Izvorni znanstveni radovi

UDK 81'221.24:159.95
159.953.35:81'23
Izvorni znanstveni članak
Prihvaćen za tisak: 14. listopada 2024.
https://doi.org/10.22210/suvlin.2024.098.02

Emilija Mustapić Malenica
Department of English Studies
University of Zadar
emustapic@unizd.hr

The (in)congruence effect of co–speech gestures on language processing

The aim of this paper is to explore the phenomenon of co–speech gestures in language processing. As gestures have often been analysed predominantly within the paradigm of the rhetorical tradition, this paper will try to portray their psycholinguistic dimension, which has become increasingly important within the domain of linguistic research. Through use of experimental psycholinguistic methodology and the priming paradigm, I have compared the speed and accuracy of processing of linguistic utterances occurring only in the audio–visual modality accompanied by congruent and incongruent co–speech gestures. The results of the experiment confirmed the hypothesis that incongruence of co–speech gestures has an inhibitory effect on speed and accuracy of processing of the multimodal linguistic message. The participants' reaction times were slower and accuracy rates lower in conditions in which the semantic congruence between the verbal and the gestural modality was nullified, and the same effect was noticed in both the native language and the first foreign language. This result speaks in favour of the claim that the verbal and gestural representational systems are interconnected and constitute a holistic picture of the mental process.

1. Introduction

At its very essence, face–to–face communication is multimodal as speech signals are accompanied by a series of visual articulators, such as facial expression, posture and gestures (Vigliocco, Perniss and Vinson 2014). Gestures have become a prominent topic in the domain of various linguistic branches, such as psycholinguistics, cognitive linguistics, applied linguistics etc. Our comprehension of gestures has a theoretical value for understanding the cognitive and emotional processes, but there are also numerous practical implications for research on gestures, which is why it has become an increasingly interdisciplinary field of study.

Co–speech gestures are one of the most common types of gestures (McNeill 1992, 2005; Kita, van Gijn and van der Hulst 1998). They represent spontaneously

and naturally occurring body movements in a particular communication setting. These unplanned gestures convey a meaning correlated with speech at a semantic, pragmatic and discourse level (Kita, van Gijn and van der Hulst 1998). They can refer to objects and actions of different levels of semiotic complexity and take on various communication roles, while their semantic connection to speech can vary from conveying completely redundant information to expressing information which complements speech (Özyürek 2014). Listeners process the words uttered by the speaker, simultaneously integrating their co–speech gestures to better understand what the speaker is trying to say (Goldin–Meadow 2006; Kendon 1994). According to McNeill (1992), gestures do not belong to the outside world but are instead located in the internal world of memories, thoughts, and mental images. Conceptualizing gestural images does not coincide with, for instance, conceptualization of pictures or photographs in the outside world. Gestural images are complex and interconnected and they open up a new pathway for interpreting thought processes, languages and human interactions. Gestures cannot be exhaustively explained in kinetic terms alone as they are not mere body movements. During gesture production, speakers' hands no longer represent a part of their body but manual symbols which describe the meaning predefined by the speakers themselves. Co–speech gestures can convey a wide array of meanings as they can refer to objects, actions, persons or places, but they are still symbols which largely differ from the verbal language forms as they occur simultaneously with speech. Even though they act in synergy with words and sentences, they are qualitatively different from them and they make up a separate system of symbols that are realized in distinct form and manner in every individual (McNeill 1992).

Co–speech gestures are semiotically divided into four major categories: iconic, metaphoric, deictic and beat gestures. Iconic gestures depict images of concrete objects or activities, i.e. they visually and structurally resemble the entity or action they refer to (McNeill 1992). For example, an iconic co–speech gesture showing the act of climbing while the speaker says "I am climbing" visually represents the speaker's mental image of climbing. Metaphoric gestures, on the other hand, describe abstract concepts and help imagine the unimaginable. For example, if a speaker is holding an object but not presenting it as an object, rather as an idea, the co–speech gesture conveys a metaphorical meaning (McNeill 1992). Deictic or pointing gestures describe spatial relations. A prototypical deictic gesture is usually an extended finger which shows the position of an entity in space. Finally, beat gestures or batons are rhythmic hand movements that follow speech prosody. They differ from iconic and metaphoric gestures as they retain the same shape, regardless of the speech content they refer to (McNeill 1992). In his later work, McNeill (2005) questions the aforementioned classification since none of those categories are actually categorical as none of them occur in isolation. For example, a particular body movement may predominantly refer to an iconic co–speech gesture, but also

contain some other gestural elements (pointing or beat). For that reason, McNeill introduces dimensions rather than categories of gestures (McNeill 2005).

This paper aims to explore the factors which influence language processing when accompanied by the co–speech gestures[1] in a multimodal communication setting through use of psycholinguistic methodology. The paper is structured as follows: in section 2, I provide a brief overview of some theoretical and empirical studies on co–speech gestures, focusing mainly on the experimental research; in section 3, I describe the methodology used in this paper, specifically the participants and materials used, design of the experiment and the data–collection procedure. In section 4, I describe the results of the conducted research and compare them against the results of relevant previous research outlined in section 2, and in section 5, I provide the main conclusions of my research along with suggestions for avenues of further research.

2. Previous theoretical and empirical research

Despite the differences between speech and gestures, the perception of the two phenomena as an integrated mental process implies their close connection which serves to convey the meaning of an utterance (McNeill 1992, 2005). This conception of an integrated system largely differs from the concept of *body speech*[2], a communication process in which signals are bodily movements that are regarded as separate from the language (McNeill 1992: 11–12, 105–109). McNeill (1992, 2005) provides several arguments for the claim that co–speech gestures and speech make up a single cognitive system: semantic and pragmatic co–expressiveness and complementarity of the two modalities; temporal synchrony and the effect of co–speech gestures on speech (more details in Mustapić Malenica 2021). However, the matter whether co–speech gestures really play a significant role in language perception is largely an open question. While one could argue that McNeill's claims are plausible and empirically verifiable, there is a certain degree of disagreement among scholars about his idea of a single integrated system which defines the role of co–speech gestures in communication. Kita (2000) and Krauss, Chen and Gottesman (2000) consider gestures and speech to be the products of two independent representational systems. However, what both sides agree on is that co–speech gestures are an important factor in language processing, which offers a more complex approach to the study of multimodal communication.

The amount of empirical research on co–speech gestures in linguistics conducted in the domain of multimodal communication has thus far been relatively modest, which resulted in several mutually contrasted approaches, three of which have dominated the recent publications: communicative, cognitive and dual approach.

1 Co–speech gestures with a predominantly iconic dimension.

2 According to McNeill (1992: 11), the concept of body speech is a result of a very narrow analysis and should be taken with a grain of salt.

