



REGULAR ARTICLE



The effect of input sensory modality on the multimodal encoding of motion events

Ezgi Mamus ^{a,b}, Laura J. Speed^a, Aslı Özyürek ^{a,b,c} and Asifa Majid^d

^aCentre for Language Studies, Radboud University, Nijmegen, The Netherlands; ^bMax Planck Institute for Psycholinguistics, Nijmegen, The Netherlands; ^cDonders Centre for Brain, Cognition, and Behaviour, Nijmegen, The Netherlands; ^dDepartment of Experimental Psychology, University of Oxford, Oxford, UK

ABSTRACT

Each sensory modality has different affordances: vision has higher spatial acuity than audition, whereas audition has better temporal acuity. This may have consequences for the encoding of events and its subsequent multimodal language production—an issue that has received relatively little attention to date. In this study, we compared motion events presented as audio-only, visual-only, or multimodal (visual + audio) input and measured speech and co-speech gesture depicting PATH and MANNER of motion in Turkish. Input modality affected speech production. Speakers with audio-only input produced more PATH descriptions and fewer MANNER descriptions in speech compared to speakers who received visual input. In contrast, the type and frequency of gestures did not change across conditions. Path-only gestures dominated throughout. Our results suggest that while speech is more susceptible to auditory vs. visual input in encoding aspects of motion events, gesture is less sensitive to such differences.

ARTICLE HISTORY

Received 22 April 2022
Accepted 21 October 2022

KEYWORDS

Motion events; iconic gestures; visual perception; auditory perception; spatial language



Introduction


We usually receive spatial information via multiple channels. For example, while seeing someone walking away, we may also hear the fading sound of footsteps echoing in the corridor. Each sensory modality has different affordances that contribute to our overall experience of an event. At the same time, we can express events in language using different modalities, as in the verbal and manual modalities, each of which has its own channel restrictions. It is possible, therefore, that the expressibility of multisensory events into multimodal language may differ according to the constraints of both input and output channels. To test this, we investigate whether perceiving events through vision or audition influences the way we express spatial events in speech and gesture.

Vision has the unique advantage of providing simultaneous (i.e., holistic) information about features of objects and events in both close and distant space (e.g., Eimer, 2004; Thinus-Blanc & Gaunet, 1997). It is continuously accessible and thus allows perceivers to update information about motion, location, and spatial relations. Like vision, audition is a distant sense, however, it provides better temporal information than

vision across locations. Audition is found to dominate in temporal processing, such as discriminating rhythmic changes (e.g., Recanzone, 2003; Repp & Penel, 2002; Shams et al., 2000; Spence & Squire, 2003), and in contrast to the holistic nature of visual information, auditory information is sequential. Even though audition provides information about objects and events, vision typically dominates over conflicting auditory information in spatial perception (e.g., Alais & Burr, 2004; Howard & Templeton, 1966). Therefore, vision is widely considered the primary source of spatial perception (e.g., Ernst & Bühlhoff, 2004; Welch & Warren, 1980).

It has been claimed that language reflects this asymmetry between vision and audition. Vision appears to have privileged status, especially in languages of Western societies (e.g., Levinson & Majid, 2014; Lynott et al., 2020; Majid et al., 2018; San Roque et al., 2015; Speed & Majid, 2017; Viberg, 1983). This is reflected in the fact that vision-related verbs (e.g., *see*, *look*) are more frequent and numerous than non-vision related verbs (e.g., *smell*, *feel*) in the perceptual lexicons of languages of the world (e.g., Floyd et al., 2018; Lynott et al., 2020; San Roque et al., 2015; Speed & Majid, 2017; Winter et al., 2018). Although we see differences

CONTACT Ezgi Mamus  ezgi.mamus@mpi.nl  Max Planck Institute for Psycholinguistics, Wundtlaan 1, Nijmegen 6525XD, The Netherlands

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/23273798.2022.2141282>

© 2022 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

in the number and frequency of words across the senses, no study has experimentally investigated the role of input modality on the language used to describe events. Moreover, there is little known about its multimodal expression, particularly co-speech gesture.

From first principles, one might speculate the sequential format of speech is best suited to express event information perceived through the auditory modality, while gesture might best express information from the visual modality. Gesture production theories do indeed share an assumption that gesture derives from visuospatial imagery (Sketch Model, de Ruiter, 2000; Postcard Model, de Ruiter, 2007; Gesture as Simulated Action Framework, Hostetter & Alibali, 2008, 2019; Information Packaging Hypothesis, Kita, 2000; Interface Model, Kita & Özyürek, 2003; Lexical Retrieval Hypothesis, Krauss et al., 2000; Growth Point Theory, McNeill, 1992; McNeill & Duncan, 2000), with iconic gestures in particular considered an effective tool to convey visuospatial information (Alibali, 2005; Hostetter & Alibali, 2008, 2019). While there is nothing in these theories precluding the expression of auditory information in gesture, the emphasis on the “visual” has meant there are very few studies that have investigated the spatial affordances derived from non-visual information and expressed through gesture (although see, e.g., Holler et al., 2022).

To be able to address the question of whether input sensory modality affects multimodal language production, it is important to situate this work in the broader study of motion events and language typology. This is important as speakers of different languages package the same spatial experience in different ways focusing on, and conversely omitting, certain event components in speech and gesture. Slobin (1996) proposed that speakers encode aspects of events depending on distinctions in their language. For example, unlike a satellite-framed language such as English, Turkish is considered a verb-framed language, which primarily encodes PATH in the main verb and optionally encodes MANNER in a subordinated verb or adverbial phrases (Talmy, 1985). Turkish speakers use PATH and MANNER in separate clauses (e.g., *koşarak eve girdi* “she entered the house running”, see Table 1), whereas English speakers conflate these in a single clause (e.g., *she ran into the house*) with MANNER as the main verb.

These language-specific patterns in speech are also reflected in co-speech gesture (Kita, 2000; Kita & Özyürek, 2003; Özçalışkan et al., 2016, 2018). Turkish speakers gesture PATH and MANNER separately, whereas English speakers are more likely to produce conflated gestures. In addition, given the focus on PATH in verb-framed languages, Turkish speakers have a tendency

to gesture only about PATH, even in cases where they mention both PATH and MANNER in speech (Özyürek et al., 2005; Ünal et al., 2022; for a similar tendency in Farsi, Mandarin Chinese, and French respectively, see also Akhavan et al., 2017; Chui, 2009; Gullberg et al., 2008). To account for this, Kita and Özyürek (2003) proposed that gesture derives partly from language typology and partly from visuospatial imagery in their interface model.

With respect to our main research question concerning the role of input modality on the expressibility of motion events, most previous studies have relied overwhelmingly on visual stimuli as input (e.g., video-clips, cartoons, line drawings, paintings; Gennari et al., 2002; Gullberg et al., 2008; Papafragou et al., 2002; Slobin et al., 2014; Ter Bekke et al., 2022). A notable exception is the work of Özçalışkan et al. (2016) who examined cross-linguistic differences in motion event descriptions in congenitally blind, sighted, and blindfolded speakers of Turkish and English. In order to elicit descriptions, blind and blindfolded participants explored scenes haptically while sighted speakers explored them visually. Scenes consisted of landmark objects (e.g., toy house, crib), where static dolls in different postures were posed to create the impression of motion (e.g., a girl running into a house). All participants were instructed to describe the scenes and were explicitly encouraged to gesture at the same time. Özçalışkan et al. (2016) found that both blind and sighted speakers (blindfolded or not) of Turkish and English expressed events in speech and co-speech gesture according to the typology of their language. In a follow-up study, Özçalışkan et al. (2018) showed that blind and sighted speakers of Turkish and English do not display typological differences in gesture when produced without speech (i.e., silent gesture), in line with the claim that only co-speech gesture reflects language-specific packaging (Goldin-Meadow et al., 2008).

