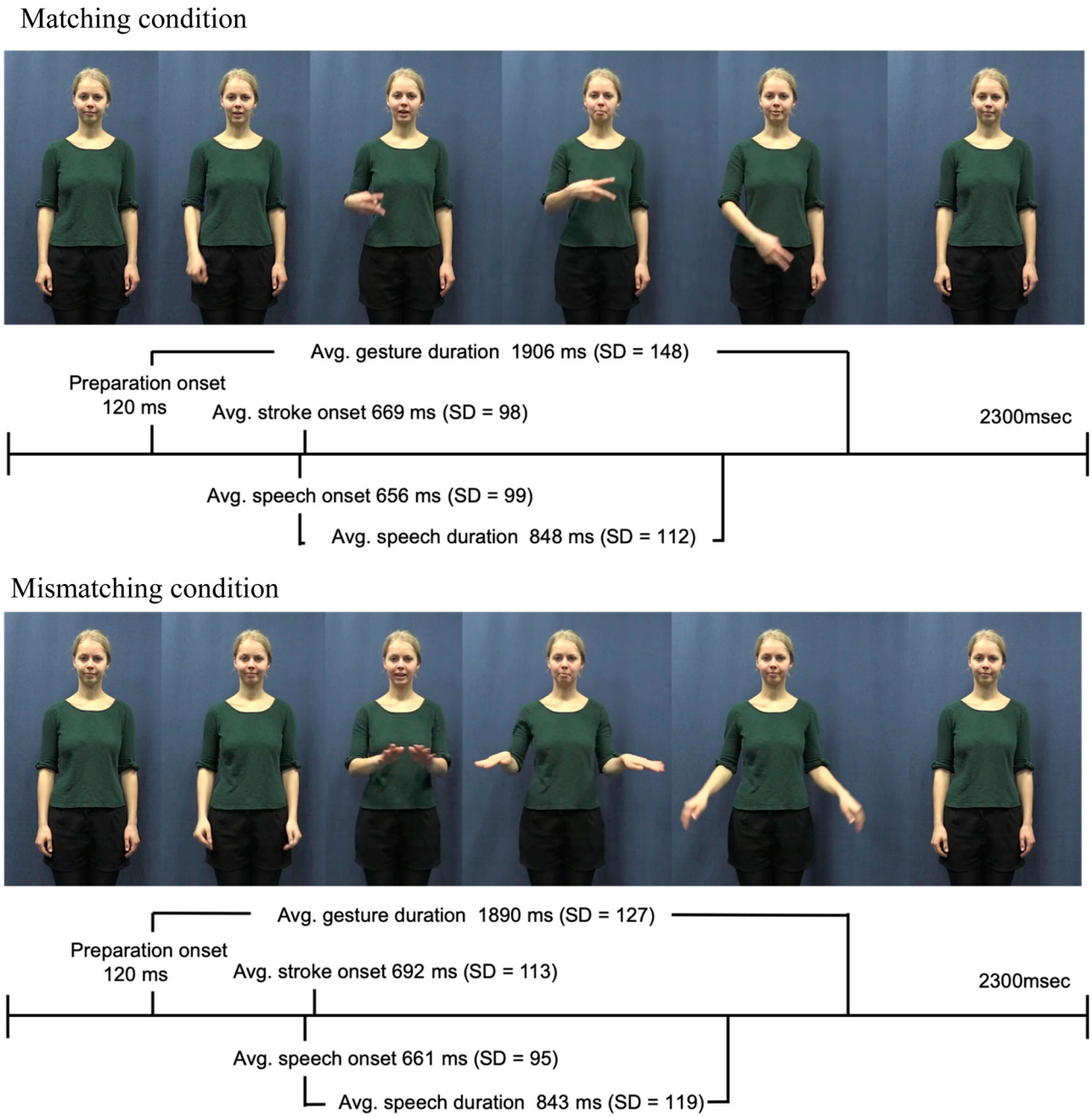

Figure 1. Time-line of the video clip exemplifying a matching speech-gesture combination (top panel) and a mismatching speech-gesture combination (bottom panel). The actor produced "knippen" ("to cut") in speech with a cutting gesture for the matching condition and with a swimming gesture for the mismatching condition.

