The acquisition of multimodal cues to disbelief

Meghan E. Armstrong[1], *Núria Esteve-Gibert*[2], *Pilar Prieto*[3,2]

[1] Department of Languages, Literatures and Cultures, University of Massachusetts, Amherst, Massachusetts, USA
[2]Departament de Traducció i Ciències del Llenguatge, Universitat Pompeu Fabra, Barcelona, Spain
[3]Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain
armstrong@umass.edu, nuria.esteve@upf.edu, pilar.prieto@upf.edu

Abstract

In this study, we examine how 3-, 4-, and 5-year-old Catalan-acquiring children are able to make use of audio (intonational) and visual (facial gesture) modalities in the comprehension of speaker disbelief, as well as the role of the child's developing Theory of Mind. Our results suggest that in this case, facial gesture provides children with scaffolding for linguistic meaning, and that explicit belief-reasoning also helps children to infer speaker disbelief. We discuss the implications of these findings for the study of intonational development.

Index Terms: intonational development, multimodal comprehension, acquisition, prosodic meaning

1. Introduction

When inferring meaning, adults may rely on different modalities (auditory and/or visual) in different ways. [1] showed, for example, that when incredulity or *disbelief* meaning is marked through intonation and facial gesture in polar questions, Catalan listeners rely more on the visual modality than Dutch listeners, since Dutch encodes incredulity intonationally in a way that is quite different from other types of questions. Children, then, must learn to make use of both the audio and visual modalities to guide them to meaning. Often times the information from the visual modality *reinforces* the message from the audio modality. [2] found that when preschoolers and kindergartners were faced with the task of comprehending a syntactically complex message, reinforcing gestures facilitated comprehension for preschoolers, but not for kindergartners. They suggested that reinforcing gestures serve as "scaffolding", by guiding the child toward the intended meaning of the utterance. They also pointed out that when a message is conveyed through two modalities (e.g. audio + visual), younger children might disregard one channel altogether for working memory reasons. Older children, on the other hand, are more capable of integrating both modalities. While [2] investigated manual gestures, to our knowledge no studies have investigated the reinforcing nature of facial gestures with respect to speech comprehension in children.

One limitation in the literature is the precision with which the audio modality is described, especially with respect to the prosodic cues that guide listeners to meaning, and that children learn to attend to. Most recently, [3] refer to children's understanding of "emotional cues in the voice", but describe their materials as neutral versus "affectively inflected stimuli" with no acoustic characterization of the prosodic differences between the different emotions. Such descriptions do not allow us to understand the types of prosodic cues that children learn to attend to. Recent work has applied the Autosegmental Metrical framework [4,5] in order to investigate children's comprehension of the intonational aspect of the audio modality. [6] showed that 4-, 5- and 6-year-old children performed at above-chance levels in a linguistic comprehension task where children had to identify a "disbelieving" speaker based on differences in the ¡H* L% and L* HL% nuclear configurations in Puerto Rican Spanish [7]. 6-year-olds, however, significantly outperformed the 4- and 5-year-olds. Therefore, in the absence of visual information, children used linguistic information to perceive belief-state meaning. Linguistic meaning associated with belief states is of particular interest since it involves Theory of Mind (ToM) reasoning [8]. That is, in order to fully comprehend linguistic forms that encode information about speaker and hearer belief states, children must have some ability to "mind-read", i.e. infer the belief states of others. One measure of ToM in children is the false belief task [9, 10]. In recent work assessing ToM and emotion understanding, [11] pointed out that awareness of false belief is needed to understand the human state of surprise since one needs to recognize that something must have contradicted the beliefs of a speaker. They compared children's (ages 3-5) scores on a facial expression task where the child had to select a target label about how a person in a picture felt, based on his/her facial expression. They used a battery of ToM tasks, and found a relationship between belief-based emotion labeling such as surprise, and ToM. The children found the false belief tasks in their study easier than labeling facial gestures of surprise or fear. They claim that children exhibit the ability to pass explicit false belief tasks before they are able to label belief-based facial gestures.

