# Audiovisual Alignment in Child-Directed Speech Facilitates Word Learning

*Alexandra Jesse* [1], *Elizabeth K. Johnson* [2]

[1] Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
[2] Department of Psychology, University of Toronto, Toronto, Canada
Alexandra.Jesse@mpi.nl, Elizabeth.Johnson@utoronto.ca

## Abstract

Adult-to-child interactions are often characterized by prosodically-exaggerated speech accompanied by visually captivating co-speech gestures. In a series of adult studies, we have shown that these gestures are linked in a sophisticated manner to the prosodic structure of adults' utterances. In the current study, we use the Preferential Looking Paradigm to demonstrate that two-year-olds can use the alignment of these gestures to speech to deduce the meaning of words.
**Index Terms**: speech perception, audiovisual alignment, word learning

## 1. Introduction

When learning their native language, children must not only learn the sounds, words, and rules of their native language, but also the relationship between labels and the world. For a child to infer the correct reference when hearing the label *cat*, the child has to realize that the label refers to the object "cat", and not, for example, to another object, action, state, or abstract concept. In addition, the child needs to learn whether *cat* is the label for a particular object, its whole class, or a subpart or feature of the object. Despite these difficulties children attach labels to objects in a rapid and seemingly effortless manner. By 30 months, the average child has a productive vocabulary of approximately 550 words [1]. Receptive vocabularies are undoubtedly far larger at this age [2].

Trying to decipher how children deduce the meanings of words in such an efficient way is an active area of study. Children have many different types of strategies for working out the referent of new words, and the types of strategies they rely on change over the course of development [see 3, for a review]. Some strategies take advantage of information in the auditory speech signal whereas others are based on information extracted from the visual scene. Ten-month-olds use visual perceptual salience to attach meanings to words, i.e. when a novel word is uttered, it is attached as a label to the most perceptually exciting and novel object in the visual environment [4]. Later in development, children begin exploiting syntactic information to narrow down the potential meaning of a word [5]. Note that just as the strategies that children use to attach meanings to words change over the course of their development, so too does the apparent depth of word learning that children engage in. Six- to eight-month-olds, for example, most likely begin by simply detecting mere word-object associations [6, 7]. Older children, in contrast, appear to possess a much more sophisticated level of comprehension [8]. Children learn, for example, to extend a label to other members of the same category even if these members are perceptually dissimilar [e.g., 9].

During their early preverbal stage of development when infants are first detecting word-object associations, they also make use of audiovisual temporal relations to extract word meanings [6, 10, 11]. As shown with the switch procedure, seven-month-old infants learn to associate an object with a single-vowel utterance, when the motion of the object starts and ends simultaneously with the onset and offset of the utterance (*temporal synchrony*), but not when the object moves in asynchrony with the label or does not move at all [6]. Children remember these learned associations for at least four days [10]. Analyses of mothers' teaching behavior in a word-learning setting suggest that this audiovisual temporal synchrony of labeling and object motion is indeed produced by mothers and done so more often for the words intended to be taught than for others [12]. This simple form of temporal audiovisual synchrony is therefore available as a cue to the child in a naturalistic word learning setting. Furthermore, the degree to which this audiovisual linking of onset and offset of labeling and motion is produced by the mothers correlates with their six- to eight-month-old infants' success in learning word-object associations [11].