average speech onset was 656 ms (SD = 99.0) after the start of the video. The mean fundamental frequency (F0) of speech was 231.9 Hz (SD = 9.99), the average speech duration was 848 ms (SD = 112), and the average gesture duration from the preparation to retraction phase was 1906 ms (SD = 148). For the mismatching condition, the average onset of gesture stroke was 692 ms (SD = 113) and the average speech onset was 661 ms (SD = 95.4) after the start of the video, the mean fundamental frequency of speech was 230.9 Hz (SD = 11.26), the average speech duration was 843 ms (SD = 119), and the average gesture duration from the preparation to retraction phase was 1890 ms (SD = 127). There was no significant difference in the average onset of gesture stroke, $t(125) = -1.70$, $p = .092$, speech onset, $t(125) = -.56$, $p = .580$, mean fundamental frequency, $t(125) = 1.79$, $p = .076$, speech duration, $t(125) = 1.32$, $p = .189$, or gesture duration, $t(125) = 1.07$, $p = .288$, between the matching and mismatching trials. Figure 1 shows the average gesture preparation and

stroke onset, gesture duration, speech onset, and speech duration for each condition. As shown in Figure 1, onset of speech and gesture stroke started very close to each other in time.

2.3. Design

Each child received 6 practice trials. For the rest of the 120 experimental trials, 60 verbs were presented in the matching condition, and the other 60 verbs were in the mismatching condition. Each of the 120 video clips was presented only once, and all combinations were counter-balanced to ensure that no gesture would occur twice (either in the matching or the mismatching condition). Each trial started with a fixation-cross (500 ms), followed by a grey transition screen (500 ms, for baseline measure). The video clip was played (2300 ms), and after a short delay period (1000 ms) a fixation-cross appeared again on the screen. In all trials, we measured EEG while children were watching video clips. In addition, one third (40) of the trials also included a behavioral task in order to make sure children were attending to the video stimuli.

In the videos that included the behavioral task, 1000 ms after the video ended (and EEG was recorded), children were asked whether they heard a word in the previous video, for example, like "Hoorde je *'boksen'*?" ("Did you hear *'to box'*?") with a black screen. The verb in the auditory question differed on every trial. Children had to respond to questions by pressing a "Ja" (*Yes*) or "Nee" (*No*) button. Children were instructed to press the "yes" button, if they thought they did hear the word mentioned (e.g. "*boksen"*), and the "no" button if they thought they did not hear the word. Out of 40 behavioral trials, half of them were *related questions* where the word that was asked indeed appeared in the previous video, and the other half were *unrelated questions* where the word that was asked about did not appear in the previous video. The target verb in unrelated questions did not occur in the video stimuli list. If children failed to respond to the question, the following trial started after a 5000 ms delay. We presented behavioral trials at randomised positions throughout the experiment. Each child received three "behavioral trials" in the practice phase before the start of the experiment.

2.4. Procedure

The participant's parent filled out a consent from, the Edinburgh Inventory of Handedness (Olfield, 1971), and a general demographics information sheet. The child sat in front of a mirror, so they could see themselves while we fitted the EEG cap (actiCap, Brain Products, Gilching, Germany). After the impedance check, we walked the child into an electrically and acoustically shielded room to sit in front of a computer monitor, which was 60 cm away. We asked the child to hold the two-button box like a game-controller so that the child could press the corresponding button with their left or right thumb while watching short video clips of a women. We also explained that sometimes they would hear a question, which they would have to answer with either yes or no by pressing the corresponding buttons on the button box. We presented the video stimuli on the monitor using Presentation software (Version 19.0, Neurobehavioral Systems, inc.). Behavioral trials were presented randomly throughout the experiment and occurred after the video clip was played. The order of the video and behavioral trials was pseudo-randomised and presented in four blocks of 40 trials, lasting around 4 min per block. Each block consisted of 30 video clips (15 from each matching and mismatching condition) and 10 behavioral trials (five yes- and five no-responses). After each block, a student assistant entered the room, and the child took a break. During the break, the child played with a maze. The EEG recording procedure, including the breaks, lasted around 40 min. After completion of the experiment, children received stickers and a certificate of participation.

