Epistemic intonation and epistemic gesture are mutually co-expressive: Empirical results from two intonation-gesture matching tasks

Joan Borràs-Comes a, b, *, Evangelia Kiagia b, Pilar Prieto b, c

a Universitat Autònoma de Barcelona, Barcelona, Spain
b Universitat Pompeu Fabra, Barcelona, Spain
c Institució Catalana de Recerca i Estudis Avançats (ICREA), Spain

A R T I C L E   I N F O
Article history:
Received 6 January 2019
Received in revised form 3 June 2019
Accepted 8 July 2019

Keywords:
Gesture-speech integration
Gesture-intonation co-dependencies
Epistemic meaning
Gesture
Intonation
Speaker commitment

A B S T R A C T
While several studies have investigated the temporal relationship between co-speech gestures and prosodic structure, little is known about their potential interaction at the level of their encoding of pragmatic meaning. Here we report the results of two complementary intonation-gesture matching tasks which investigate the potential co-dependencies between intonation patterns related to epistemic commitment operators and their associated gestures in Catalan. In Experiment 1, participants were shown audiodemutted videos in which a speaker performed gestures conveying epistemic information of certainty or uncertainty while uttering statements and questions. The subjects were then asked to produce a stipulated target word, the goal being to examine whether they would produce the word with a tune that was semantically consistent with the gestures they had seen. In Experiment 2, participants were primed by hearing intonation patterns conveying epistemic information (certainty-uncertainty) and were then asked to select one of two silent videos which seemed to best match the intonation they had heard. The results suggest converging positive effects in both matching tasks and suggest a close interrelation between the pragmatic representations of intonation and gesture that needs to be taken into account when investigating multimodal pragmatic encoding.

© 2019 Elsevier B.V. All rights reserved.

1. Introduction

Researchers in the few last decades have shown that gestures and speech form an integrated system from both a temporal and a semantic standpoint (Goldin-Meadow, 1998; Graziano and Gullberg, 2018; McNeill, 1992; among many others). From a temporal point of view, crosslinguistic investigations have shown that the prominent parts of co-speech gestures (e.g., gesture strokes and gesture apexes) tend to temporally align with prosodically prominent positions in speech (e.g., pitch-accented syllables and peaks of rising pitch accents; see Wagner et al., 2014, for a review). From a pragmatic point of view, most of the research has traditionally focused on the relationship between representational gestures (e.g., iconic and metaphorical gestures) and speech, and it has been only in recent years that researchers have begun to systematically explore how gestures function at the level of pragmatic encoding and how they relate to the prosodic encoding of the same pragmatic functions.

* Corresponding author. Universitat Autònoma de Barcelona, Barcelona, Spain.
E-mail addresses: joanborrascomes@gmail.com (J. Borràs-Comes), evangeliamaria.kiagia@upf.edu (E. Kiagia), pilar.prieto@upf.edu (P. Prieto).

https://doi.org/10.1016/j.pragma.2019.07.004
0378-2166/© 2019 Elsevier B.V. All rights reserved.
The present study empirically explores the potential co-dependencies between the epistemic functions of gesture and prosody by using two complementary intonation-gesture matching tasks.

Recent priming studies have demonstrated the existence of obligatory and mutual interactions between speech and iconic gestures during the semantic processing of lexical items, leading to Kita and Özyürek's (2003) interface model, which suggests that gestures arise from visuospatial images in the active memory communicating in a bidirectional manner with the speech component. Semantic priming studies involve exposing a subject to information (the prime) and then seeing if this exposure facilitates or otherwise affects the subject’s cognitive processing of other related information (the target) presented immediately afterwards. For example, Bernardis and Caramelli (2007) used words as primes and iconic gestures as targets. The authors found that when the two modalities displayed congruent meanings they were processed more easily. Similarly, Yap et al. (2011) ran a set of priming experiments with iconic gestures as primes and related or unrelated words as targets. They found a clear facilitation effect for related semantic targets and thus argued for a very close semantic link between the two systems. Building on this, So et al. (2013) showed that semantically related gestural primes could speed up word processing. In an experiment along the same lines, Kelly et al. (2010) showed subjects video primes of an action being performed (e.g., someone chopping vegetables) and then offered them three gesture-speech pairs as targets, which were either congruent, weakly incongruent, or strongly incongruent. Their results revealed that participants could not ignore one modality when processing the other, and the authors argue on this basis that speech and gesture form an integrated system and that they mutually and obligatorily interact with each other, in what is known as the integrated systems hypothesis. Finally, these behavioral responses have also been supported by neuroscientific evidence. In an ERP study by Wu and Coulson (2005), participants watched cartoon segments paired with soundless videos of congruous and incongruous gestures followed by words. They found that incongruent words following the cartoon—gesture pairs elicited an N400 effect, a component sensitive to contextual integration. On a similar track, an experiment by Willems et al. (2009) showed that gesture and speech share various neural mechanisms during language processing.

While all the above-mentioned studies focused on a bimodal interaction at the lexical-semantic level, much less is known about the potential interaction between gestures and speech at the pragmatic level, and specifically between gestures and other sources of pragmatic information like intonation. First, following Kendon (1995, 2004; 2017), we acknowledge a distinction between gestures that have a substantive or referential function (in other words, gestures that are related to the utterance or lexical content) and those that have a pragmatic function, that is, gestures which guide the discourse organization or express “aspects of utterance structure, including the status of discourse segments with respect to one another, and the character of the ‘speech act’ or interactional move of the utterance” (1995: 247). In general, little is known about the pragmatic functions encoded by co-speech gestures, and how closely alike they are to those performed by intonation. In the last two decades, studies within the audiovisual prosody approach have claimed that prosody and gesture encode a similar variety of semantic-pragmatic functions, such as marking speech act (assertion, question, etc.), information status (focus, given vs. new information), speaker commitment (epistemic stance of the speaker with respect to the information exchange), politeness, and affective and emotional states, as well as indexical functions related to gender, age, and the sociolinguistic or dialectal status of the speaker (see Krahmer and Swerts, 2009, for an overview of the audiovisual prosody framework; see Brown and Prieto, in press, for a review).