In this study we sought to understand the role that facial gestures might play in children's belief state comprehension, specifically how they use audio, visual or audiovisual cues in the comprehension of a speaker's state of *disbelief*. If facial gesture provides scaffolding to linguistic meaning, there should be an ordered path of acquisition - children should be more successful at comprehending disbelief meaning from facial gesture cues than they are for audio cues. Audiovisual cues could, on the one hand, be more useful than audio only cues because of the presence of the facial gesture, but might also be more difficult since the child must integrate the two modalities. Thus it is possible that children use different strategies for the audiovisual condition. When no visual information is available at all (i.e. audio-only modality), we hypothesize that comprehension of belief states should be more difficult for younger children. If children do not have access to the "scaffolding" they use for meaning, they might simply not have access to the meaning. Further, we hypothesize that children with explicit false belief reasoning should be more successful at the task, regardless of the condition. We tested these hypotheses using the tasks outlined below.

10.21437/SpeechProsody.2014-217

2. Methods

2.1.Participants

Seventy-seven Central Catalan-speaking children participated in the experiment, which consisted of two tasks: a ToM false belief task and a comprehension task. The age range was between 34 and 75 months (M = 53.6 mo.). For the comprehension task, there were three conditions, with a between subjects design: Audio Only (AO), Visual Only (VO) and Audiovisual (AV). A total of twenty-six children received the AO condition: one 2-year-old (35 mo.), six 3-year-olds (range 40-44 mo., M=41 mo.), ten 4-year-olds (range 49-59 months, M=53.5 mo.), six 5-year-olds (range 62-68 mo., M=64 mo.) and three 6-year-olds (range 73-75 mo., M=74 mo.). Twenty-three children received the VO condition: two 2-year-olds (both 34 mo.), six 3-year-olds (range 36-47 mo., M=41.5 mo.), six 4-year-olds (range 52-58 mo., M=54.7 mo.), nine 5-year-olds (range 60-71 mo., M=68 mo.) and two 6-year-olds (range 74-75 mo., M=74.5 mo.). Finally, twenty-eight children received the AV condition: one 2-year-old (34 mo.), eight 3-year-olds (range 37-47 mo., M=41.8 mo.), seven 4-year-olds (range 48-59 mo., M=47.8 mo.), eleven 5-year-olds (range 60-71 mo., M=65.4 mo.) and one six-year-old (72 mo.). The participants were all students at Catalan public elementary schools where Catalan was the primary language of instruction.

2.2.Materials

2.2.1. False belief task

The false belief task was an adaptation of the Sally Ann task [8], a classic ToM task. The materials for this task consisted of a short video (0:54) featuring two puppets. At the beginning of the video a princess puppet appears with a ball, announcing that she will leave her ball in one of two containers in front of her. She leaves the ball in one of the containers and subsequently states that she will leave for school. The princess then leaves for school, disappearing from the video. While the princess is gone, a lion puppet appears. The lion takes the ball out of the container and transfers it into the other container in the scene, covering it so that the ball cannot be seen. The lion laughs in a sneaky way and leaves. The princess puppet then appears, announcing that she has returned from school.

2.2.2. Comprehension task

For all three conditions mentioned above, the materials consisted of a Powerpoint presentation containing AO, VO or AV materials. The premise of the task was such that the child had to decide which member from a set of twins did not believe his or her friend about an animal that the friend claimed to have seen while on vacation. Thus, for each slide (which constituted one trial), participants saw four images: the two twins (upper left and lower left in Fig. 1), the twins' friend (lower right) and an image of the animal the friend claimed to have seen (upper right). The images seen in the upper left and lower left regions of Figure 1 were either a female child or a male child, depending on the block of presentation (2 blocks per participant).