Children's use of temporal synchrony can be interpreted within the theory of *intersensory redundancy* [13]. The theory assumes that early in development a child is more sensitive to the salient intersensory redundant information than unimodally presented information. For the word learning setting, this means that audiovisual temporal synchrony as a form of intersensory redundancy increases the object's perceptual salience and focuses the child's attention on the relationship between the visual event (object) and auditory event (label) rather than on their unimodal properties. This form of perceptual salience consequently facilitates the learning of intermodal word-object relationships. Intersensory redundant information, such as temporal synchrony, also bootstraps an understanding of the unitary nature of multisensory events. With the development of this understanding of unity, the child can then later during infancy also learn unimodal properties of multisensory events. According to this theory, intersensory redundancy loses its importance with age while other cues become important [13]. This view is supported by the finding that mothers decrease their use of this type of temporal synchrony as their children mature, i.e. mothers decrease their use of linking the onset and offset of co-speech gestures to labeling as their children's lexical competency increases [12]. While mothers of preverbal infants (i.e., 5 to 8 months) use this strategy more often than mothers of children in an early lexical acquisition period (i.e., 9 to 17 months), mothers of older children (i.e., 21 to 30 months) display even less of the type of temporal synchrony between labeling and gestures studied by Gogate and her colleagues. Mothers of these older children engage in more interactions were the child holds and manipulates the object [12]. In addition, the use of a mother's eye gaze to guide her child's attention increases with the child's age [14]. Around the age of 18 months, joint attention with the speaker becomes

critical. For example, even though children are already able to exploit social cues, such as eye gaze, to map a label to an object at an earlier age [e.g., 15], children at a later age rely on social cues even if the labeled object is not the perceptually most salient object [16, 3].

As mentioned above, studies have indicated that as children mature, their caretakers decrease their use of co-speech gestures that temporally link to their labeling utterances in a simplistic onset/offset fashion. However, it is also possible that Gogate and her colleagues may have observed the evolution rather than the reduction of temporal synchrony in child- compared to infant-directed speech. Other types of potentially more subtle and complex alignments of co-speech gestures to speech may likely persist, or even increase in use, as children mature. Some evidence for this hypothesis comes from analyses of adult-to-adult interactions that suggest that co-speech gestures may be linked to the prosodic structure of speech and that this alignment is informative in adult communication [17, 18, 19]. We therefore hypothesized that only the use of a more simple alignment of co-speech gesture and speech in child-directed speech may decrease as children mature.

In line with these predictions, we have shown in a series of perceptual experiments with adults, that there is a perceptible relationship between the prosodic structure of child-directed speech and its accompanying gestures. Our evidence suggests that toddler-directed communications contain co-speech gestures that appear to be linked in a complex manner to the prosodic structure of the utterance [20]. For this series of studies, eight female Dutch speakers were video recorded as they attempted to teach 24-month-old children the proper names of novel creatures (see Figure 1). As speakers labeled and described the toys, they naturally gestured with the toys they were labeling. The motion trajectories of the toys in the recordings were used to animate photographs of the toys. Hence, in these animations, the speakers were not visible. These animated pictures of the toys were then used to test whether there was a perceptible relationship between the speaker's utterances and the way they moved the toy they were labeling. In order to test this, a reversed version of each animated video was created, and paired up with its corresponding forward version. Adult participants saw these forward and reversed versions of the same animated picture side-by-side, and were asked to indicate which creature the speaker was referring to. Note that the content of the audio track was not informative in this regard. Listeners performed well in this task. Importantly, they performed equally well when the speech was low-pass filtered. Low-pass filtering destroys the phonetic content and only leaves prosodic information intact [21, 22]. Simple cross-modal simultaneity of onsets/offsets or of rate changes cannot sufficiently explain our results. When the audio track was reversed and therefore the backwards played competitor video became the target, performance dropped. If temporal synchrony between the onset and offset of words were sufficient to explain participants' performance, then participants should have scored equally well in the backwards version of the task as in the original version. In summary, these results suggest that in a word-learning setting, adults produce motion that is aligned with the prosodic structure of their utterances and not simply with the onset and offset of the label. Adults are sensitive to this cross-modal alignment and at least in a laboratory setting, use it to resolve potential referential ambiguities.