The main goal of the present study is to empirically investigate whether intonation and gesture are semantically co-dependent by focusing on one of their pragmatic functions, namely the expression of epistemic commitment. Epistemic commitment can be understood as a form of stance-taking, or epistemic stance, where speaker certainty is dynamically constructed, making use of interrelated epistemic and evidential forms. In this context, human languages have been shown to use a variety of linguistic strategies, including intonation and gesture, for the expression of a speaker’s epistemic stance towards a proposition (e.g., Palmer, 2001). In this paper, epistemic commitment will be regarded as equivalent to speaker epistemic stance, which includes interrelated but separate notions of epistemicity (or speaker certainty or belief about the proposition expressed) and evidentiality (or the source or evidence that the speaker has to back up the proposition expressed; see Prieto and Borràs-Comes, 2018, for a review).

On the one hand, intonation is clearly an important feature of stance marking. Prosodic studies have documented a variety of epistemic intonation contours across languages expressing degrees of commitment in both statements and questions. For example, Heritage (2013) noted that rising intonation at the end of a question or statement is used in English to express low commitment, thus triggering with greater ease a response from the interlocutor. Similarly, low commitment statements are typically produced with a sentence-final rising contour in most varieties of Catalan (Prieto et al., 2015; Roseano et al., 2016). For questions, speakers of the Bari Italian dialect have been shown to convey their commitment by using different intonation patterns: while unmarked information-seeking questions are generally produced with a final rise, high commitment questions (also called confirmation questions) are expressed with a falling tune (Grice et al., 1997). Central Catalan and Majorcan Catalan speakers can also use different intonation patterns to distinguish between information-seeking and confirmation-seeking questions (Vanreel et al., 2013; Prieto and Borràs-Comes, 2018). In a recent acceptability judgment task, Prieto and Borràs-Comes (2018) showed that a set of question intonation contours were pragmatically interpreted by Catalan listeners as epistemic operators capable of encoding (a) the speaker’s commitment to the proposition expressed (e.g., low and high commitment), and (b) the speaker’s agreement with the interlocutor for a given discourse (e.g., low and high agreement).
On the other hand, gesture too has been shown to mark epistemic stance. Several studies have documented a variety of gestural patterns and facial expressions involved in the expression of low epistemic commitment, such as head tilting or shaking, shrugging, or eyebrow furrowing. Dijkstra, Krahmer, and Swerts (2006) found that co-speech gestures and facial expressions conveyed information related to the degree of knowledge held by a speaker. They investigated the cues that participants used to assess the degree of certainty of a speaker when answering factual questions. Their results showed that visual cues played the most important role for estimating the speaker’s knowledge and degree of certainty. Similarly, Krahmer and Swerts (2007) showed that listeners can draw conclusions about the low commitment state of their interlocutor based on a combination of gestural and auditory information.

An increasing number of studies have documented how epistemic stance in questions and statements can be produced multimodally through combinations of specific gestural and intonational patterns in languages like Catalan, Dutch, or Portuguese (e.g., Borràs-Comes and Prieto, 2011; Borràs-Comes et al., 2011; Crespo-Sendra et al., 2013; Roseano et al., 2016; Cruz et al., 2015). Similarly, other studies have focused on the relative contribution of pragmatic intonation on the one hand and gestures and facial cues on the other in the interpretation of epistemic information, with a set of contradictory results (see Borràs-Comes, 2012, for a review). Borràs-Comes and Prieto (2011) tested the importance of facial cues relative to intonational cues to signal contrastive focus statements and echo questions in Catalan. Their results showed that listeners gave priority to facial expressions when deciding between a contrastive focus statement and echo question interpretation, indicating that auditory intonational cues played a secondary role. Similarly, Borràs-Comes et al. (2011) found that visual information was more important than prosody and lexical cues in the interpretation of uncertainty in Central Catalan statements. Crespo-Sendra et al. (2013) investigated the difference in the perceptual processing of incredulity in Dutch and Catalan interrogative sentences. Here, however, while Catalan participants again preferred the facial expression cues, Dutch participants relied more on intonational differences.

Although the abovementioned studies offer information about a partial overlap between epistemic information expressed in prosody and gesture, to our knowledge no study in the literature has empirically investigated whether speakers are able to congruently associate the two modalities in matching tasks. In short, little is known about the potential semantic co-dependencies between pragmatic intonation and gestures. Here we describe how two cross-modal matching tasks involving intonation and gesture were used to explore the mutual co-dependencies between intonation patterns that convey epistemic stance and co-speech gestures that do the same. The language involved was Catalan, a language particularly well suited for this because it has been shown to make use of clear distinctions in intonation for the expression of epistemic meanings (Vanrell et al., 2013; Prieto and Borràs-Comes, 2018). We made the two experiments complement each other by reversing the order of modalities in terms of input vs. output. In the first experiment, we explored the priming effect of observing gestures on the production of intonation, while in the second we explored the priming effect of hearing intonation on the selection of observed gestures. Note that the two experiments were not strictly speaking mirror images of each other (e.g., matching auditory input to visual cues and vice versa), given that in the first experiment participants were asked to produce verbal output in reaction to a prime, while in the second they were asked to merely select from a set of options rather than actually produce a gesture. This is because we felt that more open association responses would reveal stronger association results in terms of intonation than merely offering participants a closed set of options from which to choose. We hypothesized that exposure to one stimulus (be it a gesture or intonation) would directly influence the participants’ subsequent output in another modality. We thus expected participants in the first experiment to produce utterances with intonation that would match the message conveyed by the gestural primes they have seen. We likewise expect the intonational primes in the second experiment to guide the participants’ eye-movements and responses towards the matching set of gestures. The confirmation of both predictions would provide support for the hypothesis that epistemic gestures and intonation patterns work together from a pragmatic point of view in speech production and perception.

2. Experiment 1

Experiment 1 consisted of a cross-modal matching task which would test whether priming participants with silent epistemic gestures would prompt a matching effect at the intonational level in the subsequent production of sentences. Since primes are in the visual modality and targets are in the auditory, positive results would suggest a tight semantic relationship between the two.