These findings suggest sensory modality (in this case, visual vs. haptic) does not strongly influence the way speakers express events in speech or co-speech gesture, with language typology playing a more critical role. However, this conclusion may be premature. While Özçalışkan et al. (2016, 2018) developed a clever paradigm to compare people with and without visual access to stimuli, the conditions were not controlled in all respects. People could have spent longer exploring haptic scenes than visual ones, and this could have affected descriptions. Moreover, there was no direct comparison between descriptions of blindfolded and sighted speakers, so it is possible that within language there were differences between visual and haptic conditions. Finally, in both Özçalışkan et al. (2016) and

Özçalışkan et al. (2018) speakers were explicitly asked to gesture while describing events. Encouraging gesturing might affect the encoding of events and possibly increased speakers' gesture frequency (e.g., Cravotta et al., 2019). Therefore, it remains unclear whether sensory modality of input affects the rate and type of spontaneous gesture production.

There is, in fact, evidence that sensory input could affect multimodal language production for spatial scenes (Iverson, 1999; Iverson & Goldin-Meadow, 1997), which in turn could have implications for motion event encoding. Iverson and Goldin-Meadow (1997), for example, compared blind and sighted English speakers during a route description task and found blind children described *PATH* in a more segmented fashion with more landmarks in their speech than sighted children. For example, a blind child described a route description as: "Turn left, walk north, then you'll see the office, then you'll see 106, then 108, then 110, 112, then there's a doorway. Then there's a hall ...", whereas a sighted child said: "when you get near the staircase you turn to the left" (p. 463). Interestingly, when children gave segmented verbal descriptions, regardless of their visual status, they produced fewer gestures. Iverson and Goldin-Meadow (1997) claimed that gesture frequency decreases with segmented speech due to the process of gesture generation. As gestures express an image as "a global whole" (McNeill, 1992), when speech is represented sequentially, it is not as well-suited for gesture. So, while speech might be more suitable for expressing information from non-visual input, gesture might be less well suited to do so.

To summarise, previous studies provide contradictory evidence about whether sensory modality could influence the way information is expressed in speech and gesture (Iverson, 1999; Iverson & Goldin-Meadow, 1997; Özçalışkan et al., 2016, 2018). However, no study has directly varied the input sensory modality of motion events—while also controlling for duration and event type—to test whether it affects speech and gesture.

The present study

We explore the effect of sensory modality of input on multimodal language use by focusing on motion events. Motion events provide a good test bed as there is a large body of previous speech and gesture production studies to build upon (e.g., Akhavan et al., 2017; Brown & Chen, 2013; Chui, 2009; Gennari et al., 2002; Gullberg et al., 2008; Papafragou et al., 2002). Importantly, *PATH* and *MANNER* components of motion events can be perceived from both visual and auditory inputs

(Geangu et al., 2021; Mamus et al., 2019) and each may be differentially mapped to speech and gesture. Focusing on Turkish in particular allows us to situate our results with respect to previous studies in this language (e.g., Aktan-Erciyes et al., 2022; Allen et al., 2007; Kita & Özyürek, 2003; Özçalışkan et al., 2016, 2018; Ter Bekke et al., 2022) which together provide an important corrective to the dominance of English language studies in the literature (cf. Thalmayer et al., 2021).

We compared Turkish speakers' speech and gesture for *PATH* and *MANNER* of motion events that were presented as audio-only, visual-only, or multimodal (visual + audio) input. Our main goal was to compare audio-only to visual-only input. Including a multimodal condition allowed us to examine the additional boost, if any, multiple sources of information provide. In particular, it is interesting to compare the visual-only to the multimodal condition to see if auditory information provides additional spatial information to language production processes.

In speech, there are a number of specific predictions we can make. First, based on the observation that vision dominates in the perceptual lexicons of languages (e.g., San Roque et al., 2015; Winter et al., 2018), it is possible that vision also influences linguistic encoding for motion events. If so, we would predict that participants in the visual conditions (i.e., visual-only and multimodal conditions) would provide more motion event descriptions than participants in the audio-only condition.

In addition, we can make specific predictions about the encoding of *PATH* vs. *MANNER* in speech. With regard to *PATH*, if the previously attested differences in encoding of *PATH* information from non-visual input (i.e., segmented *PATH* descriptions in blind vs. non-blind, Iverson, 1999; Iverson & Goldin-Meadow, 1997) are caused by the sensory modality of input at encoding, we would predict that participants in the audio-only conditions would describe *PATH* of motion in a more segmented fashion than in other conditions because auditory input is more sequential. This would lead to more mentions of *PATH* within each description in speech in the audio-only condition than in the visual conditions.

As for *MANNER*, it is possible that vision is of advantage here too. For example, in order to differentiate particular *MANNERS*, such as walk vs. run, vision provides richer information than audition about biomechanical properties (e.g., Malt et al., 2014), as well as providing information about speed and direction of motion. So, participants in the visual conditions might describe *MANNER* more often than participants in the audio-only conditions. On the other hand, audition is also good at providing temporal information—such as rhythm of motion (e.g., Recanzone, 2003; Repp & Penel, 2002), so

it is also possible that auditory information might be as rich as visual information and lead to a comparable MANNER encoding of motion.

Regarding co-speech gesture, there are two main possibilities that can be predicted from the previous literature, either visual input is also advantageous for gesture or there is no impact of modality on gesture production. There are three reasons to expect gesture frequency for MANNER and PATH gestures would be higher for visual conditions than the auditory condition. First, gestures—due to the affordances of the visual modality and the possibilities of more easily mapping visuospatial information from vision to gesture—might be more suited for expressing visual information than auditory information (Macuch Silva et al., 2020). For example, signing children use more MANNER and PATH expressions in Turkish sign language than their Turkish speaking peers because of the visually motivated linguistic forms available to sign languages (Sümer & Özyürek, 2022). Second, one might expect gesture to parallel speech patterns (e.g., Kita & Özyürek, 2003; Özyürek et al., 2005), thus leading to more MANNER gestures in the visual conditions than the auditory condition. Finally, PATH gestures might be more difficult to produce with the segmented speech predicted for the auditory condition because gestures are less suited for segmented expressions (Iverson, 1999; Iverson & Goldin-Meadow, 1997), leading to higher rates of PATH gestures in the visual conditions. For all these reasons, visual input may be particularly suited to elicit gestures.

On the other hand, it is possible that there is no difference in the frequency of gesture production between different input conditions. Gesture production theories focusing on the role of mental imagery in gesture, such as the GSA framework, have suggested that “any form of imagery [such as auditory or tactile imagery] that evokes action simulation is likely to be manifested in gesture” (Hostetter & Alibali, 2019, p. 726). This suggests the type of input does not matter for how much gesture is elicited, as long as spatial imagery can be generated. Thus, on this account, participants in all conditions could produce comparable gestures.