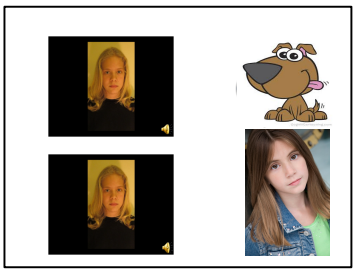

Figure 1. *Slide from AO condition (neutral faces) depicting the two twins (left panels), their friend (lower right panel) and the animal the friend claimed to have seen (upper right panel)*

2.2.2.1. Stimuli

The stimuli for the AV condition were created first. Two child actors, a female (11 years old) and a male (13 years old) were video recorded producing two types of echo questions: neutral echo questions and disbelief echo questions. While there can be variation in terms of the specific pitch contour that might appear for these contexts, the actors produced two different contours labeled L+¡H* L% (neutral echo question) and L* H% (echo question with disbelief marking) in the Cat_ToBI system [12]. (1) and (2) show examples of neutral versus disbelief echo questions:

(1) A and B are talking about what time to leave in the morning in a noisy restaurant. B strains to hear what A has said.

A: I think we should be on the road by eleven.

B: **Eleven**?

(2) B knows A has been a vegetarian for the last ten years.

A: I had the best filet mignon last night. It was cooked to perfection.

B: **Filet mignon!?!** When did you start eating meat?

In (1) B simply repeats an element from A's prior turn and does not convey any type of attitudinal information towards the proposition. In (2) B produces an echo question that repeats linguistic information from the prior discourse, and at the same time expresses her belief state about the propositional content – that she can't believe it.

The intonation contours used for the neutral vs. disbelief echo questions are presented in Figures 2 and 3. Each stimulus was phonetically analyzed in Praat to confirm that the appropriate contour was produced by the actors.

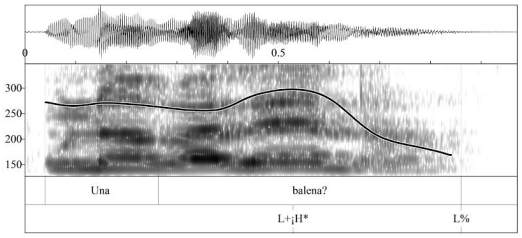

Figure 2: *Pitch track, spectrogram and waveform for the neutral echo question* ***Una balena?*** *'A whale?'*

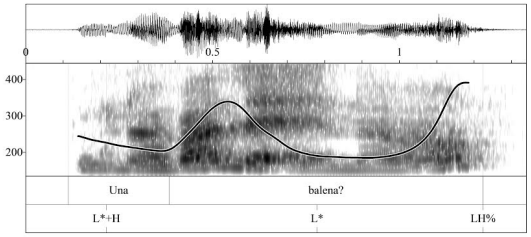

Figure 3: *Pitch track, spectrogram and waveform for the disbelief echo question* ***Una balena?!*** *'A whale?!*

For the AO condition, the audio was extracted from the original AV videos. For the VO condition, the audio was removed so that only the visual information was available. Figures 4 and 5 show typical facial gestures that were presented for the AV and VO stimuli: brow raising and eye-widening for the neutral echo questions, and brow furrowing accompanied with a backwards movement of the head for the disbelief echo questions. Children that received the AO condition saw two identical neutral faces (as shown in Figure 1) presented as screenshots (as in Figure 1).

Figure 4: *Screenshots of facial expressions for neutral echo (left panel) and disbelieving echo (right panel)*

2.3 Procedure

The experimenter was a female native speaker of Central Catalan (the second author of this paper). First, the children were given the ToM task. The experimenter was seated with the child in a quiet room in the child's school. The child was told to watch a video presented on a laptop computer and listen carefully to what the characters said, because afterwards s/he would have to answer some questions. After watching the video, the child was asked two questions: 1.) *On buscarà la pilota, la nena?* 'Where will the girl look for the ball?'; and secondly 2.) *On és la pilota, en realitat?* 'Where is the ball really?' 1.) was considered correct if the child responded that the girl would look for the ball in the container where she left it. 2.) was considered correct if the child said that the ball was in the container where it was moved to.