2.5. Behavioral data analysis

We analyzed the behavioral data with RStudio (RStudio Core Team, 2015). No participants were excluded based on low accuracy scores (mean accuracy 98.48%, range 92.31%–100.00%). We removed non-responses from the dataset and identified outliers in all reaction times (RTs). We removed those data points that fell above or below two and a half standard deviations from the grand mean. We analyzed our log-transformed (normalised) RT data with linear mixed effects models with *participants* and *items* as cross-random effects. For the analyses, we considered the following factorial predictors in the fixed effects structure of the model: congruency of the speech-gesture video (*speech-gesture congruency*; 2 levels: matching or mismatching conditions), congruency of the behavioral trial (*audio relation*; 2 levels: related or unrelated question). In addition, we considered random slopes of these predictors by participant and by item.

We performed a stepwise variable selection procedure to obtain the best fitting model. We added one predictor at a time. For each significant predictor or interaction, it was evaluated whether inclusion of this predictor or interaction resulted in a better model (i.e. had a lower AIC compared to when this predictor was not

part of the model). The final model contained the following predictors in the fixed effects part of the model: *trial, speech-gesture congruency*, and *audio relations*nd contained random intercepts for participant and item in the random part of the model.

2.6. EEG data acquisition and analysis

We recorded the EEG continuously throughout the experiment from 32-AG-agCl electrodes. Twenty-seven electrodes were mounted in a cap according to the 10–20 international system, four electrodes were used for bipolar horizontal and vertical electrooculograms (EOG) and one electrode was placed on the right mastoid. The reference electrode was placed on the left mastoid and re-referenced offline to the average of the left and right mastoid electrodes. Electrode impedance was kept below 5 KΩ. The EEG was filtered through a 0.02–100 Hz band-pass filter and digitised on-line with a sampling frequency of 500 Hz (BrainVision Recorder, Brain Products, Gilching, Germany).

We pre-processed the EEG data with using the Fieldtrip toolbox (Oostenveld et al., 2011) running under MATLAB (Mathworks, 2018). First, we re-referenced the EEG data offline to the average of the left and right mastoid and filtered the data with a high-pass filter at 0.01 Hz and a low-pass filter at 35 Hz. We further segmented the data into epochs from 200 ms before to 1900 ms after the onset of the videos. We applied a baseline correction of 200 ms before the onset of the video.

We removed artefacts in three steps. First, we removed trials with noise that were not related to eye-movements from this analysis. Rejection of trials was based on the minimal and maximum amplitude (in our case, amplitude should not exceed 70 and −70 microvolts, respectively), whether the maximal difference of values in the segment exceeded a certain value (in our case, 70 microvolt), and the variance within each channel. Next, we corrected for eye-movement artefacts using an ocular independent component analysis (ICA). In the ICA, we decomposed the data in independent components and removed the components that represented the eye artefacts. These steps were done by Fieldtrip toolbox running under MATLAB. Lastly, we again rejected trials with any remaining artefacts by a semi-automatic artefact rejection routine. The rate of artefacts was rather high in our child population. As a criterion for inclusion in our sample of trials to be analysed, we decided to accept participants with minimally 20 analysable trials per condition after artefact rejection. This resulted in the exclusion of 7 participants from the total sample of 23. A further participant was excluded due to hardware malfunction, leaving 15 participants whose trials were analysed. For those 15, the average number of trials was 28.53 in the matching condition, and 28.67 in the mismatching condition (i.e. 47.6% of their total number of trials on average).

To evaluate the differences between the matching and the mismatching condition we used a non-parametric cluster-based permutation test (Maris & Oostenveld, 2007) by using the Fieldtrip toolbox and MATLAB. The calculation of this cluster-based test statistic was based on the following steps: For every data point (a combination of channel and time), the experimental conditions were compared by means of a t-value. Then, all samples were selected whose t-value was larger than 0.025. These selected samples were then clustered in connected sets on the basis of temporal and spatial adjacency. Cluster-level statistics were calculated by taking the sum of the t-values within every cluster. The significance probability was calculated by means of the Monte Carlo permutation. To calculate this, a participant's time-locked average was randomly assigned (5000 times) to one of the two conditions to calculate the largest cluster-level statistic for every permutation. The highest cluster-level statistic from each randomised calculation was entered into the Monte-Carlo permutation distribution, and cluster-level statistics were calculated for the data. The statistics were then compared against this permutation distribution. Only those clusters that fell into the highest or lowest 2.5th percentile of the distribution were considered significant (see Maris & Oostenveld, 2007).