Method

Participants

We recruited 90 native Turkish speakers with normal or corrected-to-normal vision from Boğaziçi University. We randomly assigned 30 participants to each of three conditions: audio-only ($M = 21$ years, $SD = 2$, 17 female), visual-only ($M = 22$ years, $SD = 3$, 16 female, 1

nonbinary), and multimodal ($M = 21$ years, $SD = 2$, 10 female, 2 nonbinary). We tested participants in a quiet room on Boğaziçi University campus. They all received extra credit in a psychology course for their participation and provided written informed consent in accordance with the guidelines approved by the IRB committees of Boğaziçi and Radboud Universities.

Stimuli

We made video- and audio-recordings of locomotion and non-locomotion events with an actress. We created 12 locomotion events by crossing 3 MANNERS (walk, run, and limp) with 4 PATHS (to, from, into, and out of) in relation to a landmark object (door or elevator)—such as “*someone runs into an elevator*”. So, participants either only listened to the sound of someone running into an elevator or watched the event with or without the sound. A video and audio recorder were placed next to the landmark objects. For *to* and *into* events, the actress moved towards landmarks, with the PATH direction approaching the audio recorder. For *from* and *out of* events, the actress moved away from landmarks, with the PATH direction away from the audio recorder.

We created 12 non-locomotion events with the same actress performing two-participant “transitive” actions with different objects (e.g., cutting paper, eating an apple), and the video and audio were recorded across from her at a fixed distance. Locomotion events served as the critical items, whereas non-locomotion events were included as fillers. Thus, we did not investigate the non-locomotion events.

There were 24 trials per person, including a total of 12 locomotion ($M_{\text{duration}} = 11.3\text{s}$, $SD_{\text{duration}} = 3.6$) and 12 non-locomotion ($M_{\text{duration}} = 7.7\text{s}$, $SD_{\text{duration}} = 2.3$) events presented in different random orders across participants (see Appendix I for a list of all events and their durations). All stimuli are available at https://osf.io/qe7dz/?view_only=d202c274a186461381c09dc70db6ad39.

The experiment used a between-subjects design with three levels of input modality (audio-only vs. visual-only vs. multimodal).

Procedure

Using a laptop and Presentation Software (Version 20.0, Neurobehavioral Systems, Inc., Berkeley, CA, www.neurobs.com), events were presented as audio-clips to participants in the audio-only condition, as silent video-clips to participants in the visual-only condition, and as video + audio clips to participants in the multimodal condition. All participants regardless of the

condition wore headphones during the task. The instructions were the same across the conditions except the opening sentence (i.e., *in this task, you will “watch video clips” / “listen to sound clips”*). Participants were then asked to describe each event at their own pace without any instructions about gesture use. They were told other participants would watch their descriptions and watch/listen to the same events in order to match descriptions with events.

At the beginning of the experiment, participants performed two practice trials with non-locomotion events. Further clarification was provided, if necessary, after the practice trials. Event descriptions were recorded with a video camera that was approximately 1.5 m across from participants. The experimenter sat across from participants and next to the camera. After each event description, participants proceeded with the next trial at their own pace by pressing a button on the laptop. Participants also filled out a demographic questionnaire on another laptop after the event description task. The experiment lasted around 15 min.

Coding

Speech

All descriptions of locomotion and non-locomotion events were annotated by two native Turkish speakers using ELAN (Wittenburg et al., 2006), but only descriptions for the locomotion events were transcribed and coded. These descriptions were then split into clauses. A clause was defined as a verb and its associated arguments or a verb with gerund phrases. Clauses including locomotion descriptions (e.g., *someone is walking towards the door*) were coded as relevant, whereas clauses including a transitive event—such as *opening a door* or *ringing the bell*—or other information—such as *a person is wearing high heels*—were coded as irrelevant to the target event. Each relevant clause was coded according to the type of information it contained: (a) PATH (trajectory of motion), and (b) MANNER (how the action is performed)—see Table 1 for an example. We calculated the Interclass Correlation Coefficient (ICC) between two coders to measure the strength of inter-coder agreement for PATH and MANNER of motion in speech (Koo & Li, 2016). Agreement between coders was .97 for PATH and .94 for MANNER of motion.

Table 1. An example of coding a description.

| | Clause 1 | | Clause 2 | | | |
|---------------------|-----------------|--------------|-------------------------------|-----------|------------|------------|
| Turkish description | <i>koş</i> | <i>-arak</i> | <i>ev</i> | <i>-e</i> | <i>gir</i> | <i>-di</i> |
| Glossing | run | connective | house | dative | enter | past |
| Coding | MANNER | | PATH | | | |
| English translation | “while running” | | “(someone) entered the house” | | | |

Gesture

Participants’ spontaneous iconic gestures were coded for each target motion event description. We coded gesture strokes (i.e., the meaningful phase of a gesture) that co-occurred with descriptions (Kita, 2000). Each continuous instance of hand movement was coded as a single gesture. Iconic gesture representing trajectory and/or MANNER of motion were further classified into the following categories (see Figure 1 for gesture examples):

- PATH-only** gestures depict trajectory of movement without representing MANNER
- MANNER-only** gestures show the style of movement without representing trajectory
- PATH + MANNER** gestures depict both trajectory and MANNER of movement simultaneously

We calculated the ICC between two coders to measure the strength of inter-coder agreement for identifying a gesture and coding each type of gesture. Agreement between coders was .98 for identifying gestures and between .92–.95 for type of gesture (i.e., .95 for coding PATH only, .92 for MANNER only, and .95 for PATH + MANNER gestures).

Results

To analyse the data, we used linear mixed-effects regression models (Baayen et al., 2008) with random intercepts for participants, items, path type, and manner type, using the packages lme4 (Version 1.1–28; Bates et al., 2015) with the optimiser *nloptwrap* and lmerTest (Version 3.1–3; Kuznetsova et al., 2017) to retrieve *p*-values in R (Version 4.1.3; R Core Team, 2022). We conducted linear mixed effects models on distinct motion elements (PATH and MANNER) in speech and gesture. To assess statistical significance of the fixed factors and their interaction, we used likelihood-ratio tests with χ^2 , comparing models with and without the factors and interaction of interest. For post-hoc comparisons and to follow-up interactions, we used *emmeans* (Version 1.7.3; Lenth, 2022). Data and analysis code are available at https://osf.io/qe7dz/?view_only=d202c274a186461381c09dc70db6ad39.

Speech

Overall differences in the amount of speech produced for visual and auditory motion events

First, we tested whether participants differed in the speech they produced for motion events based on



Figure 1. Example gestures depicting (a) PATH only, (b) MANNER only, and (c) both PATH and MANNER. The full event descriptions are split into clauses (Cl.) and translations are given under each gesture example. The gesture stroke occurred during the underlined speech.

audio-only, visual-only, or multimodal input. Table 2 provides the descriptive statistics for the average number of all clauses, motion event clauses, all gestures, and relevant gestures.

We ran a glmer model with the fixed effect of input modality (audio-only, visual-only, multimodal), the fixed effect of manner type (walk, run, limp), and their interaction term on binary values for mention of motion event clauses in speech (0 = no, 1 = yes) as a dependent variable. See Appendix II for the model summary table. It revealed an effect of input modality,

$\chi^2(2) = 42.43, p < .001, R^2 = .042$. Participants in the audio-only condition had fewer motion event descriptions compared to participants both in the visual-only ($\beta = -1.07, SE = .170, z = -6.32, p < .001, R^2 = .031$) and multimodal ($\beta = -1.07, SE = .170, z = -6.29, p < .001, R^2 = .031$) conditions. There was no difference between participants in the visual-only and multimodal conditions, ($\beta = .006, SE = .178, z = .032, p = .99$). Figure 2 shows the ratio of motion event descriptions (i.e., clauses including locomotion descriptions) in all descriptions.