In the present study, we used the Preferential Looking Paradigm to investigate whether 24- to 26-month-old infants are like adults in that they can use cross-modal alignment to detect speaker intent. At the same time, we also tested whether the detection of speaker intent through co-speech gestures is used by children to attach labels to novel objects. Note that this latter step is important. There is substantial evidence that both children and adults are sensitive to intersensory redundancies [see 13, for an overview]. However, there is at present no evidence that we are aware of that toddlers use this information to work out referential ambiguities in a word learning setting. A subset of the experimental materials presented in the adult perceptual studies was used as training trials. In the Preferential Looking Paradigm [e.g., 23], learning is assessed by presenting children with two objects and asking them to look at one or the other. Children's looking times to the target or distractor were evaluated as a measure of word learning. If the child succeeded in learning the label-object relationship during training, then the infant would look longer toward the target than the distractor object. To encourage children to enter a word learning mode, we also included two familiarization trials at the beginning of the experiment. During these introductory trials, the creatures were unambiguously labeled once. We predicted that labeling the object once would be unlikely to induce word learning. However, in order to ensure that a single labeling was not sufficient for explaining any learning we observe in the test phase, we included a control group. In the control group, the creatures were named incorrectly during the two initial familiarization trials, i.e., the label-object relationships given during familiarization were inconsistent with the ones to be learned during training. We predicted that if child-directed co-speech gestures help to focus children's attention on the correct referent for new word and thereby facilitate word learning, then children in the experimental group should look longer to the target during the test phase. At the same time, we predicted that those children in the control group should fail to show any looking time preferences during the test phase because they have received conflicting information about the names of the creatures. As an added control to ensure that there were no differences between our experimental and control populations, we also included trials where children were asked to find familiar words (e.g., *dog*). Since children in both the control and experimental group were tested on the same familiar words, this allowed us to directly compare how similar our two populations were with regard to their language skills. Given that children were randomly assigned to the experimental and control conditions, we predicted that both control and experimental participants should perform identically on these familiar word trials.

# 2. Experiment

## 2.1 Participants

Forty-six Dutch-learning toddlers were tested (24 females, 22 males). Their average age at testing was 108 weeks old (ranging from 24 months and 10 days to 25 months and 28 days). All children were monolingual native Dutch learners. Data from eleven additional children (four in the experimental and seven in the control group) were excluded from the experiment due to fussiness (4), parental interference (2), experimenter error (2), or because the child did not complete

the experiment (3). Participants received ten euro or a book as a thank you for taking part in the study.

## 2.2 Stimuli

Three female Dutch-speakers were video recorded as they taught two novel names of toy creatures to a video of two-year-olds. Two novel proper names that were taught were *kag* ([kɑx]) and *zeut* ([zøt]). These items are phonotactically legal nonwords in Dutch. Speakers were given one toy creature to hold (see Figure 1 for picture of a typical recording session). Toys were novel creatures. Speakers were shown silent videos of two-year-olds watching TV and were instructed to try to teach the names of the toys to the children. To encourage speakers to use a lively attention-getting voice, they were instructed to think of themselves as being situated in a distracting environment. Speakers were naive with regard to the purpose of the study. Recordings were not scripted. Speakers were, however, told not to refer to features of the toys that would aid in discriminating them (e.g., their color). Furthermore, speakers never referred to an action they imposed on the toy. The video shown to the speaker to elicit child-directed speech consisted of six 20-second clips of children watching TV. These clips were specifically chosen because the children were distracted and not fully engaged by the TV show they were watching. We chose to use clips where children looked distracted in order to encourage speakers to try to get the children's attention while labeling the creature. We used a video of two-year-olds rather than live two-year-olds to elicit child-directed utterances from our speakers because producing recordings with live infants in a controlled way was not feasible. In addition, using silent videos enabled us to obtain recordings where only the speaker is audible.

