Journal of Experimental Child Psychology 145 (2016) 1–10

Contents lists available at ScienceDirect

Journal of Experimental Child Psychology

journal homepage: www.elsevier.com/locate/jecp

# Audiovisual alignment of co-speech gestures to speech supports word learning in 2-year-olds

Alexandra Jesse [a,*], Elizabeth K. Johnson [b]

[a] Department of Psychological and Brain Sciences, University of Massachusetts, Amherst, MA 01003, USA
[b] Department of Psychology, University of Toronto Mississauga, Mississauga, Ontario L5L 1C6, Canada

ARTICLE INFO

*Article history:*
Received 4 June 2015
Revised 8 December 2015
Available online 5 January 2016

*Keywords:*
Word learning
Referent resolution
Audiovisual perception
Language acquisition
Developmental speech perception
Multisensory perception

ABSTRACT

Analyses of caregiver–child communication suggest that an adult tends to highlight objects in a child's visual scene by moving them in a manner that is temporally aligned with the adult's speech productions. Here, we used the looking-while-listening paradigm to examine whether 25-month-olds use audiovisual temporal alignment to disambiguate and learn novel word–referent mappings in a difficult word-learning task. Videos of two equally interesting and animated novel objects were simultaneously presented to children, but the movement of only one of the objects was aligned with an accompanying object-labeling audio track. No social cues (e.g., pointing, eye gaze, touch) were available to the children because the speaker was edited out of the videos. Immediately afterward, toddlers were presented with still images of the two objects and asked to look at one or the other. Toddlers looked reliably longer to the labeled object, demonstrating their acquisition of the novel word–referent mapping. A control condition showed that children's performance was not solely due to the single unambiguous labeling that had occurred at experiment onset. We conclude that the temporal link between a speaker's utterances and the motion they imposed on the referent object helps toddlers to deduce a speaker's intended reference in a difficult word-learning scenario. In combination with our previous work, these findings suggest that intersensory redundancy is a source of information used by language users of all ages. That is, intersensory redundancy is not just a word-learning tool used by young infants.



* Corresponding author.
*E-mail address:* ajesse@psych.umass.edu (A. Jesse).

http://dx.doi.org/10.1016/j.jecp.2015.12.002
0022-0965/

## Introduction

The typical 2-year-old child expands her or his rudimentary 300-word vocabulary by about two new words a day (Fenson, Dale, Reznick, Bates, & Thal, 1994). This is particularly impressive given the ambiguity between word form and referents present in most naturalistic word-learning settings (e.g., Gillette, Gleitman, Gleitman, & Lederer, 1999; Medina, Snedeker, Trueswell, & Gleitman, 2011). How do children develop their lexicons so efficiently? Certainly the use of social cues (e.g., Baldwin, 1991) and grammatical cues (e.g., Bernal, Lidz, Millotte, & Christophe, 2007; Naigles, 1990; Paquette-Smith & Johnson, in press) aid in word learning, but what other crucially important cues might children rely on? A growing body of work has emphasized the importance of audiovisual information to guide children's attention to the correct word–object label pairings (e.g., Gogate & Hollich, 2010; Yu & Ballard, 2007; Yurovsky, Yu, & Smith, 2013). In the current study, we tested toddlers' use of a specific type of audiovisual cue to word meaning—the temporal alignment of co-speech gestures to speech.

Young infants are highly sensitive to simple cross-modal temporal relationships between sights and sounds (e.g., Bahrick, Flom, & Lickliter, 2002; Hyde, Jones, Flom, & Porter, 2011; Lewkowicz, 1996; Lewkowicz, 2003). As proposed in the intersensory redundancy hypothesis, sensitivity to these types of relationships might set the stage for word learning by helping to unify events and objects in the world with their arbitrary, but conventionalized, verbal labels (Bahrick & Lickliter, 2002). Recordings of mother–infant interactions support this view by demonstrating that mothers of young children synchronize the acoustic onset of a word referring to an object with the onset of motion they impose on the labeled object (Gogate, Bahrick, & Watson, 2000). This simple form of synchronization is most frequent in speech directed toward young infants (5–8 months) and declines as children age. The degree to which mothers produce this form of temporal synchronization is correlated with their infants' success in a word-learning task (Gogate, Bolzani, & Betancourt, 2006; Gogate, Maganti, & Bahrick, 2015). Moreover, infants as young as 2 to 7 months can learn the link between a novel label and its referent if both are presented to the children in this temporally synchronous manner (Gogate & Bahrick, 1998; Gogate & Bahrick, 2001; Gogate, Prince, & Matatyaho, 2009). Thus, temporal synchronization helps to establish a basic link between objects and their labels. However, in these latter experiments, referential ambiguity was limited because the infants were presented with only one label (spoken in isolation) and one referent at any given point in time, with the object held in the speaker's hand. And even in the more naturalistic studies recording interactions between mothers and infants, little ambiguity was present because mothers imposed motion on only one object at a time and mothers always held the intended referent. By holding the objects, mothers provided a touching cue that can guide word learning (Seidl, Tincoff, Baker, & Cristia, 2014). Thus, although it is clear that audiovisual synchrony helps infants to link a word to an object, whether or not temporal synchronization can help toddlers, and can do so in more ambiguous situations where the children must decide between multiple equally interesting moving objects to attach a label to, has yet to be explored.

Furthermore, there is no evidence that children over 1 year of age still use this synchronization. The intersensory redundancy hypothesis postulates that intersensory redundant cues, such as temporal synchrony, cease in their importance during children's second year of life when children have learned that words are arbitrarily linked to referents in the world (e.g., for a review, see Gogate, Walker-Andrews, & Bahrick, 2001). As children's reliance on audiovisual synchrony decreases with age, other cues such as eye gaze and grammatical structure become more important (e.g., Bloom, 2001). In line with this development, caregivers indeed stop producing temporal synchronization between motion and labels as their infants become toddlers (Gogate et al., 2000, 2001). However, speakers might not eliminate alignment but rather change the way in which it is realized. We recently showed that adults describing objects to 2-year-olds tended to align the movements they imposed on the objects they were referring to with the prosodic structure (i.e., rhythm, intonation, and/or stress; Jesse & Johnson, 2012). In that same study, we demonstrated that adults are sensitive to this prosodic alignment, using it to work out a speaker's referential intent in an otherwise ambiguous labeling scenario.