The model also revealed an effect of manner type, $\chi^2(2) = 7.77, p = .021, R^2 = .002$. Participants had more motion event descriptions for the run than limp events ($\beta = 0.29, SE = .102, z = 2.83, p = .013, R^2 = .002$). But, there was no difference between the walk and limp events ($\beta = 0.09, SE = .102, z = .91, p = .63$) and the run and walk events ($\beta = 0.20, SE = .107, z = 1.82, p = .16$) in terms of the motion event descriptions. The model did not reveal a significant interaction between input modality and manner type, $\chi^2(2) = 9.46, p = .051$.

Table 2. The average number (M) of clauses and gestures across participants with standard deviations (SD, in parentheses).

| Group | All clauses M (SD) | Motion event clauses M (SD) | All gestures ^a M (SD) | Relevant gestures M (SD) |
|-------------|-----------------------|-----------------------------------|--|--------------------------------|
| Audio-only | 35.83 (12.77) | 20.33 (5.25) | 12.03 (6.61) | 9.57 (5.40) |
| Visual-only | 32.83 (9.48) | 25.53 (5.43) | 11.13 (8.86) | 8.83 (7.91) |
| Multimodal | 32.40 (8.41) | 25.27 (4.65) | 11.00 (8.38) | 9.07 (7.28) |

^aAll iconic gestures (relevant or irrelevant) produced within a motion event clause.



Figure 2. Ratio of motion event descriptions. Coloured dots represent the data for each participant. Black dots represent the group mean.

Differences in reference to *PATH* and *MANNER* in speech

Next, we examined whether participants differed in how much they expressed *PATH* and *MANNER* in speech. To account for baseline differences in the number of motion event descriptions produced, we calculated the ratio of mention of *PATH* and *MANNER* per motion event description for each participant and item. We ran a lmer model with the fixed factors of input modality (audio-only, visual-only, multimodal) and type of description (*PATH* vs. *MANNER*) and their interaction term using the ratio of mention of *PATH* and *MANNER* per motion event description as the dependent variable (see Figure 3). The model revealed no fixed effect of input modality, $\chi^2(2) = 1.37$, $p = .50$, but a fixed effect of type of description, $\chi^2(1) = 15.95$, $p < .001$, $R^2 = .008$, showing that *MANNER* was mentioned more than *PATH* in speech. However, the model also revealed an interaction between input modality and type of description, $\chi^2(2) = 31.25$, $p < .001$, $R^2 = .023$. To follow-up the interaction, we first used *emmeans* function to compare the use of *PATH* vs. *MANNER* within each group. There was more mention of *MANNER* than *PATH* in the visual-only ($\beta = .141$, $SE = .028$, $t = 5.03$, $p < .001$) and multimodal conditions ($\beta = .115$, $SE = .028$, $t = 4.11$, $p < .001$), but more reference to *PATH* than *MANNER* in the audio-only condition, $\beta = .068$, $SE = .029$, $t = 2.33$, $p = .020$. That is, *MANNER* and *PATH* were differentially salient in the visual versus auditory conditions.

Second, to follow-up the interaction effect, we also compared reference to *MANNER* and *PATH* separately across input modalities. *PATH* was mentioned more often in the audio-only than visual-only ($\beta = .090$, $SE = .029$, $t = 3.15$, $p = .005$) and multimodal ($\beta = .101$, $SE = .029$, $t = 3.51$, $p = .002$) conditions. Conversely, *MANNER* was mentioned less often in the audio-only than visual-only ($\beta = -.12$, $SE = .029$, $t = -4.15$, $p < .001$) and multimodal ($\beta = -.08$, $SE = .029$, $t = -2.89$, $p = .011$) conditions. There was no difference between the visual-only and multimodal conditions for references to *PATH* ($\beta = .010$, $SE = .028$, $t = 0.36$, $p = .93$) or *MANNER* ($\beta = .036$, $SE = .028$, $t = 1.29$, $p = .41$). See Appendix III for the summary of post-hoc comparisons with *emmeans*.

Gesture

Overall differences in the amount of gesture produced for visual and auditory motion events

We investigated whether participants differed in how much they gestured about different elements of motion events based on input modality (see Table 2 for the descriptive statistics). Because the amount of gesture changes as a function of the rate of motion event descriptions, we first calculated the gesture ratio per motion event description. We compared the groups in terms of their overall gesture ratio using a one-way between-participants ANOVA. There was no significant difference in the gesture ratio between participants in the audio-only

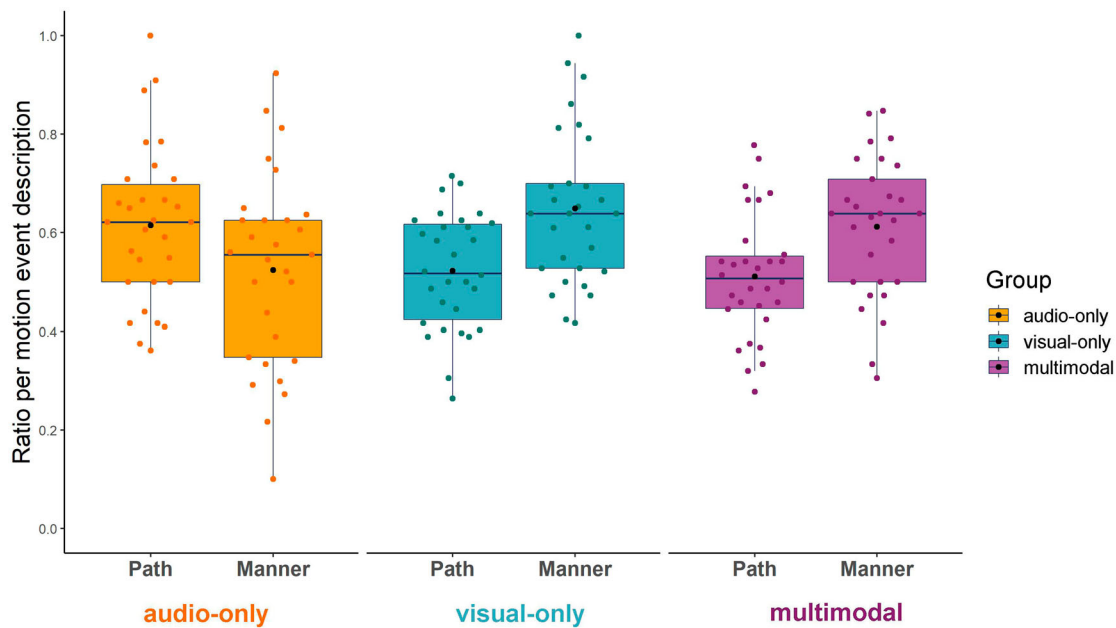


Figure 3. PATH and MANNER in speech. Coloured dots represent the average data for each participant. Black dots represent the group mean.