2.1. Methods

2.1.1. Participants

We recruited 20 Catalan native speakers (8 males, 12 females; mean age = 20.9 years, SD = 1.7), all undergraduate students at the Universitat Pompeu Fabra in Barcelona. Since the experiment was to be carried out in Catalan, prior to enrollment, the participants, all Catalan-Spanish bilinguals, were asked to complete a language questionnaire to determine the degree of dominance of Catalan as their mother tongue. The results showed that this was indeed the case, the average proportion for
daily use of Catalan relative to other languages being 74.2% (SD = 8.94). All participants gave their prior informed consent and received a small stipend for their participation in the task.

2.1.2. Materials

The experiment followed a $2 \times 2$ design, with the resulting target four epistemic patterns pertaining to two sentence-type conditions (statement, question) and two commitment conditions (high, low), as shown in Table 1 below. We gave each of these four pragmatic contexts a label characterizing its features, shown in the third column ('Pragmatic condition') of the table. The fourth column provides statements to clarify the epistemic meaning of each of the four conditions, where 'X' refers to any particular proposition. (For more information about these epistemic conditions, see Prieto and Borrás-Comes, 2018).

To obtain representative gesture-intonation pairs for these four epistemic meanings, six native speakers of Central Catalan were recorded at the Universitat Pompeu Fabra using a Panasonic HD AVCCAM camcorder. In these recordings, the speakers were presented with 16 different discourse contexts, four for each epistemic condition. By way of illustration, Table 2 shows the English translation for one of each of the four discourse contexts used for each condition.

Table 1
Epistemic conditions used in the study.

<table>
<thead>
<tr>
<th>Sentence Type</th>
<th>Degree of Commitment</th>
<th>Pragmatic Condition</th>
<th>Schematic Epistemic Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statement</td>
<td>High</td>
<td>neutral statement</td>
<td>I am sure that I know X.</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>uncertainty statement</td>
<td>I have little or no information about X.</td>
</tr>
<tr>
<td>Question</td>
<td>High</td>
<td>Confirmation question</td>
<td>I am quite sure I know X but I am asking for confirmation.</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>neutral question</td>
<td>I have no information about X and I am asking to find out.</td>
</tr>
</tbody>
</table>

Table 2
English translations of sample discourse contexts for each of the four epistemic conditions used to elicit appropriate intonation and gestures from six native speakers. An example of the target output utterance is shown for each context in square brackets.

<table>
<thead>
<tr>
<th>Sentence Type</th>
<th>High Commitment (neutral statement)</th>
<th>Low Commitment (uncertainty statement)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statement</td>
<td>You work as a chef at a restaurant, and a customer asks you which spice you have used in a dish.</td>
<td>You are talking with a fellow student from university about the courses you're currently taking. He asks you what the name of your language professor is. You would say that her name is Barbara, but you're not 100% sure.</td>
</tr>
<tr>
<td></td>
<td>You know the recipe very well, and you answer that the basic spice ingredient of that dish is black pepper. [Expected utterance example: Its basic spice ingredient is black pepper.]</td>
<td>Tell him that you think that's her name. [Expected utterance example: I think it's Barbara.]</td>
</tr>
<tr>
<td>Question</td>
<td>You are with Laura, a friend of yours from university, having dinner at a restaurant near the university. You notice that Laura just pecks briefly at her food and then ignores it.</td>
<td>The weather is a little bit humid and you don't know if the wet laundry you left out to dry has dried yet. Your brother went to take a look and now comes towards you.</td>
</tr>
<tr>
<td></td>
<td>It seems clear that she isn't hungry, but you check with her that anyway. [Expected utterance example: You're not hungry, right?]</td>
<td>Ask him whether the clothes are dry. [Expected utterance example: Is the laundry dry yet?]</td>
</tr>
</tbody>
</table>

Once all six speakers had been recorded, we checked the intonational and gestural features displayed in the videos for each of the four conditions against the results of previous studies on the multimodal expression of commitment in Catalan, as summarized in Table 3, and found them to be consistent. High commitment statements (e.g., certainty statements) were characterized by the use of the two falling pitch contours reported in the literature, L* L% or L+H* L%, and the use of head nods (Hübscher et al., 2017; Borrás-Comes et al., 2011; according to the literature, the use of one or the other of these pitch contours is dependent on the specific prosodic structure of the sentence, such that L+H* L% is used when an utterance consists of one prosodic group, and L* L% when there are two prosodic groups). Low commitment statements (uncertainty statements) were produced with a rising tune (specifically L* H%, and also a variant of this tune L+H* H%) together with facial cues such as squinted eyes, raised eyebrows, and head tilts (Hübscher et al., 2017; Borrás-Comes et al., 2011). High commitment questions (i.e., confirmation questions) were produced with a L+;H* L% intonation contour and accompanied by a head nod and raised eyebrows. Finally, low commitment questions (e.g., neutral questions) were expressed with a L* H% contour and accompanied with a head tilt-backwards movement and some degree of eyebrow furrowing (see Crespo-Sendra et al., 2013).
The video recordings were then used as reference models for the creation of the stimuli used in Experiment 1. These materials were created by video-recording a native speaker of Catalan (the first author of the study) as he spoke a set of four utterances in each of the four epistemic conditions while performing the appropriate intonation and gestures found in the literature and in the pilot examination conducted previously. These four utterances varied in their grammatical nature (i.e., two were noun phrases and two were subject-implicit full sentences) and prosodic structure (i.e., one or two prosodic words) to control for possible variation issues. The two noun phrases were matemàtiques (‘mathematics’) and el gelat de vainilla (‘the vanilla ice-cream’), and two sentences were ballaran (‘[they] will dance’) and aniran a Badalona (‘[they] will go to Badalona’). All four could potentially occur as stand-alone utterances in Catalan.

Once again, the 16 video recordings (four utterances × four conditions) were made at the laboratory facilities of the Department of Translation and Linguistic Science at the Universitat Pompeu Fabra using a Panasonic HD AVCCAM camcorder. For each utterance, the speaker first silently read each discourse context by way of orientation and was then recorded producing the elicited utterance at a normal speaking rate, facing the camera. Only his head and the upper part of his body were recorded. Fig. 1 shows stills taken from the videos to illustrate the representative gesture–intonation information displayed in each condition. Schematic diagrams represent the intonation pattern(s) used in the accompanying audio track of the recording.