3. Results

3.1. Behavioral results

As mentioned in "EEG data acquisition and analysis" section, we excluded 8 participants from EEG analyses. These 8 participants were excluded from the behavioral analyses as well. The analysis of the behavioral trials as a check for attention to the task showed that children did pay attention to content of the videos. Accuracy means ranged from 92% to 100% (Table 1), suggesting that children performed near ceiling. Because there was minimal variation in our data with respect to accuracy, the linear mixed effects model could not be fitted (due to convergence errors), and we used a two-way

Table 1. Mean and standard deviation (in parentheses) of accuracy scores.

		Audio congruency (behavioral trials)	
		Match	Mismatch
Speech-gesture congruency	Matching	0.986 (0.117)	0.980 (0.140)
	Mismatching	0.993 (0.083)	0.993 (0.085)

Table 2. Mean and standard deviation (in parentheses) of RT scores.

		Audio relation (behavioral trials)	
		Related	Unrelated
Speech-gesture congruency	Matching	1239.15 (319.89)	1405.99 (336.19)
	Mismatching	1244.71 (294.60)	1405.94 (350.15)

repeated measure analysis of variance (ANOVA) to measure significance. We did not observe a significant main effect in the accuracy scores (*speech-gesture congruency* F (1, 14) = 1.214, $p = 0.28$; *audio relation,* F (1,14) = 0.128, $p = 0.73$) and no significant interaction (*speech-gesture congruency * audio relation.* F (1, 14) = 0.110, $p = 0.75$).

We observed a main effect of RTs in *audio relation,* F (1, 14) = 40.81, $p < .001$ (Table 2). These results show that children were significantly faster in responding to words that matched the speech of the previous video clip regardless of the congruency between gesture and speech. There was no significant interaction between *audio relation* and *speech-gesture congruency* (this interaction was removed from the final model), indicating that children were also able to allocate their attention to one modality when asked to do so. The final model of RTs revealed a significant effect of *Trial,* showing that children's RTs became faster during the experiment. In addition, the model suggested that items showed a different sensitivity in response to *audio relation* (Table 3).

3.2. EEG results

We included the whole time window (0 ms at video onset to 1900 ms) for the EEG analysis. The ERPs of both conditions were time-locked to the video onset, and we compared the data of both conditions for the 15 participants. The cluster-based permutation test revealed a significant negative cluster in the time windows from 1040 ms to 1194 ms after video onset and showed a significant difference in the mean amplitude between the gesture-speech matching and the mismatching conditions ($p = 0.0166$). Given that the average of speech onset in video stimuli was around 660 ms after the video onset, this test indicates that the significant effect occurred around 400 ms after speech onset. This result shows that 6-7-year-olds' online integration of semantic information from speech and gestures is similar to that of adults.

The grand average ERPs of matching and mismatching stimuli on electrode Pz are plotted in Figure 2A. This figure shows that the N400 was larger for the mismatching stimuli compared to the matching stimuli. The electrodes showing significance based on the cluster-based permutation test were C4, CP5, CP1, CP2, P7, P3, Pz (see Figure 2B)

Table 3. Summary of the model predicting RTs.

Predictor	β	Standard error	t
Intercept	7.060	0.035	**203.29**
Trial	−0.081	0.013	**−6.47**
Speech-gesture congruency	0.007	0.017	0.43
Audio relation	0.128	0.021	**6.09**
Random effects	Var.	Standard deviation	
Item (intercept)	0.009	0.096	
Participant (intercept)	0.014	0.120	
Residual	0.038	0.195	

Note: $t < -1.96$ and $t > 1.96$ is significant, printed in bold. For speech-gesture congruency we used *matching condition* as the reference in the intercept; for audio relation we put *related question* in the intercept.

4. Discussion

Children frequently observe meaningful gestures along with the speech addressed to them. Our understanding of the processing of these multimodal messages by children is still limited to a few studies. As stated in the Introduction, previous studies using offline behavioural measures with children between 3 and 5 years have shown a developmental trend in comprehending iconic gestures in the context of short sentences (e.g. Sekine et al., 2015; Stanfield et al., 2013). Furthermore, the few fMRI studies with older children (8–11 years) have found that some but not all children integrate iconic gestures with speech in longer stretches of discourse behaviourally and neurally (e.g. Demir-Lira et al., 2018; Dick et al., 2012). However, nothing is known about the online process involved in how children integrate information from both modalities. Furthermore, although previous EEG research reported that online picture-word integration could be observed from infancy (e.g. Friedrich & Friederici, 2004, 2010), it still remains an open question as to whether this generalises to speech and gesture integration, as an iconic gesture is more ambiguous and dependent on co-occurring speech compared to a picture. Thus, we examined 6- and 7-year-olds' online integration abilities of simultaneously presented speech and gesture utterances using EEG measures, focusing on the N400 component–as previously used in adults.