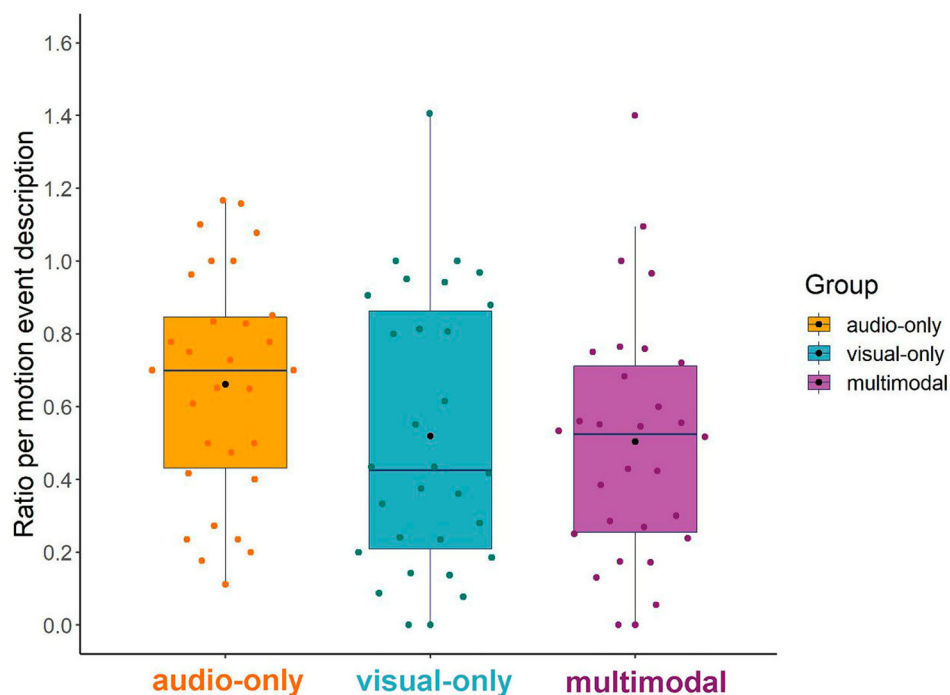


Figure 4. Ratio of gesture for motion event descriptions. Coloured dots represent the data for each participant. Black dots represent the group mean.

($M = 0.59$, $SD = 0.28$), visual-only ($M = 0.44$, $SD = 0.32$), and multimodal ($M = 0.42$, $SD = 0.30$) conditions; $F(2,87) = 2.67$, $p = .08$ (Figure 4).

Differences in PATH and MANNER gestures

To investigate the type of iconic gestures participants produced, we again calculated the ratio of PATH only,

MANNER only, and PATH + MANNER conflated gestures per motion event description for each participant and item. For these calculations, total counts of PATH only, MANNER only, and PATH + MANNER gestures were divided by the number of motion event descriptions for each trial. The data was analysed in the same way as for speech. We ran a lmer model with fixed factors of

input modality (audio-only, visual-only, and multimodal) and type of description (PATH-only, MANNER-only, and PATH + MANNER) using the ratio of PATH and MANNER gestures per motion event description as dependent variable (see Figure 5). The model revealed a fixed effect of type of description, $\chi^2(2) = 531.82$, $p < .001$, $R^2 = .156$. Regardless of input modality, speakers produced more PATH-only gestures than MANNER-only ($\beta = .230$, $SE = .011$, $z = 20.59$, $p < .001$, $R^2 = .107$) and PATH + MANNER gestures ($\beta = .236$, $SE = .011$, $z = 21.14$, $p < .001$, $R^2 = .113$). There was no difference between MANNER-only and PATH + MANNER gestures ($\beta = .006$, $SE = .011$, $z = .55$, $p = .58$). The model revealed no fixed effect of input modality, $\chi^2(2) = 3.64$, $p = .16$, and no significant interaction between input modality and type of description on PATH and MANNER gestures, $\chi^2(4) = 9.29$, $p = .054$. See Appendix IV for the model summary table.

Discussion

Our goal was to investigate whether sensory modality of input influences the multimodal linguistic encoding of spatial information in motion events in speech and co-speech gesture. To determine this, we first examined the quantity of motion event descriptions in speech to establish whether the dominance of vision shown in perception lexicons (e.g., San Roque et al., 2015; Winter et al., 2018) is reflected in the linguistic encoding of motion events under experimental conditions. We found speakers produced more motion event descriptions when they watched events—either multimodal or visual-only—in comparison to when they only listened to events, i.e., audio-only. So, speakers provide richer linguistic information about spatial components of motion events when visual information is available. There was no difference in the amount of motion event descriptions between the visual-only and multimodal conditions, which suggests having auditory input on top of visual input does not further enrich speakers' motion event descriptions. These findings support the proposal that vision dominates in language, extending it to the domain of motion events.

There was, however, a qualitative difference in the verbal expressions of different spatial aspects of motion drawn from visual vs. auditory input. Speakers within the visual conditions mentioned MANNER more than PATH of motion, whereas speakers within the auditory condition mentioned PATH more often than MANNER. In addition, in the audio-only condition speakers mentioned PATH more often than they did in the visual conditions. This finding is in line with earlier studies of space showing non-visual input at encoding might lead to segmented PATH descriptions when describing

routes (e.g., Iverson, 1999; Iverson & Goldin-Meadow, 1997). This might arise from the fact that non-visual spatial information is represented sequentially in contrast to holistic visual information. It is also possible that auditory input foregrounded PATH more than MANNER because information about MANNER of motion is less accessible without visual information. Although audition can provide high temporal acuity to differentiate rhythmic changes of movements (e.g., Recanzone, 2003; Repp & Penel, 2002), it might not provide detailed information to differentiate MANNERS of motion to the same degree as vision (Malt et al., 2014). On the other hand, we used only three simple MANNERS—i.e., *walk*, *run*, and *limp*, which may have been difficult to discriminate between based on auditory input alone. Our findings showed that participants, regardless of the condition, had more difficulty describing the limp than run events. A study using a more diverse set of MANNERS could better test the affordances of audition vs. vision.

Interestingly, Turkish speakers in the visual conditions mentioned MANNER more often than PATH in their speech. Considering the typology of Turkish, this is interesting since Turkish speakers might be expected to omit MANNER more often in motion event descriptions (e.g., Kita & Özyürek, 2003; Özçalışkan et al., 2016, 2018; Slobin, 1996; Talmy, 1985). Our findings suggest there may be universal processes at work, such that vision always provides more detailed information about MANNER of motion than audition, and therefore MANNER of motion might be more salient in visual input, even in a PATH language like Turkish. This suggests the sensory modality of input could influence speakers' encoding of spatial event components independently of the well-established tendencies of speaking a particular language (e.g., Slobin, 1996; Talmy, 1985). Future cross-linguistic studies could tease apart these possibilities systematically.