### Table 3
Description of the epistemic intonation patterns and gestures in the experimental materials and previous reports in the literature. For each of the four epistemic conditions, matching intonational patterns appear in boldface. Please note that some of them are considered variations of the same nuclear tune configurations.

<table>
<thead>
<tr>
<th>Sentence-Type</th>
<th>Degree of Commitment</th>
<th>Intonation patterns reported in the literature</th>
<th>Gesture patterns reported in the literature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statement</td>
<td>High</td>
<td>( L^* L% ) (as well as other similar low-fall tunes, e.g., ( L^+H^* L% ); Prieto et al., 2015; Hübscher et al., 2017)</td>
<td>Head nods (Hübscher et al., 2017; Borràs-Comes et al., 2011)</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>( L^* H% ) (as well as other similar rising tunes, e.g., ( L^+H^* H% ); Borràs-Comes et al., 2011; Hübscher et al., 2017)</td>
<td>Squinted eyelids, shoulder shrugs, stretched lips, head tilt (Hübscher et al., 2017)</td>
</tr>
<tr>
<td>Question</td>
<td>High</td>
<td>( L^+H^* L% ) (as well as other similar rising-falling tunes, e.g., ( L^+-H^* L% ) and ( H^* L^* ); Prieto and Borràs-Comes, 2018)</td>
<td>Repeated head nods, eyebrow raising (Crespo-Sendra et al., 2013)</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>( L^* H% ) (Prieto and Borràs-Comes, 2018; Prieto et al., 2015)</td>
<td>Head tilt or back, eyebrow furrowing (Crespo-Sendra et al., 2013; Cruz et al., 2015)</td>
</tr>
</tbody>
</table>

The video recordings were then used as reference models for the creation of the stimuli used in Experiment 1. These materials were created by video-recording a native speaker of Catalan (the first author of the study) as he spoke a set of four utterances in each of the four epistemic conditions while performing the appropriate intonation and gestures found in the literature and in the pilot examination conducted previously. These four utterances varied in their grammatical nature (i.e., two were noun phrases and two were subject-implicit full sentences) and prosodic structure (i.e., one or two prosodic words) to control for possible variation issues. The two noun phrases were matemàtiques (‘mathematics’) and el gelat de vainilla (‘the vanilla ice-cream’), and two sentences were ballaran (‘[they] will dance’) and aniran a Badalona (‘[they] will go to Badalona’). All four could potentially occur as stand-alone utterances in Catalan.

Once again, the 16 video recordings (four utterances × four conditions) were made at the laboratory facilities of the Department of Translation and Linguistic Science at the Universitat Pompeu Fabra using a Panasonic HD AVCCAM camcorder. For each utterance, the speaker first silently read each discourse context by way of orientation and was then recorded producing the elicited utterance at a normal speaking rate, facing the camera. Only his head and the upper part of his body were recorded. Fig. 1 shows stills taken from the videos to illustrate the representative gesture–intonation information displayed in each condition. Schematic diagrams represent the intonation pattern(s) used in the accompanying audio track of the recording.

![Fig. 1](image_url) Stills from the video recording used as stimulus material for Experiment 1. The speaker is shown producing gestural cues for the four epistemic conditions, as described in adjacent text columns. Inset schematic diagrams exemplify the intonation contour corresponding to each condition.
Following the recording session, the audio tracks were stripped out of each of the 16 recordings to create a set of soundless videos. The resulting 16 silent video clips (each roughly 1 s in duration) could then serve as gestural primes.

2.1.3. Procedure

The 16 silent clips were embedded into a PowerPoint presentation in a randomized order, with each clip repeated twice in quick succession. Each pair of video slides was immediately preceded by a text slide showing the verbal content of whatever utterance the speaker was about to produce. This would eliminate any need for participants in the experiment to decode what the speaker was saying by lip-reading since they would already know what he was going to say. The two video clips in each stimulus were followed by a black screen lasting 1 s and then a prompt screen asking the subject to say the utterance they had read before watching the clips. This sequence is illustrated in Fig. 2.

![Fig. 2](image)

Fig. 2. Illustration of a trial sequence of Experiment 1. The English translation for the test in the last slide would be “Produce the word/sentence spoken with a tune that you think matches the previous video”.

Participants were tested individually in the same laboratory facilities where the materials had been recorded. They were first seated in front of a laptop computer and given a brief training session by one of the researchers to familiarize them with the experimental task. They then carried out the experimental task itself.

The basic procedure, followed for each of the 16 utterance-condition videos, was for the participant to read the text slide showing the utterance, view twice the corresponding silent video of the speaker speaking that utterance in one condition, then speak that utterance him/herself with the intonation they thought the speaker had signaled gesturally. Thus, participants were first exposed to a gestural prime and then prompted to produce an intonational contour.

The full experiment lasted approximately 15 min per participant. The entire session was video-recorded with a Panasonic HD AVCCAM camcorder, located in front of the participant (though only the audio component was of interest in Experiment 1).

2.2. Results

Experiment 1 yielded a total of 320 audio recordings (20 participants × 16 utterances). Using Cat_ToBI notation (Escudero et al., 2012; Prieto, 2015), the 320 recorded utterances were prosodically coded for their nuclear pitch configuration by the first author, who is an expert prosodic labeler of the Catalan language. After coding, in order to homogenize the variation present in the tunes into one common nuclear pitch configuration, all phonetic variation was merged into basic patterns. For example, the range of rising-falling nuclear pitch configurations (e.g., L+H* L%, L+H* L% L%, H* L%) were merged into one category of a rising-falling tune, L+H* L% (see Table 3). After merging, the following basic pitch patterns emerged: [1] L* H% (low-rise tune), [2] L+H* L% (rise-fall tune), [3] L* L% (low fall), [4] H* LH% (rise-fall-rise tune), and [5] L* HL% (fall-rise-fall tune).
Because the goal of the experiment was to assess whether a visual prime showing gestures conveying particular epistemic stances would induce subjects to produce an utterance with intonation matching that stance, we first classified all nuclear pitch configurations obtained as “matching” or “non-matching” depending on whether they were consistent with the gestural message depicted in the stimulus video or not. For an utterance to qualify as “matching”, an exact match had to be obtained between the intonation pattern it displayed and the corresponding priming gesture (see Table 3 for the four target gesture—intonation matches).