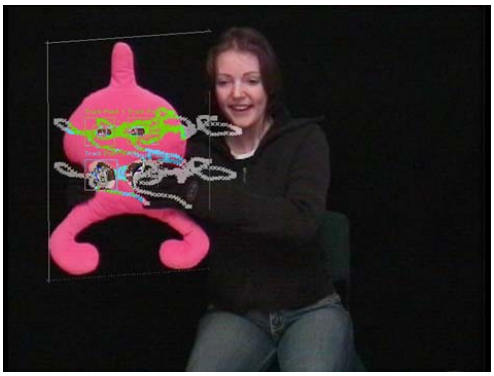

Figure 1: *Example of the set-up of a typical recording session. Dots show the motion path of three tracking points over time. A fourth tracking point is inferred by the software to form a parallelogram (see larger rectangle).*

Videos were digitized as uncompressed avi files and the motion paths of the creatures were extracted using Adobe After Effect Professional 6.5. We tracked two stickers that had been attached to the toy above its eyes as well as one of the eyes with the Parallel Corner Pin method. This method estimates a fourth point so that a rectangle connecting all four points consists of parallel lines. This tracking method skews, scales, and rotates the object, but does not estimate perspective. Rather, it assumes relative distance to be constant. Figure 1 shows an example of the four tracking points and their motion paths over time. Motion trajectories were then applied to photos of the toys. That is, the speaker was no longer visible in the final materials. A second version of each animation was created with the motion trajectory applied reversed in time. These versions served as competitor items alongside the targets. This method of creating competitor videos was chosen since it controls for the amount of overall motion of target and competitor. An object that moves more may be inherently more interesting to the children than one that moves less and therefore this difference in overall motion could have introduced a bias. Each animation was done once with each of the two creatures as target. A respective target and its competitor video were then pasted side-by-side (see Figure 2 for an example) and saved along with the original soundtrack as *training trials*. All combinations of color (green or pink), item status (target or competitor), and side (left or right) were produced in order to control these variables across participants.

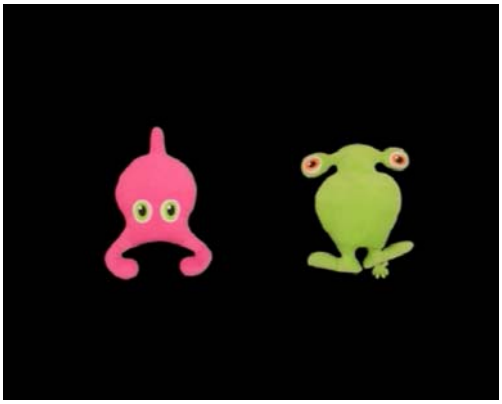

Figure 2: *Screenshot of a typical video frame in the training and test trials of the experiment.*

Six test trials were created (see Figure 2). On test trials, both creatures grew in size very slowly in an identical fashion, i.e. the movement of both creatures was not matched to the accompanying soundtrack. Test trial soundtracks consisted of the same sentence asking in a child-directed manner to find one of the toys (*"Kijk eens, kan je ___ vinden?*", "*Look, can you find ___?*"). This sentence was followed by a second simple sentence (e.g., *" Wat is ie mooi.", "How cute he is!*"). Furthermore, additional filler test trials with pairs of objects familiar to Dutch children at this age were created (dog, boat, fish, bike). In addition, *familiarization trials* were created in which only one of the creatures was shown but not animated. This was accompanied by a simple naming of the toy, following the format "*Dit is ___. Vind je ___ mooi?*" ("*This is ___. Do you find ___ cute?*"). All videos in the experiment were separated by a video of a zooming smiley or a star, accompanied by a tone.

Table 1. *Overview of trials in the experiment.*

| | Object familiarity | Number of trials | Number of objects present on a trial |
|---|---|---|---|
| Test | familiar | 2 | 2 |
| | novel | - | - |
| Familiarization | familiar | - | - |
| | novel | 2 | 1 |
| Training | familiar | - | - |
| | novel | 6 | 2 |
| Test | familiar | 2 | 2 |
| | novel | 6 | 2 |

## 2.3 Procedure and design

Figure 3 shows the typical experimental set-up for this study. Children sat on a caregiver's lap during the experiment. Caregivers listened to music over Sennheiser Noiseguard headphones during the experiment to prevent the potential introduction of parental bias. All stimuli were presented on a 192cm Sony LCD TV screen with built-in speakers. The screen was approximately 1 m away from the child. Light in the testing booth was dimmed. Children's eye movements were recorded to digital video for subsequent offline coding.