Although the finding that speakers in the visual conditions mentioned MANNER more than PATH seems discrepant with the usual typological patterns, we are not the first to report a reversed speech pattern in Turkish (e.g., Allen et al., 2007; Ter Bekke et al., 2022). Recently, Ter Bekke et al. (2022) also found that Turkish speakers used more MANNER than PATH when describing motion events presented as silent videos. To explain their findings, they highlighted the fact that they used salient MANNERS—such as *tiptoe*, *twirl*, and *hop*—that are not “default” ways of changing location. Yet, this explanation does not hold for our findings, as the MANNERS in our study were not particularly salient—i.e., *walk*, *run*, and *limp*. Alternatively, Allen et al. (2007) claimed that Turkish speakers are more likely to omit MANNER in larger discourse and when it does not

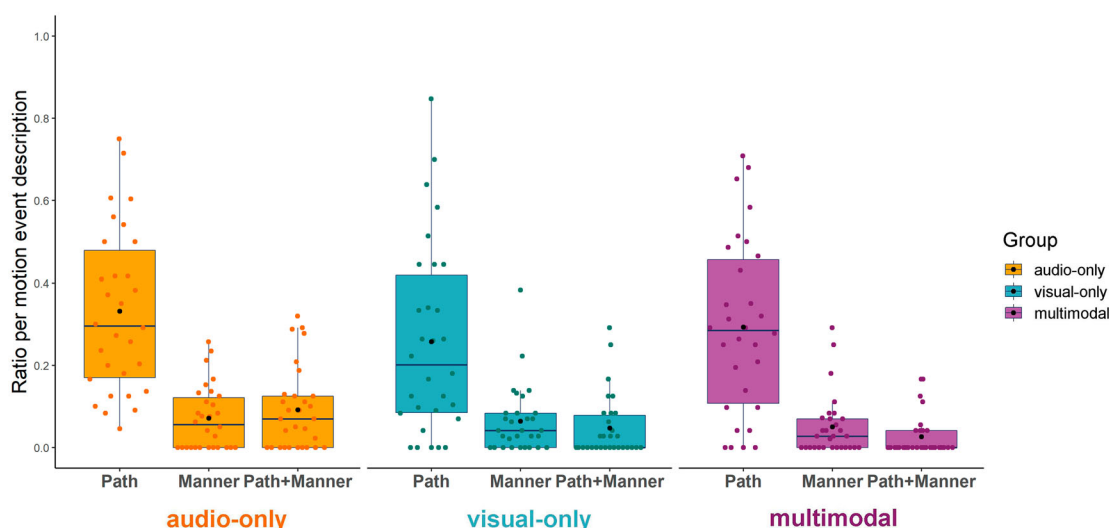


Figure 5. PATH and MANNER gestures for motion event descriptions. Coloured dots represent the average data for each participant. Black dots represent the group mean.

simultaneously occur with PATH in motion events, as used in earlier studies. When MANNER and PATH are simultaneously present in motion events—as in the present study—Turkish speakers mention both elements in their event descriptions. Further studies should examine whether the saliency of MANNER or the ease of expression modulate linguistic encoding of MANNER, particularly in PATH-dominant languages (i.e., verb-framed languages; Talmy, 1985).

For gesture, we predicted that gesture frequency for both PATH and MANNER might decrease in the audio-only condition compared to the visual conditions because of the affordances of the visual modality. Due to the available mapping between gesture and vision (Macuch Silva et al., 2020), gesture production might be easier in the visual conditions than the audio-only condition. However, this was not the case in the present study. We found auditory input alone can elicit similar gesture frequency and gesture types—PATH and MANNER—as visual input. This suggests auditory input can lead to spatial imagery just as visual input does, as explicitly claimed by Hostetter and Alibali (2019). In line with this, Holler et al. (2022) found speakers produce spontaneous co-speech gesture depicting metaphorical spatial features of auditory pitch when describing sounds—e.g., producing a gesture higher in space to depict high pitch notes. Thus, our results support the argument that auditory information can also elicit gesture if it triggers spatial imagery.

Unexpectedly, the difference between PATH and MANNER expressions across input modalities found in speech was not reflected in co-speech gesture. Based on prior work (e.g., Iverson, 1999; Iverson & Goldin-Meadow, 1997), if speech for PATH is segmented, it may

be ill-suited for PATH gesture, and consequently gesture frequency for PATH may decrease. Contrary to this, we found that although participants in the audio-only condition segmented PATH of motion more (i.e., made more reference to PATH in speech) than participants in the visual conditions, the frequency of their PATH gestures did not differ to those produced in the visual conditions. This discrepancy between our results and earlier findings could arise from the fact that these events only had single PATHS. So, although speech for PATH was segmented into smaller units, the amount of segmentation possible might be diminished since we are dealing with smaller-scale PATHS—as in our motion events—compared to larger-scale route description with multiple PATHS. Indeed, Iverson (1999) showed segmentation in PATH descriptions decreases with the diminishing size of a spatial layout.

We found the same discrepancy between speech and gesture for MANNER. Even though speakers in the visual-only and multimodal conditions mentioned MANNER more often in speech, there was no increase in the frequency of MANNER gestures. Regardless of the sensory modality of input, speakers produced more PATH only gestures than MANNER gestures, including PATH + MANNER, even in cases where they mentioned both PATH and MANNER in speech. One might hypothesise that expressing manner in speech was easier than in gesture, and participants might have chosen the modality strategically to avoid confusion for potential addressees who, according to our instructions, would go on to match descriptions to motion events. However, we think this is unlikely since earlier gesture studies of Turkish find that Turkish speakers typically gesture more about path than manner of motion

(Aktan-Erciyes et al., 2022; Mamus et al., [under review](#); Özyürek et al., 2005; Ünal et al., 2022; although see Ter Bekke et al., 2022). So, the few manner gestures observed in our study fit the broader language typology (e.g., Akhavan et al., 2017; Chui, 2009; Gullberg et al., 2008).

Taken together, our findings are more in line with predictions that language typology is the determining factor in gesture production (e.g., Özçalışkan et al., 2016, 2018) and that gestures are mostly shaped by language typology during speaking (e.g., Kita & Özyürek, 2003; Özyürek et al., 2005; Slobin, 1996) rather than sensory input. The discrepancy between our speech and gesture findings also suggests that even though speech affects gesture through language typology, gesture does not solely depend on speech contrary to the suggestions of some theories (e.g., Sketch Model, de Ruiter, 2000; Lexical Retrieval Hypothesis, Krauss et al., 2000; Growth Point Theory, McNeill, 1992), but consistent with the proposal that speech and gesture are independent, yet highly interactive systems (e.g., Gesture as Simulated Action Framework, Hostetter & Alibali, 2008; Gesture-for-Conceptualization Hypothesis, Kita et al., 2017).

Although our results imply the sensory modality of input does not affect the gesture of Turkish speakers, results may differ for a satellite-framed language that encodes MANNER in the main verb—such as English—or an equipollently-framed language—such as Mandarin Chinese (e.g., Brown & Chen, 2013). As MANNER is usually encoded in speech and co-speech gesture in such languages, the affordances of auditory vs. visual input might be more observable in gestural expressions of MANNER—e.g., auditory input may lead to fewer MANNER gestures than visual input. A cross-linguistic investigation is necessary to better understand whether and how co-speech gesture is influenced by the interaction of sensory modality of input and language typology.

Conclusion

The present study examined the role of sensory modality of input on the linguistic expression of motion event components in both speech and co-speech gesture and found they pattern in distinct ways. In comparison to the auditory modality, the visual modality appears to foreground MANNER more than PATH in speech, but gestures are generated similarly regardless of the sensory modality of input. These findings suggest the sensory modality of input influences speakers' encoding of PATH and MANNER of motion events in speech, but not in gesture.