After the child finished the false belief task, s/he was administered the comprehension task for either the AO, VO or AV condition. This was done with a between subjects design, such that each child saw only one of the three conditions. Like the false belief video, the Powerpoint was presented on a laptop computer. At the beginning of the task each child received familiarization trials. S/he was told that there was a set of twins, and their friend Marta (lower right hand panel in Figure 1). Marta was telling the twins about what she saw when she was on vacation. The children were also told that there would always be a twin that did not believe what Marta said, and that they would have to decide which twin that was based on how the twins reacted to Marta. For instance, the experimenter told the child that Marta was telling the twins that she saw a whale – *La Marta els explica que va veure una balena* 'Marta tells them that she saw a whale'. The experimenter then showed the child a reaction from each twin, one on top, and the other below. For each test trial, one twin produced a neutral echo question and the other produced a disbelief question. The children that received the AV condition saw one of the twins producing the facial gesture for a neutral echo question with the L+¡H* H% intonation contour. The other twin produced the facial gesture for the disbelieving echo question with the L* H% intonation contour. For the AO condition children saw still images and heard only the audio stimuli. For the VO condition the children saw the videos but without audio. After each twin spoke (or gestured as in the case of VO), the experimenter asked *Quin/a bessó/na no es creu la Marta, el/la de dalt o el/la de baix? Assenyala'l/-la.* 'Which twin does not believe Marta, the one on top or the one below? Point to him/her.' The child then pointed to the twin s/he thought did not believe Marta. The answer was marked correct by the experimenter if the child picked the twin that produced brow furrowing/backwards movement of head along with the L* H% for the AV condition, and the one that picked the relevant component of these for the AO (L* H%) and VO (brow furrowing/backwards head movement) conditions. Each participant received two blocks of stimuli. Children were either exposed to the female actor in Block 1, and the male actor in Block 2, or vice versa. They received four familiarization trials in Block 1, and two more in Block 2 in order to familiarize them to the new actor. For the familiarization trials, the neutral versus disbelief distinction in meaning of the test trials was maintained, but it was expressed lexically rather than intonationally (*Ah, que bé, que veiessis un gos.* 'Oh, that's nice that you saw a dog.' produced with a positive nodding head movement and *No m'ho crec, que veiessis un gos.* 'I don't believe that you saw a dog.' with a negative head shaking movement). Trials of this type were also used as fillers throughout the experiment. Therefore across the two blocks there were four filler trials in order to reorient the child were they to forget the intended meanings. For each actor, the child received six test trials and two fillers, for a total of twelve test trials and four fillers per child.

3. Analysis and Results

Two separate analyses were performed in order to account for performance on the two tasks described above. We first sought to test our predictions about children's performance for the three conditions. A total of 924 trials were analyzed. We first ran a set of mixed-effects logistic regression model in R [12] with CONDITION (three levels: AO, AV, VO), AGE (3, 4, 5[1]) and THEORY OF MIND (two levels: pass vs. not pass) as fixed effects and PARTICIPANT as a random effect. The dependent variable was CORRECT RESPONSE (correct vs. incorrect). Models were compared using ANOVAs. The best-fit model included CONDITION and AGE, but not THEORY OF MIND (pass or fail[2]). No interactions were included in the best-fit model. AGE was selected as a significant predictor of CORRECT

[1] The older 2-year-olds mentioned in 2.1 were included in the category "Age 3" and younger 6-year-olds were categorized as "Age 5".

[2] We did not include the second ToM question in the pass vs. fail decision, since all ages were shown to be at ceiling for this question.

