Infants' ability to extract verbs from continuous speech

Ellen Marklund and Francisco Lacerda

Department of Linguistics
Stockholm University, Stockholm, Sweden
ellen@ling.su.se

10.21437/Interspeech.2006-410

ABSTRACT

Early language acquisition is a result of the infant's general associative and memory processes in combination with its' ecological surroundings. Extracting a part of a continuous speech signal and associate it to for instance an object, is made possible by the structure provided by characteristically repetitive Infant Directed Speech. The parents adjust the way they speak to their infant based on the response they are given, which in turn is dependent on the infant's age and cognitive development.

It seems probable that the ability to extract lexical candidates referring to visually presented actions is developed at a later stage than the ability to extract lexical candidates referring to visually presented objects – actions are more abstract and there is a time aspect involved.

Using the Visual Preference Paradigm, the ability to extract lexical candidates referring to actions was studied in infants at the age of 4 to 8 months. The results suggest that while the ability at this age is not greatly apparent, it seems to increase slightly with age.

1. INTRODUCTION

This study is a part of the MILLE-project (Modeling Interactive Language Learning), which is a collaboration between the Department of Linguistics at Stockholm University (Sweden), the Department of Speech Music and Hearing at the Royal Institute of Technology (Sweden) and the Department of Psychology at Carnegie Mellon University (PA, USA). The project's aim is to investigate the processes behind early language acquisition and to implement them in computational models.

1.1 Infant directed speech

According to the Ecological Theory of Language Acquisition (ETLA) early language acquisition can be seen as an emergent consequence of the multi-modal interaction between the infant and its surroundings. Parents verbally interpret the world to their infant, and by doing so they provide the structure necessary for linguistic references and systems to emerge [1].

Infant Directed Speech (IDS) differs from Adult Directed Speech (ADS) mainly in that it is more repetitive, and has greater pitch variations. IDS varies with the age of the infant, as the parent adjusts to the cognitive and communicative development of the infant and adapts the language accordingly [2].

At a very early age the main goal of speaking is general social interaction – to maintain the communications link with the infant, as opposed to ADS and later IDS, where the goal mainly is to convey information of some sort [3]. Recent studies at our lab shows that early IDS is also highly focused on what's in the field-of-view of the infant; if the infant's attention shifts, the parent tend to follow the infant's lead and focus on the new center of attention (report in preparation).

1.2 Modeling early language acquisition

The principles of ETLA will be implemented in a humanoid robot, in an effort to develop a system that acquires language in a manner comparable to human language acquisition.

The system's learning will at some stage be dependent on IDS input, and thus some of the questions that need to be answered in order to accomplish this are in which way the parent adjusts the IDS with the infant's age, and on what cues those adjustments are based? To answer those questions, the infant's word-learning behavior needs to be understood.

1.3 Learning lexical labels

To a naive listener – for example an infant – speech is not perceived as a combination of words building up sentences, but rather as a continuous stream of sounds with no apparent boundaries. According to ETLA early word learning is based on general multi-sensory association and memory processes, which in combination with the structured characteristics of IDS provides opportunities for lexical candidates – smaller chunks of speech, referring to for example an object – to emerge and be extracted from a continuous speech signal.

Studies have shown that infants at an early age are able to integrate multi-modal information and extract lexical candidates from a flow of continuous speech, and learn to associate those with objects [4], and other studies show that they are capable of extracting and associating lexical candidates with attributes, such as colors and shapes [5].

Regarding lexical candidates referring to actions or events – verb-like lexical candidates, Echols has, by directing infants' attention with labeled objects or events, shown that while infants of 9 months generally focus more

on objects when an event is labeled, while older infants to a greater extent focused on events or actions [6].

The goal of this experiment is to study how infants' ability to associate visually perceived actions with linguistic references differs with age, in order to gain further understanding of the development of general cognitive processes and abilities essential for language acquisition.

2. METHOD

The method used (Visual Preference Procedure) is a variation of the Intermodal Preferential Looking Paradigm; a method originally developed by Spelke [7]. The subjects were shown a film with corresponding visual and auditory stimuli while their eye-movements were recorded. The looking time for each sequence of the film and for each quadrant of the screen was calculated and analysed using Mathematica 5.0 and SPSS 14.0.

2.1 Subjects

The subjects were 55 Swedish infants between 4 and 8 months old, randomly selected from the National Swedish address register (SPAR) on the basis of age and geographical criteria.

During the recording of the eye-movements, the infant was seated in the parent's lap or in an infant chair close to the parent. The parent wore Active Noise Reduction headphones, listening to music.