Acknowledgements

We wish to thank Ayşe Serra Kaya and Şevval Çetinkaya for transcription and annotation, and Sevilay Şengül and Eylül Eski for the reliability coding. We also wish to thank Özcan Vardar for technical support during stimulus preparation, Jeroen Geerts for processing video data, and Maarten van den Heuvel for helping with the experimental setup. In addition, we wish to thank Ercenur Ünal and two anonymous reviewers who contributed valuable feedback and suggestions on an earlier version of this manuscript.

Disclosure statement

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

ORCID

Ezgi Mamus  <http://orcid.org/0000-0001-8202-8653>
Aslı Özyürek  <http://orcid.org/0000-0002-0914-8381>

References

- Akhavan, N., Nozari, N., & Göksun, T. (2017). Expression of motion events in Farsi. *Language, Cognition and Neuroscience*, 32(6), 792–804. <https://doi.org/10.1080/23273798.2016.1276607>
- Aktan-Erciyes, A., Akbuğa, E., Kızıldere, E., & Göksun, T. (2022). Motion event representation in L1-Turkish versus L2-English speech and gesture: Relations to eye movements for event components. *International Journal of Bilingualism*. Article 13670069221076838. <https://doi.org/10.1177/13670069221076838>
- Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, 14(3), 257–262. <https://doi.org/10.1016/j.cub.2004.01.029>
- Alibali, M. W. (2005). Gesture in spatial cognition: Expressing, communicating, and thinking about spatial information. *Spatial Cognition & Computation*, 5(4), 307–331. https://doi.org/10.1207/s15427633scc0504_2
- Allen, S., Özyürek, A., Kita, S., Brown, A., Furman, R., Ishizuka, T., & Fujii, M. (2007). Language-specific and universal influences in children's syntactic packaging of manner and path: A comparison of English, Japanese, and Turkish. *Cognition*, 102(1), 16–48. <https://doi.org/10.1016/j.cognition.2005.12.006>
- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4), 390–412. <https://doi.org/10.1016/j.jml.2007.12.005>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1). Article 1. <https://doi.org/10.18637/jss.v067.i01>
- Brown, A., & Chen, J. (2013). Construal of Manner in speech and gesture in Mandarin, English, and Japanese. *Cognitive Linguistics*, 24(4), 605–631. <https://doi.org/10.1515/cog-2013-0021>
- Chui, K. (2009). Linguistic and imagistic representations of motion events. *Journal of Pragmatics*, 41(9), 1767–1777. <https://doi.org/10.1016/j.pragma.2009.04.006>

- Cravotta, A., Busà, M. G., & Prieto, P. (2019). Effects of encouraging the use of gestures on speech. *Journal of Speech, Language, and Hearing Research*, 62(9), 3204–3219. https://doi.org/10.1044/2019_JSLHR-S-18-0493
- de Ruiter, J. P. (2000). The production of gesture and speech. In D. McNeill (Ed.), *Language and gesture* (pp. 284–311). Cambridge University Press.
- de Ruiter, J. P. (2007). Postcards from the mind: The relationship between speech, imagistic gesture, and thought. *Gesture*, 7(1), 21–38. <https://doi.org/10.1075/gest.7.1.03rui>
- Eimer, M. (2004). Multisensory integration: How visual experience shapes spatial perception. *Current Biology*, 14(3), R115–R117. <https://doi.org/10.1016/j.cub.2004.01.018>
- Ernst, M. O., & Bühlhoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences*, 8(4), 162–169. <https://doi.org/10.1016/j.tics.2004.02.002>
- Floyd, S., Roque, L. S., & Majid, A. (2018). Smell is coded in grammar and frequent in discourse: Cha'palaa olfactory language in cross-linguistic perspective. *Journal of Linguistic Anthropology*, 28(2), 175–196. <https://doi.org/10.1111/jola.12190>
- Geangu, E., Roberti, E., & Turati, C. (2021). Do infants represent human actions cross-modally? An ERP visual-auditory priming study. *Biological Psychology*, 160, 108047. <https://doi.org/10.1016/j.biopsycho.2021.108047>
- Gennari, S. P., Sloman, S. A., Malt, B. C., & Fitch, W. T. (2002). Motion events in language and cognition. *Cognition*, 83(1), 49–79. [https://doi.org/10.1016/S0010-0277\(01\)00166-4](https://doi.org/10.1016/S0010-0277(01)00166-4)
- Goldin-Meadow, S., So, W. C., Özyürek, A., & Mylander, C. (2008). The natural order of events: How speakers of different languages represent events nonverbally. *Proceedings of the National Academy of Sciences*, 105(27), 9163–9168. <https://doi.org/10.1073/pnas.0710060105>
- Gullberg, M., Hendriks, H., & Hickmann, M. (2008). Learning to talk and gesture about motion in French. *First Language*, 28(2), 200–236. <https://doi.org/10.1177/0142723707088074>
- Holler, J., Drijvers, L., Rafiee, A., & Majid, A. (2022). Embodied space-pitch associations are shaped by language. *Cognitive Science*, 46(2), e13083. <https://doi.org/10.1111/cogs.13083>
- Hostetter, A. B., & Alibali, M. W. (2008). Visible embodiment: Gestures as simulated action. *Psychonomic Bulletin & Review*, 15(3), 495–514. <https://doi.org/10.3758/PBR.15.3.495>
- Hostetter, A. B., & Alibali, M. W. (2019). Gesture as simulated action: Revisiting the framework. *Psychonomic Bulletin & Review*, 26(3), 721–752. <https://doi.org/10.3758/s13423-018-1548-0>
- Howard, I. P., & Templeton, W. B. (1966). *Human spatial orientation* (p. 533). John Wiley & Sons.
- Iverson, J. M. (1999). How to get to the cafeteria: Gesture and speech in blind and sighted children's spatial descriptions. *Developmental Psychology*, 35(4), 1132–1142. <https://doi.org/10.1037/0012-1649.35.4.1132>
- Iverson, J. M., & Goldin-Meadow, S. (1997). What's communication got to do with it? Gesture in children blind from birth. *Developmental Psychology*, 33(3), 453–467. <https://doi.org/10.1037/0012-1649.33.3.453>
- Kita, S. (2000). How representational gestures help speaking. In D. McNeill (Ed.), *Language and gesture* (pp. 162–185). Cambridge University Press.
- Kita, S., Alibali, M. W., & Chu, M. (2017). How do gestures influence thinking and speaking? The gesture-for-conceptualization hypothesis. *Psychological Review*, 124(3), 245–266. <https://doi.org/10.1037/rev0000059>
- 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](https://doi.org/10.1016/S0749-596X(02)00505-3)
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>
- Krauss, R. M., Chen, Y., & Gottesman, R. F. (2000). Lexical gestures and lexical access: A process model. In D. McNeill (Ed.), *Language and gesture* (pp. 261–283). Cambridge University Press.
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). LmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(1). Article 1. <https://doi.org/10.18637/jss.v082.i13>
- Lenth, R. (2022). *Emmeans: Estimated marginal means, aka least-squares means*. R package version 1.7.3. <https://CRAN.R-project.org/package=emmeans>
- Levinson, S. C., & Majid, A. (2014). Differential ineffability and the senses. *Mind & Language*, 29(4), 407–427. <https://doi.org/10.1111/mila.12057>
- Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2020). The Lancaster sensorimotor norms: Multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods*, 52(3), 1271–1291. <https://doi.org/10.3758/s13428-019-01316-z>
- Macuch Silva, V., Holler, J., Özyürek, A., & Roberts, S. G. (2020). Multimodality and the origin of a novel communication system in face-to-face interaction. *Royal Society Open Science*, 7(1), 182056. <https://doi.org/10.1098/rsos.182056>
- Majid, A., Roberts, S. G., Cilissen, L., Emmorey, K., Nicodemus, B., O'Grady, L., Woll, B., LeLan, B., de Sousa, H., Cansler, B. L., Shayan, S., de Vos, C., Senft, G., Enfield, N. J., Razak, R. A., Fedden, S., Tufvesson, S., Dingemanse, M., Öztürk, O., ... Levinson, S. C. (2018). Differential coding of perception in the world's languages. *Proceedings of the National Academy of Sciences*, 115(45), 11369–11376. <https://doi.org/10.1073/pnas.1720419115>
- Malt, B. C., Ameel, E., Imai, M., Gennari, S. P., Saji, N., & Majid, A. (2014). Human locomotion in languages: Constraints on moving and meaning. *Journal of Memory and Language*, 74, 107–123. <https://doi.org/10.1016/j.jml.2013.08.003>
- Mamus, E., Rissman, L., Majid, A., & Özyürek, A. (2019). Effects of blindfolding on verbal and gestural expression of path in auditory motion events. In A. K. Goel, C. M. Seifert, & C. C. Freksa (Eds.), *Proceedings of the 41st Annual Meeting of the Cognitive Science Society (CogSci 2019)* (pp. 2275–2281). Cognitive Science Society.
- Mamus, E., Speed, L. J., Rissman, L., Majid, A., & Özyürek, A. (under review). Lack of visual experience affects multimodal language production: Evidence from congenitally blind and sighted people. *Cognitive Science*.
- McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago Press.
- McNeill, D., & Duncan, S. D. (2000). Growth points in thinking-for-speaking. In D. McNeill (Ed.), *Language and gesture* (pp. 141–161). Cambridge University Press.