Consistent with the results from studies with adults (e.g. Drijvers & Özyürek, 2018; Habets et al., 2011; Kelly et al., 2004; Özyürek, 2014; Özyürek et al., 2007; Wu & Coulson, 2007), the current study also found a larger N400 component in the gesture-speech mismatching condition than in the matching condition (see Figure 2A). ERPs of children and adults may not directly be comparable, since young children's brains are not fully developed and differences in neural density or myelination may affect the brain activity in different ways (DeBoer

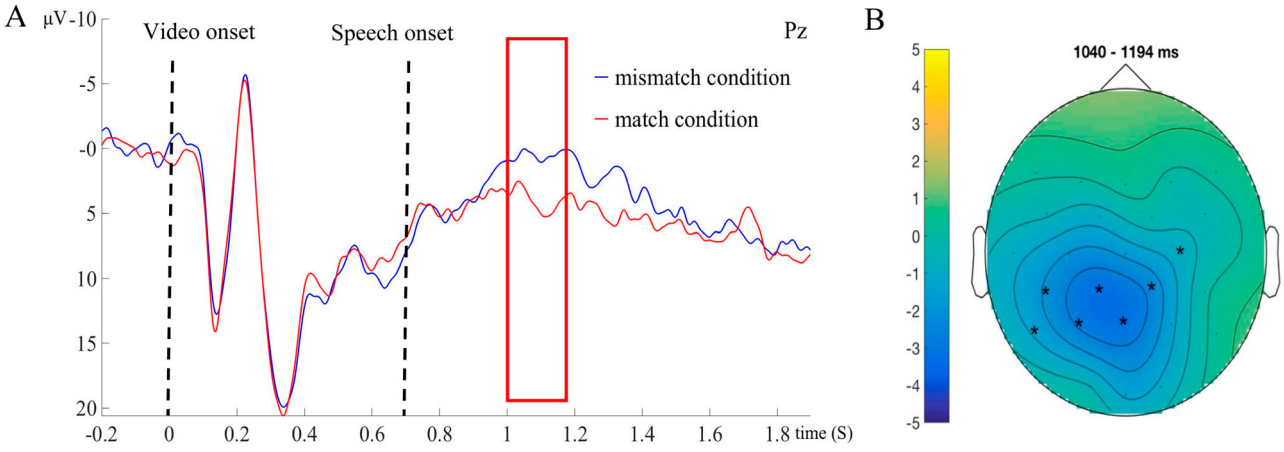

Figure 2. Figure A (left panel) shows the grand average waveforms for ERPs elicited in the match (red) and mismatch (blue) condition at electrode Pz. Negativity is plotted upward. The red squire box indicates the time windows where a significant negative cluster was found. Figure B (right panel) shows the topo plot of the significant effect at 400 ms after speech onset.

et al., 2005). However, similar brain waveforms for ERPs and topographical plots of N400 effects were observed between children (in the current study) and the previous adult study using very similar materials (see Figure 3 in Drijvers & Özyürek, 2018). This implies that the neural basis for online speech-gesture integration appears to be already in place for 6- and 7-year-old children and that children in this age period can process both types of information as soon as they are available, as adults do, rather than in a sequential manner. This suggests that at this age period, integrating semantic information from different modalities poses no greater challenge to children than to adults (at least at the one word level), which is in line with the idea that by this age, children have successfully mastered the developmental "shift" from unimodal to multimodal processing (Ramscar & Gitcho, 2007).

Comparing findings from previous behaviour and brain studies and the current one, we can infer the following developmental path for gesture-speech integration: From around 3 years old, children gradually start to develop the ability to integrate iconic gesture with simple spoken utterances (e.g. Sekine et al., 2015; Stanfield et al., 2013). By the age of 6 and 7 years, they can integrate simultaneously presented gesture and speech in an online fashion similar to adults in the context of single words and gestures. However, children continue to develop their ability to integrate gestures within larger spoken sentences and narratives until the age of 11–12 years, behaviourally and neurally, as shown by Demir-Lira et al. (2018).