From the communicative perspective, there are several ways in which co–speech gestures realize their role in communication. They primarily allow multimodal representation of the same linguistic message (Kendon 1994; Valenzeno, Alibali and Klatzky 2003), thus facilitating its comprehension (Beattie and Shovelton 1999; Kelly, Özyürek and Maris 2010). They also help the listener in solving semantic ambiguity and enable better understanding of complex semantic information (Graham and Argyle 1975; Rogers 1978; McNeil, Alibali and Evans 2000). Taking into account the embodied nature of co–speech gestures, Kelly, McDevitt and Esch (2009: 314) argue that they play a crucial role in understanding and development of language. In one of the key empirical works in the domain of communicative approaches to gestures, Beattie and Shovelton (1999) experimentally confirmed the claim from Kendon (1980) that co–speech gestures represent one aspect of the target utterance not expressed through the vocal–auditory channel and that combining the two modalities transmits a more complete meaning of the utterance to the listener, thus facilitating its comprehension. Beattie and Shovelton (1999) showed a part of adult participants video recordings of co–speech gestures and their semantic features accompanied by speech (audio–visual condition) or only the audio recordings with no co–speech gestures (audio condition). A sequence of questions was formulated to test different aspects of semantic information from the selected drawn stories, for instance, whether the table was moving in circular motion, whether the hands of the clock are moving etc. Based on the collected data, Beattie and Shovelton (1999) conclude that the participants in the audio–visual condition were in general more accurate in their answers about the semantic features of story in comparison to those participants who could only hear the pertinent content. The participants who could see the gestures received significantly more information about the story than those who did not see the gestures (Beattie and Shovelton 1999).

The advocates of the cognitive approach argue that co–speech gestures have a facilitatory effect on the content that the speaker is trying to produce as they enable easier access to the mental lexicon (Krauss, Morrel–Samuels and Colasante 1991; Morrel–Samuels and Krauss 1992; Hadar et al. 1998; Krauss, Chen and Gottesman 2000). They also claim that research which speaks in favour of the communicative role of gestures is methodologically too deficient to allow any serious conclusions about the facilitatory effect of gestures to be made. Krauss, Morrel–Samuels and Colasante (1991) deduce that co–speech gestures transfer some semantic information correlated with the semantic content of speech, but the amount of information they convey is not enough to enhance their communicative value. They confirm this through the results of two experiments – an experiment with a recognition task and a semantic category assessment task (Krauss, Morrel–Samuels and Colasante 1991). In the first experiment, the participants were expected to solve a series of tasks for recognizing segments shown in three conditions: the auditory condition (speech with no gesture), the visual condition (gesture with no speech)

and the audio–visual condition (gesture + speech). The results showed that adding gestural information to the verbal part of the utterance did not increase the participants' recognition accuracy. In the second experiment, the participants were shown a sequence of gestures they were supposed to divide into four semantic categories (*action*, *location*, *object designation* and *description*). The first group only saw the gesture video without the accompanying speech (visual condition), while the second group saw the gesture but also heard the accompanying speech (audio–visual condition). Both groups were supposed to assess which semantic category the meaning of the displayed gestures refers to. No additional explanations for semantic categories were given and the participants were supposed to use their own criteria in their assessments. Two additional groups could only hear the speech (audio content) or read the transcript and determine the semantic category of lexical content based on it. The results showed that the assessment of semantic category of gestures in the visual condition was significantly different from all other conditions as the participants exposed to visual condition only had the lowest level or accuracy in their assessments. However, it was shown that gesture in the visual condition reveals some information about the semantic category of its lexical affiliate, which confirms its minimal communication value. On the other hand, the presence of co–speech gesture in the audio–visual condition did not contribute to the semantic category assessment in comparison to the audio condition. This implies that when the listeners hear a world accompanied by a co–speech gesture, their understanding of semantic content is largely a product of what they hear, not what they see (Krauss, Morrel–Samuels and Colasante 1991).

From the perspective of dual approach to co–speech gestures, the ability to understand a spoken language is not a fixed trait that an individual has at a certain point, but a dynamic concept which varies depending of the complexity of the task (and the listener's perceptive skills) and the availability of external support (McNeil, Alibali and Evans 2000). McNeil, Alibali and Evans (2000) believe that the role of congruent co–speech gestures in comprehension depends on the features of the message being conveyed – when the message is simple, congruent co–speech gestures do not contribute to comprehension, but they contribute to comprehension when the message becomes more complex. McNeil, Alibali and Evans (2000) base their assumption on the results of research in which the adult speakers in non–interruptive conditions are able to understand the spoken message without external support, which minimizes the role of co–speech gestures in speech perception (cf. Krauss, Morrel–Samuels and Colasante 1991; Morrel–Samuels and Krauss 1992; Krauss, Chen and Gottesman 2000). However, they also take into account the studies in which the congruent co–speech gestures had a crucial role in comprehension, as in inhibiting auditory conditions (Riseborough 1981; Ross et al. 2006; Holle et al. 2010; Drijvers and Özyurek, 2017, 2018; Drijvers, Özyürek and Jensen 2018; Drijvers, Vaitonytė and Özyürek 2019; Schubotz et al. 2020).

3. Research methodology

In this paper, I present the results of empirical research conducted as a part of the unpublished doctoral dissertation (Mustapić Malenica 2021). Considering that the topic of multimodal communication is relatively under–researched from a psycholinguistic perspective, this paper provides an example of an experimental approach which might serve as the basis for similar future research. The aim of the research is focused on two main research questions:

1. Do incongruent co–speech gestures affect the speed of processing of verbal utterances in the native language and the first foreign language and if so, how?
2. Do incongruent co–speech gestures affect the accuracy of processing of verbal utterances in the native language and the first foreign language and if so, how?

With regards to the research questions, the following hypotheses were formulated:

1. Participants will provide more accurate and faster responses when the co–speech gesture is congruent with the verbal utterance than when the utterance contains a semantically incongruent co–speech gesture.
2. Semantically incongruent co–speech gestures will impede and decelerate language processing.

These hypotheses are primarily based on three studies described below. In a study conducted with children aged 17 months, co–speech gestures semantically incongruent with speech were able to obstruct comprehension of the linguistic message when the attention of the listener was distributed between the two modalities (Macnamara 1977). In a situation when the message was articulated only verbally (for example, in the sentence *Show me the shoe*), the children successfully selected the right object out of two possible ones (such as a shoe and a cup). However, when the same verbal message was combined with incongruent co–speech gestures, the children's reaction was based less on verbal and more on non–verbal (gestural) modality which displayed different content (Macnamara 1977).