Figure 3: *Photograph of an example of the set-up during the experiment. Lights are not shown as dimmed here as was the case during the real experiment.*

Participants were randomly assigned to one of two conditions: experimental or control group. The experiment always began with a video clip showing a zooming animation of both creatures accompanied by the remark "*Wat mooi, zeg!*" ("*How cute!*"). The experiment then continued with two trials with two familiar objects (dog, boat, fish, bike) shown on each trial. Here, the child was asked to look at one of the objects on each trial. The two objects the child was not asked to look at during this phase of the experiment served as targets in the test trials later. Subsequently, two familiarization trials were given where each novel toy creature was presented and labeled individually once. For the experimental group, this familiarization was consistent with what should be learned during training. For the control group, the label-object mapping during familiarization was inconsistent with the one to be learned during training. This was the only difference in the procedure between the two groups. Following familiarization, both groups were presented with six 20-second training trials during which the speakers labeled one of the two side-by-side animated toys. That is, each child was trained for each word once by each of the three speakers. On these trials, only one of the creatures followed the motion trajectory as originally produced with the utterance. Finally, children were presented with six test trials of side-by-side pictures of the two toys and asked to look at either *kag* or *zeut*. Likewise, children were also presented with two test trials of picture pairs of the familiar objects. The experiment lasted approximately four minutes.

The presentation order of familiarization and test trials was completely counterbalanced across participants. The presentation order of training trials was random for each participant in the experimental group. However, the same orders were then also used for the control group. Word-object assignment was counterbalanced across participants. Also, the creature referred to by the speaker was half of the training trials on each side.

## 2.4 Coding

Mean proportion of looks to target during test trials was calculated for each infant during a two-second time window, starting 400 ms after target word onset. Looks to target preceding this time window were not analyzed since it takes some time for children to program an eye movement [23]. Videos were digitized and then hand-coded offline. Coding was done with the audio track disabled. Onsets and offsets of test trials were marked in the video by lighting changes, since test trials had black backgrounds and inter-trial screens were white. The coder was unaware of the condition a child was tested under. For each 40ms frame in the coding window, the participant's eye gaze was either coded as look to the left or the right or as neither. Proportions of looks to target were then calculated as number of looks to the correct target side divided by the total number of looks to either side. These proportions were then averaged over frames for each trial. Looks that were coded as "neither" were not considered in this measure. Chance performance is therefore at a proportion level of .50.

## 2.5 Results and discussion

Figure 4 shows performance for both experimental and control group for the trained novel words. A one-sample t-test on mean proportions of looks to target compared performance in the experimental group to the chance level. Children in the experimental group looked significantly longer to the correct creature than predicted by chance ($M$=.59, $SD$=.15, $t(21)$=2.81, $p$<.011), indicating that children successfully learned the creatures' names. Critically, this was not solely due to the initial familiarization with displays of single creatures, since children in the control group showed no looking preference ($M$=.48, $SD$=.13, $t(23)$=-.59, $p$=.56). In addition, the looking preference for the correct target in the experimental group was larger than in the control group ($t(44)$=2.58, $p$<.013).