RESPONSE. 4-year-olds performed significantly better than 3-year-olds (Estimate=1.35 vs. .38, $p<0.05$, SE=.38, z=-2.56) and 5-year-olds performed significantly better than 4-year-olds (Estimate = 2.70 vs. 1.35, $p<0.01$, SE=0.42, z=3.24). In order to investigate further any differences between conditions, a simple regression analysis was then performed. This analysis revealed differences for conditions. The slopes for the AO condition ($p<0.01$, r=0.70) and the AV condition ($p<0.01$, r=0.65) were both significant, while the slope for VO was not significant (p=0.09, r=0.39).

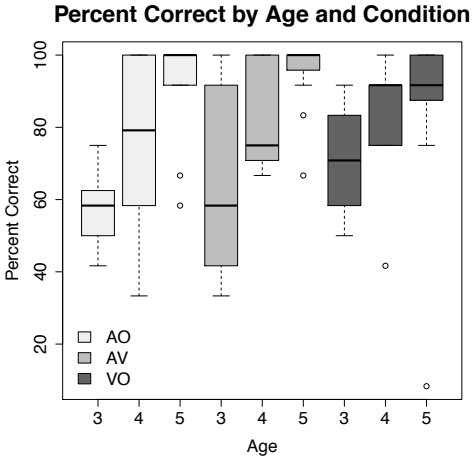

Figure 5: *Percent correct by Age for AO, AV & VO*

The fact that the slope for VO was not significant is evident if we inspect the boxplots in Figure 5, where less change between age groups is apparent for the VO condition when compared to the AO and AV conditions. Also evident in Figure 5 is the fact that the youngest children perform the worst on the AO task. 4-year-olds are better, but with a great deal of variability, suggesting that acquisition is still in progress. By Age 5, children are no longer struggling and show very little variability. A great deal of variability is observed for younger children that received the AV condition, especially for 3-year-olds. This could reflect the fact that even though children had gestural information available to reinforce the disbelief meaning in the AV condition, it may have been difficult to integrate the two cues. It is also possible that they ignored one modality because of working memory, which could have helped or hindered them depending on which one it was. Interestingly, 5-year-olds show more variability for the VO condition than for any other condition. This is perhaps because as children get older, they learn that facial cues do not always reinforce linguistic meaning.

With respect to ToM performance, Table 1 shows a general tendency observed. Across ages, children that pass the Sally Ann task tend to be more successful at the comprehension task, regardless of the condition.

Table 1. *% correct trials on linguistic comprehension task based on performance on Sally Ann task.*

Condition	Pass SA	Fail SA
AO	80	68
AV	89	73
VO	89	74

T-tests revealed that both the pass and fail groups performed at above-chance levels for all conditions. We conclude, then, based on the tendency observed in Table 1, that while explicit false belief reasoning (i.e. success on the Sally Ann task) may be helpful to children for the comprehension task, it is not a prerequisite for successful performance.

4. Discussion and Conclusions

In this experiment, each age group differed significantly from the next, showing that in general, children are making great strides in their comprehension of disbelief between the ages of 3 and 5. The simple regression analysis confirms our hypothesis that children might be better at inferring disbelief through the visual modality (significant slopes for AO and AV conditions but not VO). Inspection of the boxplot in Figure 5 reveals very different patterns for the three age groups across conditions. The amount of variability found for 3-year-olds depends on the condition. The 3-year-olds performed worst on the AO condition, but with less variability than 3-year-olds that received the AV condition, where the most variability for this age group is found. This suggests that combining the audio and visual modalities may help some children, but hinder others. The variability observed for 4-year-olds in the AO task, however, could be explained by a 'transition' stage of acquisition between the 3- and 5-year-olds. Four-year-olds are still in the process of abandoning the stage during which they rely on visual scaffolding, and are moving towards full acquisition of the linguistic meaning. For this reason, less variability is found for the 4-year-olds that received the AV condition when compared to 3-year-olds for that condition. Finally, 5-year-olds no longer require scaffolding from the visual modality, and do not seem to be thrown off by the presence of two modalities in the AV condition (there is very little variability). With respect to how explicit false belief reasoning might help children with our task, our results reveal a tendency for children that pass the Sally Ann task to perform better on the comprehension task. But across ages, those that failed the Sally Ann task still perform at above-chance levels, indicating that success on the Sally Ann task is not a prerequisite for success on the comprehension task. These results should be interpreted with caution, since our task tested *explicit* false belief judgments. Younger children may exhibit *implicit* knowledge of false belief [8]. It is obvious that some kind of mind-reading ability (i.e. ToM) must be present for children to perceive their interlocutor's state of disbelief, and this study shows that the modality through which this belief state is communicated matters. In linguistic theory, gesture is often thought to be peripheral to the study of speech. We argue that gesture is a critical component of the study of intonational development that should not be ignored.