The second study which particularly served as a methodological basis of this paper was conducted by Kelly, Özyürek and Maris (2010) through use of the priming paradigm. Kelly, Özyürek and Maris (2010) designed experimental tasks which consisted of two parts – every task started with a prime in the form of a video clip, which was followed by the second, target stimulus in the form of an audio recording of a verbal utterance followed by a semantically congruent or incongruent gesture. Both types of stimuli were recorded with the help of actors. For the video clips of stimuli used as primes, the actor conducted real everyday actions (such as chopping vegetables), based on which the actress in the target stimulus was supposed to say the sentence describing the action as spontaneous and natural as possible, while producing a co–speech gesture which accurately described the action in some cases

and inaccurately in others. One half of video clips in the target stimulus of the first experiment showed the action from the prime, while the second half were unrelated information which served as fillers and were not taken into consideration in the analysis. The results of the experiment showed that the strength of overlap of the gestural and the verbal modality is correlated with speed and accuracy of language processing (Kelly, Özyürek and Maris 2010). Kelly, Özyürek and Maris (2010) concluded that congruence of co–speech gestures with speech was correlated with faster reaction time and lower number of errors in comparison to cases where there was certain degree of incongruence between them.

As a direct continuation of the research conducted by Kelly, Özyürek and Maris (2010), Özer and Goksün (2019) examined the connection between verbal and visuo–spatial cognitive sources of processing semantic information expressed jointly in the visual modality (in the form of co–speech gestures) and the auditory modality (through speech). Using the incongruence paradigm which causes increased cognitive load during verbal and visual information processing, Özer and Goksün (2019) argued that the way in which the listeners process multimodal information and the level of benefit they obtain depends on cognitive load that they are "forced upon" (Özer and Goksün 2019). Listeners with higher processing capacity are better at overcoming the cognitive burden than listeners with a lower cognitive capacity (Paas, van Gog and Sweller 2010; Özer and Goksün 2019). For instance, complex task like multimedia learning require the listener to process and integrate multimodal information, which is why multimodal approach to coding information is often believed to enhance learning and memory (Özer and Goksün 2019, more details also in Clark and Paivio 1991). The results of the experiment in Özer and Goksün (2019) showed that incongruence of the target stimulus leads to a higher degree of incorrect answers and slower reaction time in comparison to the control condition. The participants were slower and less accurate in their responses in trials in which the co–speech gestures from target stimulus were incongruent with the action in the prime, in comparison to trials in which the verbal part of the target stimulus was incongruent with the prime. No significant difference was noted between slightly or extremely incongruent co–speech gestures (Özer and Goksün 2019).

In order to compare the speed and accuracy of multimodal language processing, I conducted an experiment with two different conditions: a) co–speech gestures being congruent with the verbal part of the target stimulus; and b) co–speech gestures being semantically incongruent with the verbal part of the target stimulus. The aim of the experiment was to determine whether semantically incongruent gestures facilitate comprehension of language content in the two languages. Tasks with semantically congruent co–speech gestures served as a control condition to determine the possible existence and the effect strength of incongruence of co–speech gestures.

3.1. Participants

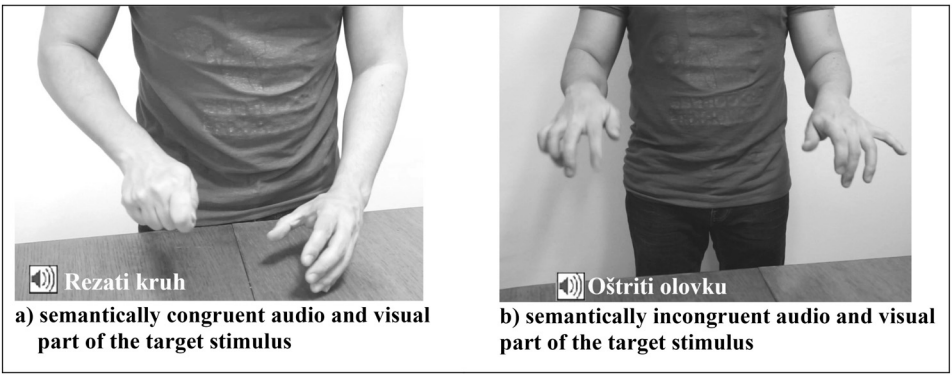
A total of 36 participants (M = 6, F = 30)[3] took part in the experiment, with the mean age of 20,3 (range from 19 to 24), with normal visual and motor skills. Three participants reported technical difficulties when running the experiment, which is why their results were excluded from subsequent analyses. All participants who took part in the experiment were students of the second year of undergraduate study of English language and literature at the Department of English studies at the University of Zadar. Their second majors included fields such as Pedagogy, Sociology, Linguistics, Theology, French, Spanish, German and Russian language and literature. As a form of compensation for their participation, the participants received additional course credit. The demographic data collected via the *Google Forms* questionnaire indicate that all participants were exposed to English in spoken and written form on a daily basis. All participants were native speakers of Croatian and they studied English as L2 for an average of 14.06 years and regarded themselves as experienced (C1) speakers of English. They had reached the high level of general English language proficiency (C1–C2 according to the Common European Framework of Reference for Languages) by passing the courses *Contemporary English Language 1* and *2*, in which they developed their advanced reading, writing, listening and speaking skills in English.

3.2. Materials

In order to control as many linguistic variables that could affect the outcome of the research, the word combinations were not selected randomly but based on several criteria. To create the materials, I used the lexical database of Croatian words designed by Erdeljac, Lendić and Sekulić Sović (2018) which used six psycholinguistic parameters: subjective frequency of words in use, imageability, abstractness/concreteness, age of acquisition, familiarity, and associative connectivity of the word in question. In order to test the frequency of co–occurrence of verbs and nouns used in the trials, a corpus analysis was conducted. The measuring of co–occurrence of verbs and nouns in Croatian was done using the *Croatian Web Corpus* (hrWaC) 2.2 corpus (Ljubešić and Klubička 2014), which contains about 1,400,000,000 tokens, while the *British National Corpus* (BNC) with approximately

3 An anonymous reviewer has asked why the sample was unbalanced in terms of gender of participants. An overview of previous empirical research (in §2) and theoretical background in (§2) did not identify gender as a variable that may affect processing in this domain, which is why this variable was not controlled for. It is also worth pointing out that while potential participants were students, an important requirement of the study was that the L2 proficiency of participants was at least at the intermediate to advanced level. This requirement could in principle be satisfied by drawing participants from a pool for which this proficiency level could be assumed (the method undertaken in the study), or by administering an L2 proficiency test, which would pose an additional burden in terms of choice of instrument, motivation, participant fatigue, and even validating that the results on the test were in fact achieved by a particular participant. Thus, not only is absolute balance in terms of gender not necessitated by previous research, achieving it would provide an unnecessary obstacle for conducting the research itself.