Previous studies on unimodal (auditory) semantic integration in children aged between 5–18 years have shown an N400 effect in response to words or pictures that mismatch the preceding context in linguistic priming tasks (e.g. Benau et al., 2011; Holcomb et al., 1992; Pijnacker et al., 2017). Our finding goes beyond unimodal integration by showing that semantic integration does not only occur during information processing of the auditory modality, but is also apparent when information is communicated simultaneously through different modalities. As the N400 effect was observed in the same time window as in previous unimodal linguistic semantic priming studies, we believe that linguistic semantic integration and multimodal integration are strongly related and are similar processes.

The observed effect in our study suggests that childrens' brains have a bias to integrate information from speech and gesture simultaneously at a relatively early age--in line with results found for simultaneous word and picture integration (e.g. Friedrich & Friederici, 2004, 2010). It also seems like the modality specific nature of gesture is not more taxing than pictures for children. Despite the attention-controlling behavioural task in our study focusing children's attention on speech, but not necessarily on gesture, the ERP findings show that children could not help but consider both the visual and auditory modalities when processing a multimodal message online, as indicated by the N400 effect (see Kelly et al., 2010 for a similar effect with adults).

Thus, the current study supports the claim that gestures have the potential to greatly contribute to language comprehension not only in adults, but also in children (Sekine et al., 2015). Based on our findings, we can conclude that a neural basis for online speech-gesture integration appears to already be in place at the age of 6–7 years at the single gesture-word level. The results of this study, however, need to be extended to other ERP studies at the sentence or discourse level and also to children who are younger than 6, critically

at the age periods of 3 and 5 where behavioral studies show a developmental effect. It is also necessary to explore whether individual differences in linguistic and cognitive abilities affect the integration between gestures and speech in children.

Finally, the findings of the current study have theoretical implication for multimodal integration at the semantic level. Research has shown that gesture and speech form an integrated system in both production and comprehension in adults (e.g. Kelly et al., 2010; McNeill, 1992). The current study supports the argument by adding new neural evidence that gesture and speech already form an integrated system in comprehension by 6–7 years of age. The findings also have practical implications for using gestures with children in noisy environments, as well as with children with hearing impairments or other cognitive deficits, and can provide a baseline for future studies.

Note

1. In this paper, we used terms "matching'" and "mismatching" to indicate cases where gesture and speech refer to similar or very different referents, in line with other N400 studies used in spoken language comprehension (e.g. Drijvers & Özyürek, 2017, 2018). Please note that other gesture studies have also used these terms but in a differently way from the current study. Goldin-Meadow and her colleagues (e.g. Goldin-Meadow, 2003) used the terms "matching" and "mismatching" to indicate whether gesture and speech semantically represent same or different aspects of the *same* referent, respectively.

Acknowledgements

This study was supported by the European Commission, Marie Curie Actions (H2020-MSCA-IF-2015) given to the first author. We would like to thank our actor for her patience in creating the video stimuli and our student assistants who helped during data collection. We further want to express gratitude to Nick Wood at the Max Planck Institute for Psycholinguistics, who has since passed away, for technical assistance and support in editing our stimuli. We also thank Linda Drijvers who offered her assistance and knowledge.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This study was supported by the European Commission, Marie Curie Actions [H2020-MSCA-IF-2015] given to the first author.

ORCID

Kazuki Sekine http://orcid.org/0000-0002-5061-1657

References

Beattie, G., & Shovelton, H. (1999a). Do iconic hand gestures really contribute anything to the semantic information conveyed by speech? An experimental investigation. *Semiotica*, *123*(1/2), 1–30. https://doi.org/10.1515/semi.1999.123.1-2.1

Beattie, G., & Shovelton, H. (1999b). Mapping the range of information contained in the iconic hand gestures that accompany spontaneous speech. *Journal of Language and Social Psychology*, *18*(4), 438–462. https://doi.org/10.1177/0261927X99018004005

Benau, E. M., Morris, J., & Couperus, J. W. (2011). Semantic processing in children and adults: Incongruity and the N400. *Journal of Psycholinguistic Research*, *40*(3), 225–239. https://doi.org/10.1007/s10936-011-9167-1

Broaders, S. C., & Goldin-Meadow, S. (2010). Truth is at hand: How gesture adds information during investigative interviews. *Psychological Science*, *21*(5), 623–628. https://doi.org/10.1177/0956797610366082