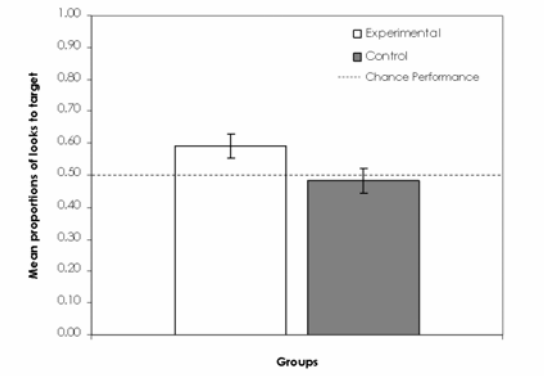


Figure 4: *Mean proportion of fixation time to target on novel toy trials during two second analysis window (chance performance = .5).*

As Figure 5 shows, performance for familiar words was above chance for both experimental ($M$=.68, $SD$=.15, $t(21)$=5.48, $p$<.001) and control group ($M$=.67, $SD$=.11, $t(23)$=7.45, $p$<.001). More importantly, an independent t-test showed that both groups performed equally well on the four familiar word trials ($t(44)$=.24, $p$=.81). This indicates that differences in performance between the two groups on test trials cannot be explained by differences between the two groups' general language comprehension abilities.

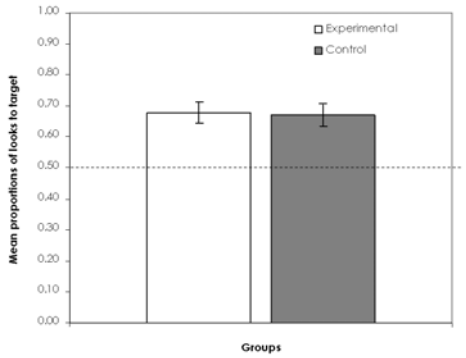


Figure 5: *Mean proportion of fixation time to target on familiar word trials during two second analysis window (chance performance = .5).*

## 3. General Discussion

Infants and young children learn their native language in an environment rife with complex intersensory stimulation. Recent work has shown that intersensory redundancies can change the way very young infants perceive the world [24], demonstrating that this type of information has a strong impact on early information processing. Intersensory redundancies also impact language processing. For example, in a laboratory setting, preverbal seven-month-olds learned word-object associations only when very salient artificially produced intersensory correlations existed between the acoustic onset and/or offset of a verbal label and the movement of the labeled object [6]. In the current paper we hypothesized that the use of intersensory redundancies to learn object labels persists into toddlerhood. Indeed, we argued that child-directed speech contains an exaggerated version of the same type of intersensory redundancies that appear to be present in adult-directed speech, and toddlers are like adults in that they can use this information as an aid to understand the communicative intent of their adult interlocutors. The results of the current study support this notion. When presented with videos showing two animated creatures, where the motion of only one matches what was originally produced along with the labeling soundtrack, toddlers used this intersensory information as a cue to determine speaker intent and consequently learned the label-object relationship.

Our word learning results dovetail nicely with those previously reported in the literature. Gogate and colleagues [11] found that very young preverbal seven-month-olds were increasingly likely to associate a verbal label with an object the more their mothers synchronized the onset and offset of verbal labels with object motion. Our results are compatible with this finding in that we have shown that linguistically savvy 24-month-olds still use intersensory redundancies to determine speaker labeling intent. Moreover, they do so in a difficult word learning task in which two moving creatures were presented at the same time with no information other than intersensory redundancies available to determine which creature the speaker was labeling. Despite the compatibility of our empirical findings with other reports in the field, our theoretical interpretation of children's use of intersensory alignment diverges from that previously published in the literature. According to the intersensory redundancy hypothesis, young infants are particularly sensitive to intersensory redundancies in the environment. This leads infants to pay close attention to events defined by such relationships. As infants mature, these intersensory redundancies are thought to decrease in saliency. In support of this notion, Gogate and colleagues have shown that as children become increasingly linguistically mature, mothers reduce their production of co-speech gestures correlated with the onset or offset of target word utterances [12]. However, we have shown that at 24 months, children are still very sensitive to intersensory redundancies.