110,000,000 tokens was used for English. In the final version of the experiment, only the word combinations with a high level of imageability and concreteness and high frequency of use were included. The corpus of actions used in the experiment also consisted of 20 actions in Croatian which were subsequently translated into English, and 20 filler items in both Croatian and English. The audio recordings of verbal utterances, video recordings of real actions and congruent gestures were edited in the *Wondershare Filmora 9* program prior to their implementation in the experiment. To eliminate the possibility of facial expressions affecting the results, the actor's face was not visible in any of the analysed stimuli, but only in a handful of filler items. Video recordings of real actions were used as primes and audio files played simultaneously with congruent or incongruent gestures were the target stimuli.[4] Every audio–visual recording lasted 2 seconds. In every task for which reaction time (RT) was measured, the audio of the target stimulus was congruent with the prime, which the participants had to recognize as fast as possible. However, the gestural part of the stimulus was semantically congruent with the audio in one half of the stimuli and semantically incongruent in the other half (Figure 1). Fillers contained completely incongruent actions between primes and targets as well as within the target stimuli. The experiment consisted of 86 trials, 6 practice trials, 40 trials used in the analysis (20 trials in English and 20 in Croatian), and 40 filler trials (20 in English and 20 in Croatian).



a) semantically congruent audio and visual part of the target stimulus

b) semantically incongruent audio and visual part of the target stimulus

Figure 1. Examples of target stimuli with a) semantically congruent and b) semantically incongruent co–speech gestures (Mustapić Malenica 2021: 145)

4 An anonymous reviewer has pointed out that the fact that the study was conducted with video recordings can significantly affect the presentation as the "load" of multimodality may not be the same in video and live situations. While I partly agree with the comment in the sense that processing of co–speech gestures seen in live situations and co–speech gestures in video recordings is not identical (though comparable), this kind of argument about the ecological validity of an experiment can be applied to most (if not all) experimental paradigms. In order for the data from all participants to be comparable and generalizable, all participants need to be exposed to the same set of stimuli presented in an absolutely identical manner. This level of comparability can only be guaranteed via video and audio recordings, which makes this compromise in terms of ecological validity acceptable, though inevitable. However, I believe such a trade–off is justified by the comparability of the results, not just from this study, but from numerous other studies in this and similar lines of research.

3.3. Design of the experiment

As the aim of the experiment was to test the (in)congruence effect of the two modalities of the simultaneously shown target stimuli on speed and accuracy of language processing in the prime, the dependent variables were reaction time and response accuracy, while the independent variables were language (Croatian and English) and congruence of the stimuli (congruent and incongruent). The experiment was designed and conducted in a controlled environment, using the priming paradigm.

Every trial started with the fixation cross (+) being shown for 500 ms, followed by a video recording of the real action as the prime in the duration of 2000 ms. This was followed by a target audio–visual stimulus in the duration of 2000 ms in two experimental conditions: a) the actor read the sentence describing the real action from the prime and simultaneously produced a congruent co–speech gesture (Figure 2), and b) the actor read the sentence describing the real action from the prime and simultaneously produced an incongruent co–speech gesture (Figure 3).[5] In one half of trials, the target stimulus was played in Croatian and in English in the other half of trials. After the target stimulus, the answers DA ("Yes") and NE ("No") were shown on screen and the participants were prompted to answer the question "Does the action you saw in the first screen match with the audio recording you heard on the second screen?". So as not to overburden the participants with redundant information, only the answers were shown on the screen, while the question was stated at the beginning of the experiment and the participants were instructed that they are supposed to answer it every time they see the prompts for answers.

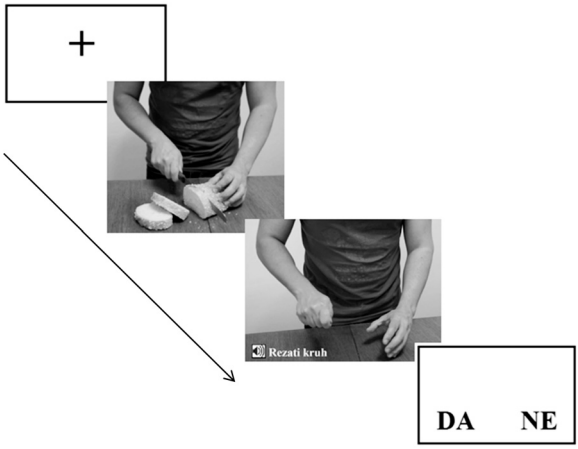


Figure 2. Schematic overview of an experiment trial with a congruent co–speech gesture (Mustapić Malenica 2021: 147)

5 It should be emphasized that experimental conditions set this way are the necessary starting point for further research of co–speech gestures and language processing. Based on the empirical results obtained in this experiment, one could further test the accuracy and reaction time in cases when the co–speech gesture in the target stimulus is congruent and the audio recording is incongruent with the real action from the prime.

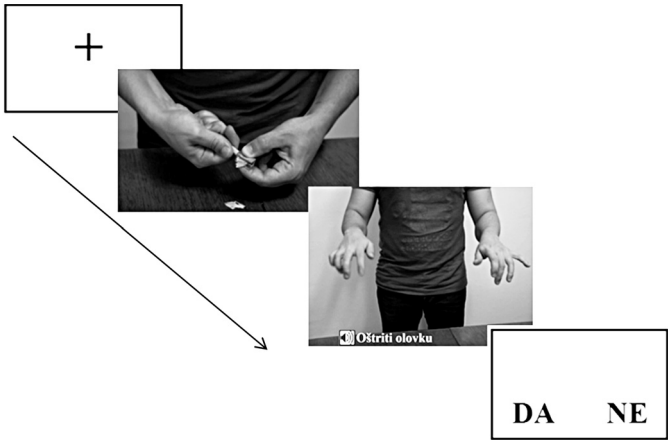


Figure 3. Schematic overview of an experiment trial with an incongruent co–speech gesture (Mustapić Malenica 2021: 147)

In all trials, the audio of the target stimulus was congruent with the real action, so the correct response from the participants in those tasks was to press the Yes button. In the filler tasks, no part of the target stimulus matched the prime so the participants were supposed to answer No to get a correct answer. This was done to minimize the response bias, i.e. to prevent the participants to provide faster responses by falling into a routine of multiple consecutive positive answers. Reaction time was measured from the moment the last screen with the YES and NO buttons appeared until the participant provided their response by pressing one of the keys.

Every target stimulus occurred in four possible combinations (English/Croatian, Congruent/Incongruent); hence, the participants were randomized into two groups so that all shown actions would be maximally balanced. Three participants from group A reported technical difficulties and were excluded from subsequent analysis, which ultimately led to 15 participants in group A and 18 participants in group B. In every group, two versions of the target stimulus were shown for one action in the prime stimulus. For instance, one trial in group A included the recording of cutting bread followed by the audio–visual target stimulus with audio in Croatian and co–speech gesture being congruent with the spoken content. In the second instance of the trial in the same group, the audio–visual target was shown but with audio content in English and an incongruent co–speech gesture. The order of presentation of conditions was balanced across the whole experiment so that every action which served as a prime was shown in two conditions in each group. The implementation of tasks and conditions into a complete experiment was conducted by using the *E–Prime* 3.0 software (Schneider, Eschman and Zuccolotto 2012). The table with all stimuli, practice trials, and initial and final instructions were entered into *E–Prime* and the trials were generated based on this. No tasks were repeated within the same group and the order of their presentation was randomized in *E–Prime*. Out of 20 trials in Croatian, 10 had a co–speech gesture congruent with the

content in the target stimulus and 10 had an incongruent gesture (the same ratio applied to 20 trials in English).