Campisi, E., & Özyürek, A. (2013). Iconicity as a communicative strategy: Recipient design in multimodal demonstrations for adults and children. *Journal of Pragmatics*, *47*(1), 14–27. https://doi.org/10.1016/j.pragma.2012.12.007

Chui, K. (2005). Temporal patterning of speech and iconic gestures in conversational discourse. *Journal of Pragmatics*, *37*(6), 871–887. https://doi.org/10.1016/j.pragma.2004.10.016

DeBoer, T., Scott, L. S., & Nelson, C. A. (2005). Event-related potentials in developmental populations. In H. Todd (Ed.), *Methodological handbook for research using event-related-potentials* (pp. 263–297). TheMIT Press.

Demir-Lira, ÖE, Asaridou, S. S., Raja Beharelle, A., Holt, A. E., Goldin-Meadow, S., & Small, S. L. (2018). Functional neuroanatomy of gesture-speech integration in children varies with individual differences in gesture processing. *Developmental Science*, *21*(5), Article e12648.

Dick, A. S., Goldin-Meadow, S., Solodkin, A., & Small, S. L. (2012). Gesture in the developing brain: Brain development of gesture. *Developmental Science*, *15*(2), 165–180. https://doi.org/10.1111/j.1467-7687.2011.01100.x

Drijvers, L., & Özyürek, A. (2017). Visual context enhanced: The joint contribution of iconic gestures and visible speech to degraded speech comprehension. *Journal of Speech, Language, and Hearing Research*, *60*(1), 212–222. https://doi.org/10.1044/2016_JSLHR-H-16-0101

Drijvers, L., & Özyürek, A. (2018). Native language status of the listener modulates the neural integration of speech and iconic gestures in clear and adverse listening conditions. *Brain and Language*, *177-178*, 7–17. https://doi.org/10.1016/j.bandl.2018.01.003

Friedrich, M., & Friederici, A. D. (2004). N400-like semantic incongruity effect in 19-month-olds: Processing known words in picture contexts. *Journal of Cognitive Neuroscience*, *16*(8), 1465–1477. https://doi.org/10.1162/0898929042304705

Friedrich, M., & Friederici, A. D. (2010). Maturing brain mechanisms and developing behavioral language skills. *Brain & Langauge*, *114*(2), 66–71. https://doi.org/10.1016/j.bandl.2009.07.004

Goldin-Meadow, S. (2003). *Hearing gesture: How our hands help us think*. Harvard University Press.

Gutmann, A. J., & Turnure, J. E. (1979). Mothers' production of hand gestures while communicating with their preschool children under various task conditions. *Developmental*

Psychology, *15*(2), 197–203. https://doi.org/10.1037/0012-1649.15.2.197

Habets, B., Kita, S., Shao, Z., Özyurek, A., & Hagoort, P. (2011). The role of synchrony and ambiguity in speech–gesture integration during comprehension. *Journal of Cognitive Neuroscience*, *23*(8), 1845–1854. https://doi.org/10.1162/jocn.2010.21462

Hagoort, P., & Van Berkum, J. (2007). Beyond the sentence given. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 801–811. https://doi.org/10.1098/rstb.2007.2089

Henderson, L. M., Baseler, H. A., Clarke, P. J., Watson, S., & Snowling, M. J. (2011). The N400 effect in children: Relationships with comprehension, vocabulary and decoding. *Brain and Language*, *117*(2), 88–99. https://doi.org/10.1016/j.bandl.2010.12.003

Holcomb, P. J., Coffey, S. A., & Neville, H. J. (1992). Visual and auditory sentence processing: A developmental analysis using event-related brain potentials. *Developmental Neuropsychology*, *8*(2–3), 203–241. https://doi.org/10.1080/87565649209540525

Holle, H., & Gunter, T. C. (2007). The role of iconic gestures in speech disambiguation: ERP evidence. *Journal of Cognitive Neuroscience*, *19*(7), 1175–1192. https://doi.org/10.1162/jocn.2007.19.7.1175

Holler, J., Schubotz, L., Kelly, S., Hagoort, P., Schuetze, M., & Ozyurek, A. (2014). Social eye gaze modulates processing of speech and co-speech gesture. *Cognition*, *133*(3), 692–697. https://doi.org/10.1016/j.cognition.2014.08.008

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*(2), 73–88. https://doi.org/10.1007/s10919-008-0063-9