What implications, if any, do our findings have for the intersensory redundancy hypothesis? One possibility is that intersensory redundancies continue to play an important role in facilitating verbal communication throughout the lifespan. In other words, there is a continuity between the type of information used in toddlerhood and that used in adulthood. Note that this is not necessarily contradicting the intersensory redundancy hypothesis. Rather, this continued use of intersensory information could exist alongside an earlier over-reliance on such cues. Support for the notion that intersensory redundancies are still important comes from studies on adult language processing demonstrating enhanced comprehension in the presence of such information [18]. But if intersensory redundancies are so important for mature language communication, then why did Gogate and colleagues observe a decline in the use of child-directed co-speech gestures as children matured? We would like to propose that certain types of intersensory alignment patterns may decline in frequency, whereas others either remain level or may even increase in frequency as children mature. Gogate and colleagues defined intersensory synchrony as simple temporal co-occurrence of the label and the gesture event [e.g., 12]. However, not all co-speech gestures need to be aligned in such a simple way with the onset or offset of a word in order to be informative. Indeed, the results of our adult perceptual experiments suggest that child-directed speech and its co-speech gestures are linked in a more complex manner. We propose that co-speech gestures in the word learning setting are most likely linked to the prosodic structure of utterances. The fact that both co-speech gestures as well as speech prosody tend to be exaggerated in child-directed speech [25, 14, 26, 27] leaves room for the possibility that the intersensory relationship between gestures and prosody are even more salient in child-directed speech than adult-directed speech.

## 4. Conclusions

These results represent the first experimental demonstration that two-year-olds use the alignment between naturally-produced child-directed speech and its accompanying gestures to determine intended word meaning. Children used the multisensory alignment to determine which of two competing moving objects the speakers referred to and consequently learned the label-to-object mapping. It is striking that infants succeeded in this mapping task despite the absence of other word learning cues (e.g., eye gaze). Our findings are compatible with previous findings in this area with far younger infants; however, it may be that this earlier work underestimated the role of intersensory redundancies in later linguistic development because it did not take into account the complexity and wide variety of realizations of multisensory alignment. In the future, in order to further understand the role of intersensory redundancies across development, it will be important to determine to what degree two-year-olds rely on this information in a more naturalistic setting when multiple cues to labeling intent are present.

## 5. Acknowledgements

This work was supported in part by a grant from the German Science Foundation to AJ, funding from the University of Toronto and Max Planck Institute for Psycholinguistics to EJ, and a Spinoza grant to Anne Cutler. The authors thank Angela Khadar and the rest of the Nijmegen Babylab Crew for assistance in recruiting, testing, and coding participants.

## 6. References

[1] Fenson, L., Dale, P., Reznick, S., Bates, E., Thal, D. and Pethick, S., "Variability in early communicative development", Monographs of the Society for Res. in Child Develop., 59 (Serial No. 242), 1994.

[2] Jusczyk, P., "The discovery of spoken language", Cambridge, MA: MIT Press, 1997.

[3] Hollich, G., Hirsh-Pasek, K. and Golinkoff, R., "Breaking the language barrier: An emergentist coalition model for the origins of word learning", Monographs for the Society for Res. in Child Develop., 65, (Serial No. 262), 2000.

[4] Pruden, S. M., Hirsh-Pasek, K., Golinkoff, R. M. and Hennon, E. A., "The birth of words: Ten-month-olds learn words through perceptual salience", Child Develop., 77: 266-280, 2006.

[5] Fisher, C., Hall, D. G., Rakowitz, S. and Gleitman, L. R., "When it is better to receive than to give: Syntactic and conceptual constraints on vocabulary growth", Lingua, 92: 333-375, 1994.

[6] Gogate, L.J. and Bahrick, L.E., "Intersensory redundancy facilitates learning of arbitrary relations between vowel sounds and objects in seven-month-old infants", J. of Exp. Child Psychol., 69:1-17, 1998.

[7] Tincoff, R. and Jusczyk, P. W., "Some beginnings of word comprehension in six-month-olds", Psychol. Sci., 10: 172-175, 1999.

[8] Golinkoff, R. M., Mervis, C. B. and Hirsh-Pasek, K., "Early object labels: The case for a developmental lexical principles framework", J. Child Lang., 21: 125-155, 1994.