The bar graph in Fig. 3 shows the proportion of appropriately matched intonation scores obtained for each of the four epistemic conditions. It can be seen that the obtained contours match the epistemic condition expressed by the gestural prime 80% of the time or more in all conditions except for high commitment questions, which triggered both the expected \( L^+H^*L^- \) pitch patterns but also the general interrogative pitch contour \( L^*H^- \).

![Fig. 3. Mean proportion of gesture-intonation matching scores broken down by sentence-type (statement, question) and degree of commitment (low, high).](image)

We then ran a Generalized Linear Mixed Model (GLMM) using IBM SPSS Statistics 23, with \textit{accuracy} as the dependent variable (Binomial distribution, Logit link). A random intercept was allowed for \textit{Participant} (not significant: \( \hat{\beta} = .697, p = .074 \)). The fixed factors were \textit{sentence type} (statement, question), \textit{Commitment} (low, high), and the interaction \textit{sentence type} \( \times \) \textit{Commitment}. A main effect was found for \textit{sentence type}, \( F(1, 316) = 25.521, p < .001 \), but not for \textit{Commitment}, \( F(1, 316) = .636, p = .426 \), and their two-way interaction was found to be significant, \( F(1, 316) = 20.117, p < .001 \). The main effect of \textit{sentence type} indicates that statements were more often primed by the corresponding gestures than were questions (\( \hat{\beta} = .237, p < .001 \)), but this effect must be called into question if we consider the interaction between \textit{sentence type} and \textit{Commitment}. In this interaction, there was no significant difference between statements and questions regarding the expression of low commitment (\( \hat{\beta} = .024, p = .639 \)), but statements more often matched the prime than questions when high commitment was being conveyed (\( \hat{\beta} = .485, p < .001 \)). A complementary view of the interaction is that statements displaying high commitment more closely matched the visual prime than statements displaying low commitment (\( \hat{\beta} = .091, p = .033 \)), and that questions displaying low commitment more closely matched the visual prime than questions displaying high commitment (\( \hat{\beta} = .370, p < .001 \)).

In general, the results of the gesture-intonation matching task in Experiment 1 confirmed our predictions that speakers would recognize epistemic meanings in silent gestures and be able to use them consistently to produce matching intonation patterns. In Experiment 2, we will test for the reverse cross-modal effect, that is, the influence of epistemic intonation in guiding visual attention towards target epistemic gestures.

3. **Experiment 2**

Experiment 2 was intended to explore whether hearing audio primes containing intonational cues to epistemic distinctions will guide a subject’s ability to choose from among several candidates a silent video showing the gestural cues that most closely match the epistemic content of the prime.
3.1. Methods

3.1.1. Participants

We recruited 30 Catalan native speakers (6 males, 14 females; mean age = 21.5; SD = 3.3), all undergraduate students at the Universitat Pompeu Fabra in Barcelona, none of whom had been a participant in Experiment 1. Since they were all Catalan-Spanish bilinguals, participants were also asked to complete the questionnaire regarding their average daily use of Catalan. The results showed that they used it on average 75.7% of the time (SD = 8.5%). All participants reported normal or a corrected-to-normal vision. They all gave prior informed consent and received a small stipend for their participation. The data from two participants had to be excluded from the final analysis because of faulty eye-tracking equipment calibration at the outset of the experimental session.

3.1.2. Materials

The experimental materials reflected the same four epistemic conditions as those used in Experiment 1, namely high commitment (neutral) statements, low commitment (uncertainty) statements, high commitment (confirmation) questions, and low commitment (neutral) questions. Like in Experiment 1, the materials were created by making video recordings of a native speaker of Catalan (the first author) speaking a set of utterances while performing gestures appropriate to the four conditions as per Table 3. During these recordings too, the speaker faced the camera, head and upper body visible, and spoke at a normal speaking rate. In this experiment, the target utterances all consisted of 20 one-word trisyllabic items (all well-known verbs and nouns) with penultimate stress (e.g., cantava ‘s/he sang’, esquena ‘back’, etc.). Fig. 4 shows an example of the target word aranya (‘spider’) as produced in the four epistemic conditions. This yielded a total of 80 video clips (20 words × 4 conditions). The audio and video tracks of each clip were separated to create audio (intonational) primes and silent video (gestural) targets respectively.

![Fig. 4. The four intonation contours recorded for the target word aranya (‘spider’) as presented in the four different epistemic conditions. From left to right: high commitment statement produced with a L+H* H% contour, low commitment statement produced with a L* H% contour, high commitment question produced with a L+H* L% contour, and low commitment question produced with a L* H% contour.](image)

Furthermore, in order to disguise the purpose of the study and prevent speakers from using matching generalizations, a set of distractor video-recordings was created by having the same speaker perform four other pragmatic meanings (namely a vocative intonation pattern accompanied by a calling gesture, a greeting intonation accompanied by a greeting gesture, a neutral statement intonation accompanied by no gesture, and a neutral statement intonation accompanied by a grooming gesture). The speaker spoke the same 20 trisyllabic words used in the target materials in each of these four conditions, and the audio and video tracks for all recordings were again separated, yielding a total of 80 filler video clips having the same form as the experimental items but representing four other pragmatic meanings. Thus, a total of 160 video stimuli were recorded.

Tobii Studio software (Tobii Technology) was used to create the sequence of stimuli that constituted the experiment. Each stimulus consisted of a sequence of audio and then video content which would be displayed on a computer screen, as illustrated in Fig. 5. The first thing to appear in the sequence was a white screen with a cross in the center. Not part of the experiment per se, this was intended to allow calibration of eye-tracking equipment. This was followed by the experiment proper, which consisted of 20 test sequences. In each sequence, first, against a black screen, the participant heard an audio recording of one of the 20 words being spoken with a particular intonation contour (the audio prime), which lasted roughly
1.3 s. Then a screen appeared with two silent video clips, side by side, which played simultaneously for about 1 s. One of the two clips—the target—showed a person speaking while performing a set of silent gestures conveying an epistemic stance that was consistent with the intonation of the utterance in the audio prime. The other clip—the competitor—showed gestures that did not match the intonation of the audio prime but rather that of the opposite degree of commitment. The reproduction speed of both videos was slowed down by 40% so that the total play time of the two videos was about 1.4 s. This was done because it was felt that 1 s would be insufficient time for participants to fully process the visual information contained in two videos at once. The relative position on the screen of target and competitor video were counterbalanced so that the target appeared on the left 50\% of the time, and the order of the stimuli was also randomized so that participants could not expect a specific sequence of conditions. Finally, a blank screen was shown for 3 s to allow time for subjects to record their choice of video clip by pushing a key on the computer keyboard.