5. Acknowledgements

We thank Page Piccinini for statistical analysis, Llorenç Andreu for help running participants, CE Jacint Verdaguer, Escola Sants Abdó, CEIP Sant Martí, the families that participated in this research and the actors, Anna and Lluís Gifra Prieto. The research was funded by a Spanish Ministry of Science and Innovation grant (FFI2012-31995 "Gestures, prosody and linguistic structure"), by a Generalitat de Catalunya grant (2009SGR-701) to the Grup d'Estudis de Prosòdia, and by the the grant RECERCAIXA 2012 for the project "Els precursors del llenguatge. Una guia TIC per a pares i educadors" awarded by Obra Social 'La Caixa'. Finally, we thank Maria del Mar Vanrell and Jill de Villiers for their helpful feedback.

6. References

[1] Crespo-Sendra, V., Kaland, C., Swerts, M. and Prieto, P., "Perceiving incredulity: The role of intonation and facial gestures", Journal of Pragmatics, 47:1-13, 2013.

[2] McNeil, N.M., Alibali, M.W., and Evans, J.L., "The role of gesture in children's comprehension of spoken language: now they need it, now they don't", Journal of Nonverbal Behavior 24(2), 131-150, 2001.

[3] Sauter, D.A., Panattoni, C., Happé, F., "Children's recognition of emotions from vocal cues", British Journal of Developmental Psychology 31: 97-113, 2013.

[4] Pierrehumbert, J., "The phonology and phonetics of English intonation". MIT PhD dissertation, 1980.

[5] Ladd, D. R., "Intonational Phonology", Cambridge University Press, 1996/2008

[6] Armstrong, M.E., "Child comprehension of intonationally-encoded disbelief", Proceedings of the 38th Boston University Conference on Language Development, accepted.

[7] Armstrong, M.E., "Puerto Rican Spanish intonation", in P. Prieto and Roseano, P., [Eds.], Transcription of intonation of the Spanish Language, Lincom EUROPA, 155-190.

[8] de Villiers, J., "The interface of language and Theory of Mind", Lingua: 117(11), 1858-1878, 2007.

[9] Baron-Cohen, S., Leslie, A. and Frith, U., "Does the autistic child have a 'Theory of Mind'?, Cognition 21: 37-46, 185.

[10] Wimmer, H., and Perner, "Beliefs about beliefs: representation and the containing function of wrong beliefs in young children's understanding of deception", Cognition: 13, 103-128, 1983.

[11] Nelson, N.L., Widen, S.C. and Russell, J.A., "The development of preschooler's Theory of Mind and emotion understanding", poster presented at the Bi-Annual Meeting of the Cognitive Development Society, 2007.

[12] Prieto, P., Borràs-Comes, J., Cabré, T., Crespo-Sendra, V., Mascaró, I., Roseano, P., Sichel-Bazin, R. and Vanrell, M., "Intonational phonology of Catalan and its dialectal varieties", in S. Frota and Prieto, P. [eds.], Intonational variation in Romance, Oxford University Press, in press, to appear in 2014.

[13] R Core Team, "R: a language and environment for statistical computing", R Foundation for Statistical Computer, Vienna, Austria, ISBN 3-900051-07-0, http://www.R-project.org/, 2013.