3.4. Procedure

Since the experiment was conducted during the COVID–19 pandemic restrictions, the *E–Prime Go 1.0* module of the *E–Prime 3.0* package (Schneider, Eschman and Zuccolotto 2012) was used for the procedure, which enabled a remote data collection. Before the main experiment was conducted, the trial version of the experiment was piloted with 6 participants who did not take part in the main research. This was done to test the clarity of instructions for installing the *E–Prime Go* module, possible technical difficulties while using the *E–Prime Go*, clarity of instructions for remote data collection, and average time needed for going through the whole procedure. Before running the experiment, the participants filled out a *Google Forms* questionnaire with sociodemographic information and questions about their use of the English language. They received e–mail instructions about securing the necessary conditions for uninterrupted participation in the experiment, technical instructions for installing and using *E–Prime Go 1.0* and procedures for filling out all research components (*Google Forms* questionnaire, installing the application and running the experiment).

The entire experimental procedure took around 30 minutes to complete, including a few minutes for filling out the sociodemographic questionnaire, 2–3 minutes for installing the *E–Prime Go* app and 15–20 minutes for going through the experiment. All participants provided their digital consent with participation by selecting the 'Pristajem ('I accept')' key at the beginning of the experiment. They were also informed about the anonymity of use regarding their data for research purposes and the possibility of reviewing their research results. In order to ensure anonymity, every participant used their own unique ID code. Before the first 6 practice trials started, the participants were informed to hold the index finger of their left hand over the A key for YES and the index finger of their right hand over the K key for NO throughout the entire experiment. Theye were instructed to respond to the question as fast as possible only after the white screen with YES/NO answers appears. If necessary, the participants could redo practice trials multiple times. After going through the practice trials and a short break, the participants would start the main part of the experiment. Once the main trials were started, the participants could not pause or go to previous screens and were therefore instructed not to press any other keys apart from the two keys for answering. When a participant completed the experiment, the final screen showed a thank you message and the notification about the end of the experiment.

4. Results and discussion

The results of 33 participants who completed the experiment without experiencing any difficulties were collected via the *E–Prime Go* module and merged in a joint database using the *E–Merge* module. The data were then exported to R (*R Core Team* 2015) for further statistical analysis. Every participant provided 40 datapoints, giving a grand total of 1320 observations. Before analysing the effects of congruence and stimulus language on reaction time, the effect of these variables on distribution of correct and incorrect answers was assessed (Table 1 and Table 2). A chi–squared test showed no statistically significant difference in terms of correct answers between stimuli in Croatian and English ($\chi2(1) = 0{,}232$, $p = .63$), and a statistically significant difference between sentences with congruent and incongruent gestures ($\chi2\ (1) = 45{,}47$, $df = 1$, $p < .001$). Table 2 shows that a larger proportion of incorrect answers, i.e. a lower degree of accuracy was noted for tasks with incongruent gestures. Somewhat expectedly, tasks for which the participants provided correct answers had slower reaction times than tasks with incorrect answers.

	English	%	Croatian	%
Correct answers	626 (623.5)	94.85	621 (623.5)	94.1
Incorrect answers	34 (36.5)	5.15	39 (36.5)	5.9

Table 1. Ratio of correct and incorrect responses across stimuli language (Mustapić Malenica 2021: 151)

	Congruent	%	Incongruent	%
Correct answers	652 (623.5)	98.79	595 (623.5)	90.15
Incorrect answers	8 (36.5)	1.21	65 (36.5)	9.85

Table 2. Ratio of correct and incorrect responses across stimuli types (Mustapić Malenica 2021: 151)

Prior to analysing the effects of congruence and language on reaction time, the collected values for reaction time were filtered so as to remove the observations with incorrect responses and outliers. The accuracy rate at the level of the whole group was generally very high (94.47 %), but one participant had an accuracy rate of 52.5 %. Since this level of accuracy could be expected from random pressing of keys on the keyboard, the observations collected from this participant (N=40) were removed from further analysis. The exclusion of observations was also applied to all observations with incorrect responses (N=54), all observations with reaction time higher than 5000 ms (N=2), reaction time lower than 50 ms (N=5), and all observations 2 standard deviations from the mean (N=16). The remaining 1203 observations were used in the analysis below.

The average reaction time of these filtered values was then calculated across congruence (Table 3) and language of the stimuli (Table 4). The data in the tables show that there is a certain difference in reaction time between congruent and in-

congruent stimuli, and between stimuli in Croatian and English. The statistical significance of these effects was tested using a random effects regression model (Bates et al. 2015). As can be seen from the histogram (Figure 4), the values were normally distributed and could be used in the regression model.

	Congruent	Incongruent
M	428.04	460.1
SD	294.53	386.19

Table 3. Reaction time across congruence of stimuli (Mustapić Malenica 2021: 152)

	CRO	ENG
M	430.49	456.24
SD	327.28	355.44

Table 4. Reaction time across language of stimuli (Mustapić Malenica 2021: 152)

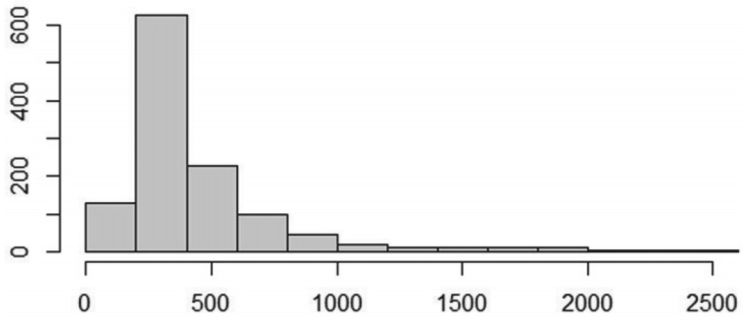


Figure 4. Histogram of reaction time values collected in the experiment (Mustapić Malenica 2021: 152)

Using the *lme4* package (Bates et al. 2015), four random effect models were created; the dependent variable was reaction time and participants and trials were entered as random effects in every model. The models were as follows:

i) The main model with congruence and language of the stimuli as fixed factors and individual participants and trials as random effects (Reaction time ~ Congruence + Language of stimulus + (1 | Participant) + (1 | Task));
ii) The model with language of the stimuli as the only fixed factor (Reaction time ~ Language of stimulus + (1 | Participant) + (1 | Task));
iii) The model with congruence of the stimuli as the only fixed factor (Reaction time ~ Congruence + (1 | Participant) + (1 | Task));
iv) The model with no fixed factors (Reaction time ~ (1 | Participant) + (1 | Task)).