[9] Gelman, S. A., "Young children's inductions from natural kinds: The role of categories and appearances", Child Develop., 58: 1532-1540, 1987.

[10] Gogate, L. and Bahrick, L.E., "Intersensory redundancy and seven-month-old infants' memory for arbitrary syllable-object relations", Infancy, 2: 219- 231, 2001.

[11] Gogate, L.J., Bolzani, L.H. and Betancourt, E.A., "Attention to maternal multimodal naming by 6- to 8-month-old infants and learning of word-object relations", Infancy, 9:259-288, 2006.

[12] Gogate, L., Bahrick, L.E. and Watson, J.D., "A study of multimodal motherese: The role of temporal synchrony between verbal labels and gestures", Child Develop., 71: 878- 894, 2000.

[13] Bahrick, L.E. and Lickliter, R., "Intersensory redundancy guides early perceptual and cognitive development", in R. Kail [Ed], Advances in child development and behavior, 30:153-187, Academic Press, 2002.

[14] Brand, R.J., Shallcross, W.L., Sabatos, M.G. and Massie, K.P., "Fine-grained analysis of motionese: Eye gaze, object exchanges, and action units in infant-versus adult-directed action". Infancy, 11:203-214, 2007.

[15] Woodward, A. L., "Infants' use of action knowledge to get a grasp on words", in D. G. Hall and S. R. Waxman [Eds], Weaving a lexicon, 149-172, MIT Press, 2004.

[16] Baldwin, D.A., "Infants' contribution to the achievement of joint reference", Child Develop., 62: 875-890, 1991.

[17] Hadar, U., Steiner, T.J., Grant, E.C. and Rose, F.C., "Head movement correlates of juncture and stress at sentence level", Lang. Speech, 26:117–129, 1983.

[18] Munhall, K.G., Jones, J.A., Callan, D. Kuratate, T. and Vatikiotis-Bateson, E., "Visual prosody and speech intelligibility: Head movement improves auditory speech perception", Psychol. Sci., 15:133-137, 2003.

[19] Yasinnik, Y., Renwick, M. and Shattuck-Hufnagel, S., "The timing of speech-accompanying gestures with respect to prosody", From Sound to Sense: 50+ Years of Discoveries in Speech Comm., 11-13 June 2004, Cambridge, MA. Online: http://www.rle.mit.edu/soundtosense/conference/pdfs/fulltext/Friday%20Posters/FA-Yasinnik-STS-MAC.pdf, accessed on 2 June, 2008.

[20] Jesse, A. and Johnson, E., "Audiovisual alignment in child-directed speech facilitates the detection of speaker intent in a word learning setting", Proc. 50th Conf. of Exp. Psychologists, Marburg, Germany, 2008.

[21] Grant, K. W. and Walden, B. E., "Spectral distribution of. prosodic information." J. Speech Hearing Res., 39: 228-.238, 1996.

[22] Pollack, I., "Effects of high pass and low pass filtering on the intelligibility of speech in noise", J. Acoust. Soc. Am., 20:259-266, 1948.

[23] Swingley, D. and Aslin, R.N., "Spoken word recognition and lexical representation in very young children", Cognition, 76: 147-166, 2000.

[24] Phillips-Silver, J. and Trainor, L.J., "Feeling the beat: movement influences infants' rhythm perception", Science, 308:1430, 2005.

[25] Brand, R. J., Baldwin, D. A. and Ashburn, L. A.., "Evidence for "motionese": Modifications in mothers' infant-directed action", Develop. Sci., 5:72-83, 2002.

[26] Fernald, A., "Intonation and communication intent in mother's speech to infants: is the melody the message?", Child Develop., 60:1497–1510, 1989.

[27] Fernald, A. and Mazzie, C., "Prosody and focus in speech to infants and adults", Develop. Psychol., 27: 209–221, 1991.