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

Kelly, S. D., & Church, R. B. (1998). A comparison between children's and adults' ability to detect conceptual information conveyed through representational gestures. *Child Development*, *69*(1), 85–93. https://doi.org/10.1111/j.1467-8624.1998.tb06135.x

Kelly, S. D., Kravitz, C., & Hopkins, M. (2004). Neural correlates of bimodal speech and gesture comprehension. *Brain and Language*, *89*(1), 253–260. https://doi.org/10.1016/S0093-934X(03)00335-3

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

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

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

Krauss, R. M., Morrel-Samuels, P., & Colasante, C. (1991). Do conversational hand gestures communicate? *Journal of Personality and Social Psychology*, *61*(5), 743–754. https://doi.org/10.1037/0022-3514.61.5.743

Kutas, M, & Federmeier, K.D. (2014). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annual review of psychology*, *62*, 621–647.

Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences*, *4*(12), 463–470. https://doi.org/10.1016/S1364-6613(00)01560-6

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*(5947), 161–163. https://doi.org/10.1038/307161a0

Lausberg, H., & Sloetjes, H. (2009). Coding gestural behavior with the NEUROGES-ELAN system. *Behavior Research Methods*, *41*(3), 841–849. http://tla.mpi.nl/tools/tla-tools/elan/ https://doi.org/10.3758/BRM.41.3.841

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Mathworks. (2018). Natick, Massachusetts: The MathWorks Inc.

McNeil, N. M., Alibali, M. W., & Evans, J. L. (2000). The role of gesture in children's comprehension of spoken language: Now they need it, now they don't. *Journal of Nonverbal Behavior*, *24*(2), 131–150. https://doi.org/10.1023/A:1006657929803

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

Nobe, S. (2000). Where do most spontaneous representational gestures actually occur with respect to speech? In D. McNeill (Ed.), *Language and gesture* (pp. 186–198). Cambridge University Press.

Obermeier, C., Holle, H., & Gunter, T. C. (2011). What iconic gesture fragments reveal about gesture–speech integration: When synchrony is lost, memory can help. *Journal of Cognitive Neuroscience*, *23*(7), 1648–1663. https://doi.org/10.1162/jocn.2010.21498

Olfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh Inventory. *Neuropsychologia*, *9*, 97–113. https://doi.org/10.1016/0028-3932(71)90067-4

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). Fieldtrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, *2011*, 1–9. Article 156869. https://doi.org/10.1155/2011/156869

Özçaliskan, S., & Goldin-Meadow, S. (2011). Is there an iconic gesture spurt at 26 months? In G. G. Stam & M. Ishino (Eds.), *Integrating gestures: The interdisciplinary nature of gesture* (pp. 163–174). John Benjamins.

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

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

Pijnacker, J., Davids, N., Van Weerdenburg, M., Verhoeven, L., Knoors, H., & Van Alphen, P. (2017). Semantic processing of sentences in preschoolers with specific language impairment: Evidence from the N400 effect. *Journal of Speech, Language, and Hearing Research*, *60*(3), 627–639. https://doi.org/10.1044/2016_JSLHR-L-15-0299

Ramscar, M., & Gitcho, N. (2007). Developmental change and the nature of learning in childhood. *Trends in Cognitive Sciences*, *11*(7), 274–279. https://doi.org/10.1016/j.tics.2007.05.007

RStudio Core Team. (2015). *RSudio: Integrated development for R*. RStudio, Inc. http://www.rstudio.com

Schaerlaekens, A. M., Kohnstamm, G. A., Lejaegere, M., & Vries, A. K. (1999). *Streeflijst woordenschat voor zesjarigen: gebaseerd op nieuw onderzoek in Nederland en België*. Swets & Zeitlinger.

Sekine, K., Sowden, H., & Kita, S. (2015). The development of the ability to semantically integrate information in speech and iconic gesture in comprehension. *Cognitive Science*, *39*(8), 1855–1880. https://doi.org/10.1111/cogs.12221

Stanfield, C., Williamson, R., & Özçalışkan, S. (2013). How early do children understand gesture- speech combinations with iconic gestures? *Journal of Child Language*, *41*(2), 1–10. https://doi.org/10.1017/S0305000913000019

Wu, Y. C., & Coulson, S. (2007). How iconic gestures enhance communication: An ERP study. *Brain and Language*, *101*(3), 234–245. https://doi.org/10.1016/j.bandl.2006.12.003