A statistically significant difference ($\chi2(2) = 7.05$. $p < .05$) was noted between the model with both fixed factors (i) and the model without any fixed factors (iv), indicating the effect of at least one of the two variables on reaction time. Furthermore, a statistically significant difference ($\chi2(1) = 4.83$. $p < .05$) was noticed between the main model (i) and the model in which congruence of the stimuli was not used as a fixed factor (ii), which indicates that congruence has a significant effect on reaction time. In figure 5, it can be seen that this difference is manifested in faster reaction times for congruent stimuli, in comparison to incongruent ones. Conversely, the comparison of the main model (i) and the model in which language of the stimuli was not used as a fixed factor (iii) showed no statistically significant difference between the models ($\chi2(1) = 2.19$. $p = .139$), which shows that language of the stimuli did not have an effect on reaction time.

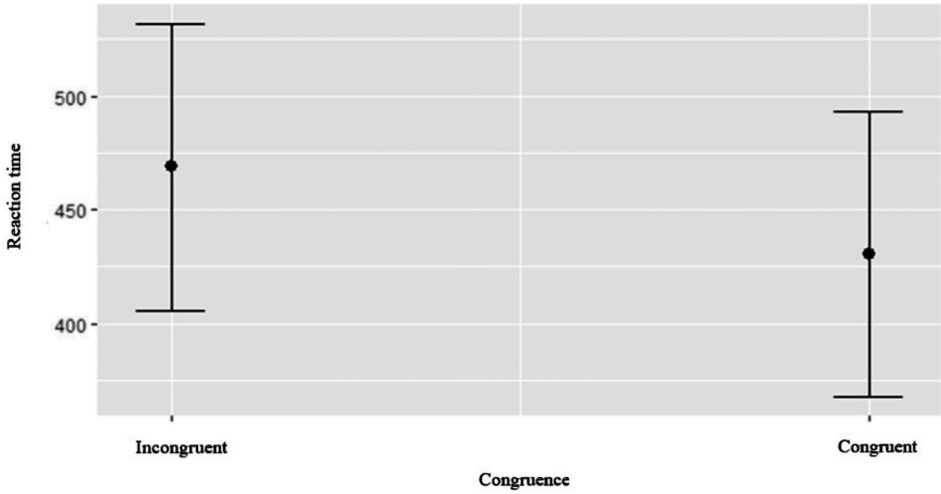


Figure 5. The effect of congruence on reaction time (Mustapić Malenica 2021: 153)

Since all visual stimuli had a recording of the action which was showed twice, it was necessary to determine whether the order of trials affects reaction time, i.e. whether the trials shown at the beginning of the experiment have faster reaction time values than tasks shown closer to the end of the experiment. To test this, another group of linear regression models with random effects was created, this time with ordinal position of the trial (from 1 to 80) and order of appearance of the action (first or second) as the fixed factors and individual tasks and trials as random factors. Three regression models were created:

i) The main model with order of appearance of the action and the ordinal position of the trial in the experiment (Reaction time ~ Order of appearance + Ordinal position + (1 | Participant) + (1 | Task));
ii) The model with order of appearance as the only fixed factor (Reaction time ~ Order of appearance + (1 | Participant) + (1 | Task));

iii) The model with ordinal position as the only fixed factor (Reaction time ~ Ordinal position + (1 | Participant) + (1 | Task)).

The comparison of models revealed a statistically significant difference ($\chi2(1) = 11.25$. $p < .001$) between the main model (i) and the model with order of appearance as the only fixed factor and no statistically significant difference ($\chi2(1 = 1.9$. $p = .164$) between the main model (i) and the model with ordinal position as the only fixed factor (iii). This indicates that the second exposure to same video of the real action did not result in faster reaction times. However, higher ordinal position of a particular trial (i.e. being shown towards the end of the experiment) did result in faster reaction times (Figure 6). The overall contribution of this effect was nullified through randomization of the trials as every combination of stimuli had an equal chance of appearing closer to the start or the end of the experiment.

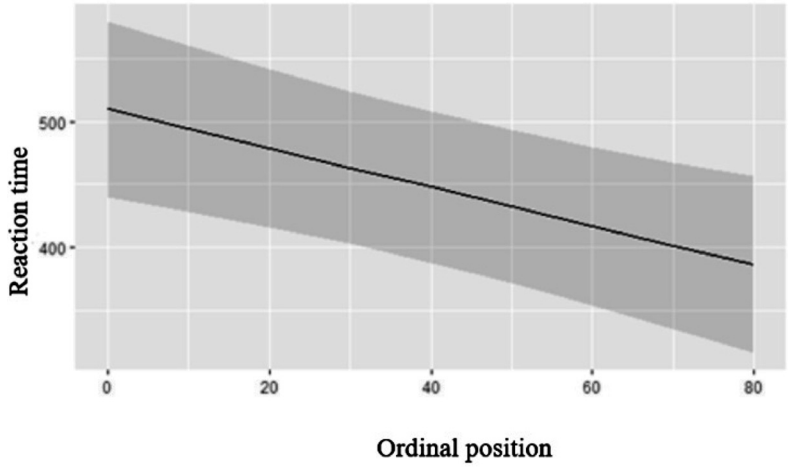


Figure 6. Influence of ordinal position of the trial on reaction time (Mustapić Malenica 2021: 155)

The results of the experiment show a larger proportion of incorrect answers and slower reaction time in both languages in trials with incongruent co–speech gestures in comparison to the trials with congruent co–speech gestures. As expected, reaction time was slower and accuracy lower in the incongruent co–speech gesture condition, which is in line with the previous results in Kelly, Özyürek and Maris (2010). The results presented here are also compatible with the results of older research, such as Macnamara (1977), which established that incongruent co–speech gestures can inhibit comprehension of the linguistic message as the listener is focused on processing of two modalities (see §2). Finally, the results presented here are also in line with the results of Özer and Goksün (2019) who established higher rates of errors and slower information processing in conditions in which co–speech gestures in target conditions were incongruent with the action in the prime, compared with the condition with congruent gestures. This convergence of the results is also noteworthy considering the fact that Özer and Goksün (2019) analysed

the data of participants with different cognitive abilities, while the group used in this experiment was homogenous in this respect and included advanced learners of a foreign language. Even though the participants with a higher visuo–spatial work memory capacity were more successful in solving tasks with incongruent modalities, Özer and Goksün (2019) noticed an inhibitory effect of incongruence with both participant profiles, which is in line with the results presented here. If we take into consideration the claim in McNeil (1992) that co–speech gestures represent mental concepts of information and their tight connection with the meaning of the verbal part of the utterance with which they create a complete mental image, it is evident that the obstructed semantic connection between speech and gesture in incongruent conditions leads to slower reaction time and lower accuracy, even with advanced speakers of a foreign language.