3.1.3. Procedure

Participants were tested in the same laboratory room at the Universitat Pompeu Fabra as was used in Experiment 1. They were seated at a distance of approximately 60 cm from a computer screen to which was affixed a Tobii X2-60 Wide Remote eye-tracker with a sampling rate of 60 Hz (Tobii Technology). The eye tracker was calibrated at the onset of testing by means of the initial fixation screen.

Participants first read the instructions and completed ten practice trials and then the experiment proper began. As described above, participants first heard a word being spoken and then had to choose the one of two video clips which they felt displayed gestures consistent with the intonation they heard in the prime. Participants saw the initial fixation slide only once, at the outset of the trial.

Fig. 5. Illustration of the sequence of materials comprising each test item in Experiment 2. A 1.3-second audio track against a black screen is followed by a screen showing two silent video clips. Then a blank screen appears to give participants time to indicate which clip showed gestures consistent with the intonation they heard in the prime. Participants saw the initial fixation slide only once, at the outset of the trial.

3.1.3. Procedure

Participants were tested in the same laboratory room at the Universitat Pompeu Fabra as was used in Experiment 1. They were seated at a distance of approximately 60 cm from a computer screen to which was affixed a Tobii X2-60 Wide Remote eye-tracker with a sampling rate of 60 Hz (Tobii Technology). The eye tracker was calibrated at the onset of testing by means of the initial fixation screen.

Participants first read the instructions and completed ten practice trials and then the experiment proper began. As described above, participants first heard a word being spoken and then had to choose the one of two video clips which they felt displayed gestures consistent with the intonation they had heard. They were asked to indicate their choice by pressing either the “F” or “J” key on the laptop keyboard, corresponding to either the left- or right-hand video clips they had seen. As soon as they did so, the next stimulus set began to play. The experiment lasted altogether about 40 min per participant.

For each participant, we recorded eye-tracking data during the presentation of the visual materials, that is, from the onset of the two-video screen (see Fig. 5) until the participant pressed a key, for each of the 20 test sequences. This data consisted of the duration of gaze in milliseconds directed to each of the two areas of interest, target and distractor, and the number of fixations directed to each one. We also recorded the participant’s reaction time (RT)—the time that elapsed between the start of video play-back and the instant s/he pressed a key. In what follows, we present two measures derived from the duration of gaze to the target and competitor. First, we analyze the effects of the total duration of gaze (i.e., the sum of the gaze duration to both target and competitor). Second, we report the results on the proportion of gaze duration directed at target vs. competitor, calculated as the gaze duration to target divided by the total duration of gaze.
3.2. Results

Data for analysis comprised keyboard responses to 80 video pairs from 28 participants for a total of 2240 keyboard responses (80 trials × 28 participants), as well as an equal number for RTs and the two types of gaze duration information described above.

We first analyzed the mean accuracy of participants’ keyboard responses by assigning intonation–gesture congruency scores. A positive accuracy score was obtained per trial if the selected area of interest showed gestures that matched the primed intonation pattern (again, see Table 3 for correspondences), whereas the response was deemed “inaccurate” if the participant chose the competitor. Fig. 6 shows the mean proportion of positive accuracy scores broken down by sentence type (statement, question) and degree of commitment (low, high). The results show that the participants’ tendency to congruently match gestures with intonation primes was high, with an average accuracy rate of 81.93%. The main difference between these results and those of Experiment 1 (see Fig. 3) lie in the behavior of the high commitment question condition. While in Experiment 1 this condition yielded a low average score of 47.5%, in the present experiment the gesture–intonation matching accuracy rate was relatively high at 76%. Nonetheless, as in Fig. 3, it is the condition showing the lowest scores.

We ran a GLMM with ACCURACY as the dependent variable (Binomial distribution, Logit link). A random intercept was allowed for both PARTICIPANT (significant: \( \beta = .323, p = .017 \)) and TEST ITEM (significant: \( \beta = .335, p = .022 \)). SENTENCE TYPE (statement, question), COMMITMENT (low, high), and the interaction SENTENCE TYPE × COMMITMENT were set as fixed factors. A main effect was found for SENTENCE TYPE, \( F(1, 1258) = 10.790, p = .001 \), indicating that statements were more accurately matched than questions (\( \beta = .069, p = .002 \)). However, COMMITMENT was not found to be significant, \( F(1, 1258) = 3.048, p = .081 \). Finally, the interaction SENTENCE TYPE × COMMITMENT was found to be significant, \( F(1, 1258) = 11.332, p = .001 \), which can be interpreted in two different ways. First, the difference between low commitment statements and questions was not significant (\( \beta = .002, p = .951 \)), but high commitment statements were appropriately matched more often than high commitment questions (\( \beta = .126, p < .001 \)). A complementary view of the interaction would be that, regarding statements, those displaying high commitment were correctly matched more often than those displaying low commitment (\( \beta = .092, p = .002 \)), but there were no significant differences between low and high commitment when it came to questions, though the direction of the effect was reversed (\( \beta = .037, p = .272 \)). The pattern of results is thus similar to that seen in Experiment 1.