- Özçalışkan, Ş., Lucero, C., & Goldin-Meadow, S. (2016). Is seeing gesture necessary to gesture like a native speaker? *Psychological Science*, 27(5), 737–747. <https://doi.org/10.1177/0956797616629931>
- Özçalışkan, Ş., Lucero, C., & Goldin-Meadow, S. (2018). Blind speakers show language-specific patterns in co-speech gesture but not silent gesture. *Cognitive Science*, 42(3), 1001–1014. <https://doi.org/10.1111/cogs.12502>
- Özyürek, A., Kita, S., Allen, S., Furman, R., & Brown, A. (2005). How does linguistic framing of events influence co-speech gestures?: Insights from crosslinguistic variations and similarities. *Gesture*, 5(12), 219–240. <https://doi.org/10.1075/gest.5.1.15ozy>
- Papafragou, A., Massey, C., & Gleitman, L. (2002). Shake, rattle, 'n' roll: The representation of motion in language and cognition. *Cognition*, 84(2), 189–219. [https://doi.org/10.1016/S0010-0277\(02\)00046-X](https://doi.org/10.1016/S0010-0277(02)00046-X)
- R Core Team. (2022). *R: A language and environment for statistical computing*.
- Recanzone, G. H. (2003). Auditory influences on visual temporal rate perception. *Journal of Neurophysiology*, 89(2), 1078–1093. <https://doi.org/10.1152/jn.00706.2002>
- Repp, B. H., & Penel, A. (2002). Auditory dominance in temporal processing: New evidence from synchronization with simultaneous visual and auditory sequences. *Journal of Experimental Psychology: Human Perception and Performance*, 28(5), 1085–1099. <https://doi.org/10.1037/0096-1523.28.5.1085>
- San Roque, L., Kendrick, K. H., Norcliffe, E., Brown, P., Defina, R., Dingemanse, M., Dirksmeyer, T., Enfield, N. J., Floyd, S., Hammond, J., Rossi, G., Tufvesson, S., Van Putten, S., & Majid, A. (2015). Vision verbs dominate in conversation across cultures, but the ranking of non-visual verbs varies. *Cognitive Linguistics*, 26(1), 31–60. <https://doi.org/10.1515/cog-2014-0089>
- Shams, L., Kamitani, Y., & Shimojo, S. (2000). What you see is what you hear. *Nature*, 408(6814), 788–788. Article 6814. <https://doi.org/10.1038/35048669>
- Slobin, D. (1996). From “thought” and “language” to “thinking for speaking”. In J. J. Gumperz & S. C. Levinson (Eds.), *Rethinking linguistic relativity* (pp. 70–96). Cambridge University Press.
- Slobin, D. I., Ibarretxe-Antuñano, I., Kopecka, A., & Majid, A. (2014). Manners of human gait: A crosslinguistic event-naming study. *Cognitive Linguistics*, 25(4), 701–741. <https://doi.org/10.1515/cog-2014-0061>
- Speed, L. J., & Majid, A. (2017). Dutch modality exclusivity norms: Simulating perceptual modality in space. *Behavior Research Methods*, 49(6), 2204–2218. <https://doi.org/10.3758/s13428-017-0852-3>
- Spence, C., & Squire, S. (2003). Multisensory integration: Maintaining the perception of synchrony. *Current Biology*, 13(13), R519–R521. [https://doi.org/10.1016/S0960-9822\(03\)00445-7](https://doi.org/10.1016/S0960-9822(03)00445-7)
- Sümer, B., & Özyürek, A. (2022). Cross-modal investigation of event component omissions in language development: A comparison of signing and speaking children. *Language, Cognition and Neuroscience*, 1–17. <https://doi.org/10.1080/23273798.2022.2042336>
- Talmy, L. (1985). Lexicalization patterns: Semantic structure in lexical forms. In T. Shopen (Ed.), *Language typology and semantic description* (pp. 36–149). Cambridge University Press.
- Ter Bekke, M., Özyürek, A., & Ünal, E. (2022). Speaking but not gesturing predicts event memory: A cross-linguistic comparison. *Language and Cognition*, 14(3), 362–384. <https://doi.org/10.1017/langcog.2022.3>
- Thalmayer, A. G., Toscanelli, C., & Arnett, J. J. (2021). The neglected 95% revisited: Is American psychology becoming less American? *The American Psychologist*, 76(1), 116–129. <https://doi.org/10.1037/amp0000622>
- Thinus-Blanc, C., & Gaunet, F. (1997). Representation of space in blind persons: Vision as a spatial sense? *Psychological Bulletin*, 121(1), 20–42. <https://doi.org/10.1037/0033-2909.121.1.20>
- Ünal, E., Manhardt, F., & Özyürek, A. (2022). Speaking and gesturing guide event perception during message conceptualization: Evidence from eye movements. *Cognition*, 225, 105127. <https://doi.org/10.1016/j.cognition.2022.105127>
- Viberg, Å. (1983). The verbs of perception: A typological study. *Linguistics*, 21(1), 123–162. <https://doi.org/10.1515/ling.1983.21.1.123>
- Welch, R. B., & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin*, 88(3), 638–667. <https://doi.org/10.1037/0033-2909.88.3.638>
- Winter, B., Perlman, M., & Majid, A. (2018). Vision dominates in perceptual language: English sensory vocabulary is optimized for usage. *Cognition*, 179, 213–220. <https://doi.org/10.1016/j.cognition.2018.05.008>
- Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). *ELAN: A professional framework for multi-modality research*, 1556–1559.