The results of the experiment presented in this paper provide empirical support in favour of one integral, multimodal system of information processing. However, the fact that incongruent co–speech gestures had an inhibitory effect on language reception is not sufficient to support the communicative approach in a more general sense. If the role of co–speech gestures were indeed a communicative one, which could be argued based on the results of this experiment, their effect should be visible not only when they are incongruent with speech but also in congruent conditions. Given the postulates of the communicative approach (§2), one would expect a facilitatory effect when co–speech gestures are congruent with speech, but this does not seem to be the case (cf. Mustapić Malenica 2024). On the other hand, if co–speech gestures only had a cognitive function, there should be no effect on language reception at all. Assuming that co–speech gestures behave in a more dynamic, dual way seems to be a viable compromise between these two positions. However, this assumption should be tested in further research by including additional dimensions of gestures, other profiles of speakers, different levels of task complexities and external support.

5. Conclusion

The aim of this paper was to empirically test the accuracy and speed of processing of multimodal linguistic information when accompanied by congruent and incongruent co–speech gestures. The results were compared across Croatian as the native language and English as the first foreign language to determine whether there is a difference in effect of gestural (in)congruence on processing in the two types of languages. Data analysis has shown no statistically significant difference in processing of multimodal information in the native language and the first foreign language with advanced speakers of English as L2. Even though the participants' reaction times were slightly faster in Croatian, the difference between the two languages was statistically negligible to make any strong conclusions. On the other hand, congruence of co–speech gestures had a statistically significant effect in

terms of accuracy and speed of processing of the linguistic message. In line with expectations based on previous research, lower proportion of correct responses was recorded in tasks in which the co–speech gesture was incongruent with the verbal part of the utterance. Slower reaction times were also noted in trials with correct responses, in comparison to trials with incorrect responses. Faster reaction times were also noticed in tasks with congruent co–speech gestures, while incongruence resulted not only in a higher number of incorrect responses, but also in slower reaction times with correct answers in both languages.

The results presented in this paper are in line with those of Kelly, Özyürek and Maris (2010) who have empirically demonstrated how incongruent co–speech gestures have a negative effect on reaction time and accuracy during language processing. These results are also in accordance with the research conducted by Özer and Goksün (2019) who maintain that incongruence between the two modalities increases cognitive load during verbal and visual information processing. Even though congruent co–speech gestures did not inhibit language processing with advanced speakers, incongruent gestures had a negative effect on speed and accuracy of processing of the multimodal message. This result speaks in favour of the hypothesis of interconnectedness between the two representational systems.

Through application of experimental research methods, the goal of the paper was to make a small contribution to a better understanding of linguistic perception of the multimodal message. Since the field of gestural studies has been relatively new in (psycho)linguistics, there are many research questions that still have to be addressed and potentially re–evaluated. Co–speech gestures have proven to have a significant role in language processing, but taking the empirical facts into consideration, it may be assumed that their function primarily depends on the features of the multimodal message. Still, that claim should be further questioned by testing how dynamic co–speech gestures can actually be. The results obtained in this study definitely open up some new research ideas in that direction. For instance, a further step in the analysis could be to include other profiles of participants (low–proficiency speakers of a foreign language, participants with language disorders, participants with hearing impairment, people of different age groups etc.) or to incorporate some other, more complex tasks. Also, it would be interesting to empirically test other gestural types (more in Mustapić Malenica 2021) used, for example, in adverse listening conditions and compare their role in language processing. Considering the limitations in the literature covering this topic and the need for further re–examination of provided conclusions, I believe this paper can serve as a solid starting point for similar future research. More comprehensive and profound understanding of role of gestures in communication and determining their speech–related features would certainly open up new pathways of their application in fields such as applied linguistics and clinical linguistics.

References

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker (2015). Fitting Linear Mixed–Effects Models Using lme4. *Journal of Statistical Software*, 67(1). doi:10.18637/jss.v067.i01

Beattie, Geoffrey, and Heather Shovelton (1999). Do iconic hand gestures really contribute anything to the semantic information conveyed by speech? An experimental investigation. *Semiotica*, 123(1–2), 1–30. doi:10.1515/semi.1999.123.1–2.1

Clark, Jim, and Allan Paivio (1991). Dual coding theory and education. *Educational Psychology Review*, 3(3), 149–210. doi:10.1007/bf01320076

Drijvers, Linda, and Asli Özyürek (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. doi:10.1044/2016_jslhr–h–16–0101

Drijvers, Linda, and Asli Özyürek (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. doi:10.1016/j.bandl.2018.01.003.

Drijvers, Linda, Asli Özyürek, and Ole Jensen (2018). Hearing and seeing meaning in noise: Alpha, beta, and gamma oscillations predict gestural enhancement of degraded speech comprehension. *Human Brain Mapping*, 39(5), 2075–2087. doi:10.1002/hbm.23987

Drijvers, Linda, Julija Vaitonytė, and Asli Özyürek (2019). Degree of language experience modulates visual attention to visible speech and iconic gestures during clear and degraded speech comprehension. *Cognitive Science*, 43(10). doi:10.1111/cogs.12789

Erdeljac, Vlasta, Anabela Lendić, and Martina Sekulić Sović (2018). Prikaz baze psiholingvističkih parametara riječi u hrvatskom jeziku, XXXII International scientific conference, JEZIK I UM, Rijeka, 3–5 May, 2018

Goldin–Meadow, Susan, David McNeill, and Jenny Singleton (1996). Silence is liberating: Removing the handcuffs on grammatical expression in the manual modality. *Psychological Review*, 103(1), 34–55. doi:10.1037/0033–295x.103.1.34

Goldin–Meadow, Susan (2006). Talking and thinking with our hands. *Current Directions in Psychological Science*, 15(1), 34–39. doi:10.1111/j.0963–7214.2006.00402.x

Graham, Jean Ann, and Michael Argyle (1975). A cross–cultural study of the communication of extra–verbal meaning by gestures. *International Journal of Psychology*, 10, 57–67.

Hadar, Uri, Dafna Wenkert–Olenik, Robert Krauss, and Nachum Soroker (1998). Gesture and the processing of speech: Neuropsychological evidence. *Brain and Language*, 62(1), 107–126. doi:10.1006/brln.1997.1890

Holle, Henning, Jonas Obleser, Shirley–Ann Rueschemeyer, and Thomas C. Gunter (2010). Integration of iconic gestures and speech in left superior temporal areas boosts speech comprehension under adverse listening conditions. *NeuroImag*e, 49(1), 875–884. doi:10.1016/j.neuroimage.2009.08.058

Kelly, Spencer D., Tara McDevitt, and Megan Esch (2009). Brief training with co–speech gesture lends a hand to word learning in a foreign language. *Language and Cognitive Processes*, 24(2), 313–334. doi.org/10.1080/01690960802365567

Kelly, Spencer D., Asli Özyürek, and Eric Maris (2010). Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension, *Psychological Science*, 21(2), 260–267.

Kendon, Adam (1980). Gesticulation and speech: Two aspects of the process of utterance. In: Key, M.R. (ed.) *Nonverbal communication and language*. The Hague: Mouton, 207–227.