In order to further understand the results, three GLMMs were conducted to see whether participants’ responses were related to online measures related to psycholinguistic processing, namely REACTION TIME, the TOTAL DURATION OF GAZE as described above, and the PROPORTION OF GAZE DURATION DIRECTED AT TARGET VS. COMPETITOR. The descriptive statistics for these measures are shown in Table 4. These three analyses shared the same fixed and random effects structures. The fixed effects were ACCURACY (inaccurate, accurate), SENTENCE TYPE (statement, question), COMMITMENT (low, high), and all their possible interactions. As for random effects, a random intercept was allowed for both PARTICIPANT and TEST ITEM, as in the previous analysis.
The GLMM results of the reaction times analysis revealed the following effects. On the one hand, there was a main effect of accuracy, $F(1, 1252) = 8.587, p = .003$, indicating that participants took shorter to react when they chose a matching gesture than when they chose an incongruent one ($\beta = 104.926, p = .004$). On the other hand, the paired interaction accuracy $\times$ commitment was found to be significant, $F(1, 1252) = 4.285, p = .039$, which can be interpreted in two ways. First, the preference for inaccurate responses to cause longer RTs than accurate ones was found to be significant for low commitment contexts ($\beta = 178.080, p < .001$), but not for high commitment contexts ($\beta = 30.787, p = .575$). Second, focusing on accurate responses, high commitment contexts displayed higher RTs than low commitment contexts ($\beta = 109.076, p < .001$), whereas for inaccurate responses there were no significant RT differences regarding commitment ($\beta = -38.216, p = .575$).

In the analysis of the total duration of gaze, only one main effect was again found to be significant. In this case it was the main effect of sentence type, $F(1, 1254) = 9.544, p = .001$, which would indicate that participants were more attentive in questions than in statements ($\beta = 251.861, p = .002$). Interestingly, an examination of the contrasts among non-significant effects indicates that this effect would be especially present for accurate responses.

Finally, in the analysis of the proportion of gaze duration directed at target vs. competitor, again only one main effect was found to be significant, the main effect of accuracy, $F(1, 1218) = 159.931, p < .001$, indicating that accurately-answered trials presented far higher proportion of gaze directed at the target than inaccurately-answered ones ($\beta = .202, p < .001$).

### 4. Discussion and conclusions

During face-to-face communication, we constantly make use of gestural and prosodic strategies that help us process information and communicate pragmatic meanings, intentions, and attitudes more effectively. In this study, we investigated the patterns of semantic co-dependencies shown by intonation and gestures signaling various epistemic meanings, namely high vs. low commitment in questions and statements, by using two complementary cross-modal matching tasks. The results showed convincingly that when Catalan participants have access to one of the modalities as initial input (either gesture or intonation), they are able to successfully produce the matched intonational patterns (Experiment 1) or to successfully associate the auditory melodic input to the corresponding gestures (Experiment 2). In more general terms, these results offer evidence of the close interconnection between gesture and intonation at the pragmatic level.

The aim of Experiment 1 was to use a cross-modal gesture-intonation matching task to investigate whether Catalan-speaking participants’ intonational output could be successfully primed by viewing silent gestures. Results for matching scores showed that for at least three of the four conditions participants produced utterances with intonation that was congruent with the gesture they had seen more than 80% of the time. The one exception to this general pattern was high commitment questions, which for produced with two different intonation contours, namely the expected rising-falling pitch contour ($L^* H^% L^+$, a general category representing rising-falling pitch contours) but also the low-rise pitch pattern $L^+ H^%$, which has been described extensively in the literature as a highly polysemic tune used for polar questions in Peninsular Catalan ($L^+ H^%$; Prieto and Borràs-Comes, 2018; Prieto et al., 2015). In our view, the fact that these question stimuli were presented without a previous discourse context might have prevented participants from easily grasping the high commitment question interpretation. The consequence of this might have been the use of a general pitch contour for questions, namely the $L^+ H^%$ pitch contour, for at least half of the elicited utterances. It remains for now unclear whether the addition of previous contextual information could have helped increase the accuracy scores obtained for high commitment questions.

The purpose of Experiment 2 was to use a complementary cross-modal gesture-intonation matching task to examine whether priming participants with intonation conveying various epistemic stances in Catalan would guide their decision when choosing between silent video recordings showing gestures either congruent or incongruent with the intonational messages they had heard. The results showed a tendency to choose congruent gestures that was consistently high across conditions, and high commitment statements were again the most often appropriately matched context. Though high commitment questions again elicited the lowest matching scores, the proportion of congruent choices in Experiment 2 was 76%, well above chance. Yet the processing difficulties involved in obtaining intonational—gestural matches for this pragmatic condition was evidenced by RT and eye-tracking data. Interestingly, the results from the RT responses provided complementary evidence that (a) inaccurate responses were associated with longer RTs than accurate ones; and (b) that, among the latter, high commitment questions were associated with the longest RTs, indicating that this pragmatic context may have
been the hardest to interpret. On the other hand, the analysis of eye-tracking data showed two additional results, namely (a) that the proportion of gaze directed to the target vs. competitor was greater for accurate responses than for inaccurate ones; and (b) that participants were more attentive when they were facing intonation—gesture pairs involving questions than when they were presented with similar multimodal data involving statements.

Nonetheless, why is it that high commitment questions were not classified as successfully as other epistemic conditions, especially in the first experiment? As noted above, this negative effect could be the result of at least two factors. First, the fact that the auditory question stimuli were presented outside a discourse context could have led to interpretation problems, since questions typically seek a response from the interlocutor and thus more directly imply a low rather than high degree of commitment. In relation to this, in most European languages, confirmation-seeking (or high commitment) questions are syntactically marked by manipulating the surface structure of the sentence (e.g., declarative-form questions in English; see Gunlogson, 2001) or by using question tags (Heritage, 2013). In Catalan, while it is possible to express differences in epistemic stance exclusively through the use of intonation (Prieto and Borràs-Comes, 2018; Prieto et al., 2015; Borràs-Comes et al., 2014), the language offers a number of strategies that can be combined with intonation, such as the sentence-initial expletive particle que (analogous to est-ce que in French and é que in Portuguese; Prieto and Rigau, 2007), the use of a question tag form, or the employment of other verbal markers (see, e.g., Prieto and Roseano, 2016). Since our materials for the high commitment question condition did not contain any syntactic markers, their intended meaning might have been difficult to grasp without context. Second, the gestural cues for high commitment questions might have also seemed ambiguous to participants, since head nods and eyebrow raising are used to signal various other meanings. In fact, while previous studies have focused on the gestures used for expressing low commitment (uncertainty) in statements (Hübischer et al., 2017; Borràs-Comes et al., 2011), less is known about the gestural articulators associated with the expression of epistemic commitment in questions. There is therefore a need for further investigation of the use of different kinds of gesture configurations and a larger variety of high commitment question contexts in order to shed more light on the specific bonds between epistemic gestures and intonational patterns in this and other linguistic varieties.

All in all, this is one of the first studies exploring the interdependence between intonation and gesture in a variety of epistemic questions and statements, and more work will clearly be needed to clarify the mechanisms that govern the combined use of intonation, gesture, and lexical and morphosyntactic strategies for the expression of complex pragmatic meanings across languages. In general, our predictions for the two experiments were borne out, namely that (a) silent epistemic gestures can successfully prime the production of the corresponding epistemic intonation (Experiment 1) and (b) epistemic intonation contours can successfully prime listeners to select the corresponding epistemic gestures (Experiment 2). The results from the two experiments reveal for the first time that adults can successfully associate congruent information coming from these two modalities, epistemic intonation and gestures, and that priming with either of these two modalities can successfully trigger matching output in the other. With respect to the discrepancy between the results of the two experiments regarding high commitment questions, it may be due to differences in experimental task design. On the one hand, Experiment 1 was a free-production experiment, as participants were asked to freely respond to visual primes by producing any possible intonation contour in their language; by contrast, the task presented in Experiment 2 was more restricted, as participants were asked to respond to audio primes by selecting one video out of two. This might explain why matching results in the production of high commitment questions (Experiment 1) are lower than matching results in choosing visual depictions of such questions (Experiment 2). As a reviewer has suggested, applying a different methodology to explore the research questions in Experiment 1 could have led to a more similar treatment of high commitment questions between the two experiments. More specifically, using a similar design to that used in Experiment 2 (i.e., presenting subjects with silent video recordings of gestures and then having them select among different pre-recorded intonational patterns) might increase subjects’ preference for a rising-falling intonational pattern in association with the gestural patterns characteristic of high commitment questions. All in all, this must be explored in the near future by specifically focusing on questions and by considering the different patterns used in Catalan to express this modality while controlling for geographical and social variation. This would constitute a natural extension of the present study as well as Prieto and Borràs-Comes (2018).

Be that as it may, the results from these two experiments add new empirical evidence to the literature on gesture-speech semantic integration. While it is well established that speech and gestures are very tightly and obligatorily integrated at the lexical-semantic levels (Kita and Özyürek, 2003; Kelly et al., 2010), less is known about the potential and mutual interaction between prosody and gestures as sources of pragmatic meaning. Though the results of the two matching experiments presented here have shown that speakers successfully match patterns between the two modalities, this does not quite constitute evidence that interactions between the two modalities during speech processing is automatic and obligatory. Thus, whether the integrated systems hypothesis applies between speech and gesture at the pragmatic level is still an open question. As one reviewer has pointed out, a modification of the methodology applied here might provide clearer data in this direction. For example, participants could be shown four types of video clips (gesture match + intonation match, gesture mismatch + intonation match, gesture match + intonation mismatch, and gesture mismatch + intonation mismatch), and then be asked whether the recorded speaker was delivering a high or low commitment statement/question. While higher accuracy in pragmatic detection of epistemicity should be obtained when both gesture and intonation match the epistemic commitment, mismatched or even unimodal conditions should reduce the performance.

Nonetheless, this study shows that speakers of a language are extremely sensitive to the multimodal linguistic cues that express pragmatic information, and when faced with the lack of one of the modalities they are able to establish a congruent selection of tunes and gestures for the expression of epistemic meaning. This is further evidence that all semantic and
pragmatic components of language, whether contributed by lexical meaning or intonational meaning, are deeply connected to gestures. Since this investigation has shown that gestural and intonational configurations are co-expressive and semantically inter-related, this pattern of results suggests that they can even be used methodologically to investigate the realization of their counterparts. More specifically, in studies that need to elicit target intonational data, static or dynamic gestural configurations could be used as a complement to written situational contexts like those used in Discourse Completion Tasks (DCTs, see Vanrell et al., 2018).

All in all, we argue for the need to incorporate a multimodal view of language into sociopragmatic inquiry and investigate prosodic and gestural systems in tandem. More work will be needed to assess how intonational and gestural meanings interact and reinforce themselves in online communication, as well as the potential facilitation effect of multimodal language interaction for both language production and language comprehension processes.

Declarations of interest

None.

Acknowledgements

This research has been funded by a research grant awarded by the Spanish Ministry of Science, Innovation and Universities (PGC2018-097007-B-I00, “Multimodal Language Learning (MLL): Prosodic and Gestural Integration in Pragmatic and Phonological Development”), and by a grant awarded by the Generalitat de Catalunya (2017 SGR 971) to the Prosodic Studies Group. Also awarded by the Generalitat de Catalunya, the first author acknowledges a research group grant to “LLETRA, Linguistic and Literary Education” (2017 SGR 1728).

References

Ozyürek, A., 2003. What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: evidence for an interface repre-
Ozyürek, A., 2003. What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: evidence for an interface repre-
Veilleux, N., 2011. What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: evidence for an interface repre-
Veilleux, N., 2011. What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: evidence for an interface repre-
Veilleux, N., 2011. What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: evidence for an interface repre-


Dr. Joan Borràs-Comes is Adjunct Professor at the Universitat Autònoma de Barcelona and works as a Research Technician for several research groups belonging to five different universities based in Barcelona (UAB, UPF, UB, UOC, and UIC). Since 2016, he is accredited as Tenured Assistant Professor by the Catalan University Quality Assurance Agency (AQU). He has published more than 30 manuscripts in journal or books, and his main research interests are related to the pragmatics of intonation and gesture, and to the study of language contact and variation with a special focus on the suprasegmental level.

Mrs. Evangelia Kiagia is a psycholinguist working in industry. Her main interests are the combination of pragmatics and gesture in human language processing. She studied Linguistics and Computational Linguistics at the universities of Athens and Saarland. She has held two research scholarship at the University of Stuttgart and at Universitat Pompeu Fabra in Barcelona.

Dr. Pilar Prieto is an ICREA Research Professor at the Department of Translation and Language Sciences at Universitat Pompeu Fabra, Barcelona, Catalunya. Her main research interests focus on the communicative role of prosody and gesture in language, as well as their significance in language development and language training for both impaired and non-impaired populations. Currently she is an Associate Editor of Language and Speech and coedited the book The Development of Prosody in First Language Acquisition with N. Esteve-Gibert (2018, John Benjamins).