Kendon, Adam (1994). Do gestures communicate? A review. *Research on Language and Social Interaction*, 27(3), 175–200. doi:10.1207/s15327973rlsi2703_2

Kita, Sotaro, Ingeborg van Gijn, and Harry van der Hulst (1998). Movement phases in signs and co–speech gestures, and their transcription by human coders. In: Wachsmuth, Ipke and Martin Fröhlich (ed.) *Gesture and Sign Language in Human–Computer Interaction*. Lecture Notes in Computer Science, vol 1371. Springer, Berlin, Heidelberg. doi:10.1007/BFb0052986

Kita, Sotaro (2000). How representational gestures help speaking. In: McNeill, David (ed.) *Language and Gesture. Language Culture and Cognition*. Cambridge: Cambridge University Press,162–185.

Krauss, Robert, Palmer Morrel–Samuels, and Christina Colasante (1991). Do conversational hand gestures communicate? *Journal of Personality and Social Psychology*, 61, 743–754.

Krauss, Robert, Yihsiu Chen, and Riki Gottesman (2000). Lexical gestures and lexical access: A process model. In: McNeill, David (ed.) *Language and Gesture. Language Culture and Cognition*. Cambridge: Cambridge University Press, 261–283.

Ljubešić, Nikola, Filip Klubička (2014). Proceedings of the 9th Web as Corpus Workshop (WaC–9) @ EACL 2014. In: Bildhauer, Felix and Schäfer Roland (ed.), *Proceedings of the 9th Web as Corpus Workshop (WaC–9) @ EACL 2014*, 29–35.

Macnamara, John (1977). From sign to language. In: Macnamara, John (ed.), *Language learning and thought*. New York: Academic Press, 11–35.

McNeil, Nicole M., Martha W. Alibali, and Julia L. Evans (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. doi:10.1023/a:1006657929803

McNeill, David (1992). *Hand and Mind*. The University of Chicago Press, Chicago and London.

McNeill, David (2005). *Gesture and thought*. The University of Chicago Press, Chicago and London.

Morrel–Samuels, Palmer, and Robert Krauss (1992). Word familiarity predicts temporal asynchrony of hand gestures and speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(3), 615–622. doi:10.1037/0278–7393.18.3.615

Mustapić Malenica, Emilija (2021). *Uloga koverbalnih gesta u jezičnoj recepciji*. Unpublished doctoral dissertation. University of Zagreb, Zagreb.

Mustapić Malenica, Emilija (2024). Multimodality of communication – Speech and co–speech gestures. *Govor*, 41 (1), 3–33. doi:10.22210/govor.2024.41.01

Özer, Demet, Tilbe Göksun (2019). Visual–spatial and verbal abilities differentially affect processing of gestural vs. spoken expressions. *Language. Cognition and Neuroscience*, 35(7), 896–914, doi: 10.1080/23273798.2019.1703016

Özyürek, Asli (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): 20130296. doi:10.1098/rstb.2013.0296

Paas, Fred, Tamara van Gog, and John Sweller (2010). Cognitive load theory: New conceptualizations, specifications, and integrated research perspectives. *Educational Psychology Review*, 22(2), 115–121. doi:10.1007/s10648–010–9133–8

Riseborough, Margaret G. (1981). Physiographic gestures as decoding facilitators: Three experiments exploring a neglected facet of communication. *Journal of Nonverbal Behaviour*, 5, 172–183.

Rogers, William T. (1978). The contribution of kinesic illustrators toward the comprehension of verbal behaviours within utterances. *Human Communication Research*, 5, 54–62.

Ross, Lars, Dave Saint–Amour, Victoria Leavitt, Daniel Javitt, and John Foxe (2006). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex*, 17(5), 1147–1153. doi:10.1093/cercor/bhl024

Schneider, Walter, Amy Eschman, and Anthony Zuccolotto (2012). *E–Prime User's Guide*. Pittsburgh: Psychology Software Tools, Incorporated.

Schubotz, Louise, John Holler, Linda Drijvers, and Asli Özyürek (2020). Aging and working memory modulate the ability to benefit from visible speech and iconic gestures during speech–in–noise comprehension. *Psychological Research*. doi:10.1007/s00426–020–01363–8

Valenzeno, Laura, Marta Alibali, and Roberta Klatzky (2003). Teachers' gestures facilitate students' learning: A lesson in symmetry. *Contemporary Educational Psychology*, 28, 187–204.

Vigliocco, Gabriella, Pamela Perniss, and David Vinson (2014). Language as a multimodal phenomenon: implications for language learning, processing and evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1651), 20130292. doi:10.1098/rstb.2013.0292

Electronic sources

Croatian Web Corpus (hrWaC)

http://nlp.ffzg.hr/resources/corpora/hrwac/

British National Corpus (BNC)

https://www.english–corpora.org/bnc/

Efekt (ne)podudarnosti koverbalnih gesta na jezično procesiranje

Cilj ovog rada bio je istražiti fenomen koverbalnih gestâ u jezičnom procesiranju multimodalne poruke. Budući da se geste i dalje nerijetko promatraju većinom u okviru retoričke tradicije, nastojali smo predstaviti i ispitati njihovu drugu, psiholingvističku dimenziju koja poprima sve veću važnost unutar polja lingvističkih istraživanja. Koverbalne geste definiraju se kao spontani pokreti tijela koji se sujavljaju s govorom te su međusobno povezani na semantičkoj, pragmatičkoj i diskursnoj razini. Njihova uloga u jezičnoj produkciji i recepciji postala je predmetom niza znanstvenih rasprava, što je posljedično rezultiralo i pojavom triju teorijskih stajališta (komunikacijskog, kognitivnog i dvojnog), koje kratko predstavljamo u prvom dijelu rada.

U drugom dijelu rada, primjenom eksperimentalnog istraživačkog pristupa i paradigme usmjeravanja, usporedili smo brzinu i točnost procesiranja jezičnog iskaza koji se javljao isključivo audio–vizualno, s podudarnom i nepodudarnom koverbalnom gestom. Rezultati eksperimenta potvrdili su hipotezu kako nepodudarnost koverbalnih gestâ ima negativan učinak na brzinu i točnost procesiranja multimodalne jezične poruke. Ispitanici su sporije odgovarali i imali veći broj netočnih odgovora kada je semantička podudarnost između verbalnog i gestovnog modaliteta bila narušena, a isti efekt zabilježen je u materinskom i prvom stranom jeziku. Takav rezultat potvrđuje tezu o isprepletenosti dvaju reprezentacijskih sustava, verbalnog i gestovnog, koji tvore cjelovitu sliku mentalnog procesa.

Ključne riječi: (ne)podudarne koverbalne geste, multimodalnost, jezično procesiranje, mentalni procesi
Key words: (in)congruent co–speech gestures, multimodality, language processing, mental processing