The 23 infants whose total looking time did not exceed 30% of the duration of the film were not considered sufficiently exposed to the stimuli; hence those recordings were excluded leaving a total of 32 recording sessions.

2.2 Stimuli

The auditory stimuli of the films were non-sense words read by a Swedish female and concatenated to three-word phrases. The films had three sequences each; a baseline, an exposure and a test. The baseline showed a split-screen with four different scenes in which cartoon figures were moving towards or away from each other in silence (Figure 1).

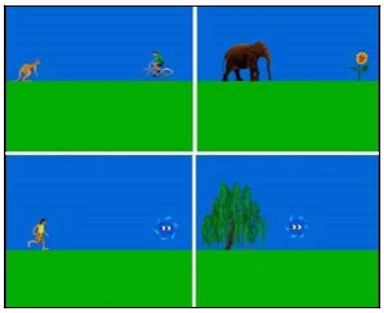

Figure 1: In each scene of the baseline one cartoon figure was moving towards or away from an other, stationary figure.

The exposure phase showed eight subsequent scenes similar to those in the baseline, where the direction of the target movements was balanced regarding actor and direction on the screen (Figure 2). In each of the scenes only two cartoon figures were involved – a kangaroo (K) and an elephant (E).

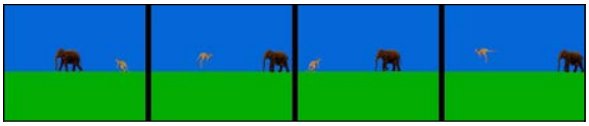

Figure 2: In the eight subsequent scenes of the exposure, only two cartoon figures were involved – a kangaroo and an elephant.

During the exposure, three-word non-sense phrases were played referring to the objects and actions involved. The target word, the one referring to the action, always held the medial position (Table 1). The phrases were repeated three times during each scene.

Scene	Audio film 1	Audio film 2	Video	Movement
1	*nima bysa pobi*	*fugga pobi bysa*	←E K	Away
2	*pobi fugga nima*	*bysa nima fugga*	K→ E	towards
3	*nima bysa pobi*	*fugga pobi bysa*	K E→	away
4	*nima fugga pobi*	*fugga nima bysa*	K ←E	towards
5	*nima fugga pobi*	*fugga nima bysa*	E→ K	towards
6	*pobi bysa nima*	*bysa pobi fugga*	←K E	away
7	*pobi fugga nima*	*bysa nima fugga*	E ←K	towards
8	*pobi bysa nima*	*bysa pobi fugga*	K E→	away

Table 1: The auditory stimuli and the visual balancing of the eight scenes in the exposure phase.

The test was visually identical to the baseline, while the non-sense word referring to the action was repeated twice (Table 2).

Film	Audio	Target
1	*bysa*	away
2	*nima*	towards

Table 2: Auditory stimuli of the test phase; the word referring to the target movement was repeated twice.

2.3 Eye-tracking

A Tobii Eye-Tracker integrated with a 17" TFT monitor was used to record the eye-movements of the infants. Near infra-red light-emitting diods (NIR-LEDs) and a high-resolution camera with a large field-of-view are mounted on the screen. The light from the NIR-LEDs is reflected on the surface of the subject's eye, which is captured by the camera, as is the position of the pupil. During a calibration the subject is encouraged to look at pre-specified points at the screen while the system measures the pupils position in relation to the reflections, providing data for the system to use in order to calculate the gaze point on the screen during the recording. Software used for data storage was ClearView 2.5.1.

3. RESULTS AND DISCUSSION

The results showed a slight increase with age in looking time at the target compared to the non-target (Figure 3). Regression coefficient is R=0,261, p<0,148.

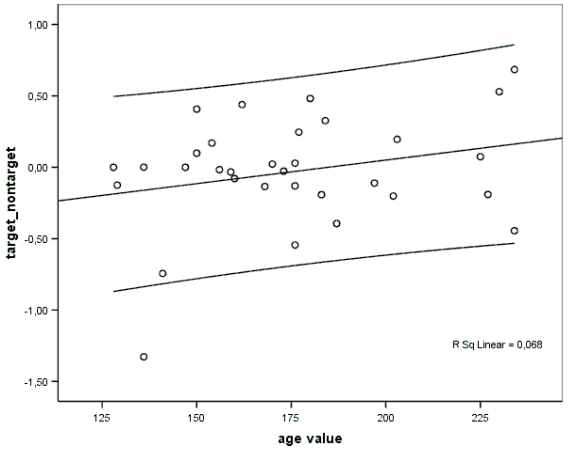

Figure 3: The y-axis shows the normalized difference in gain in percentage of looking time towards the target with non-target as baseline. The x-axis is age in days.

The results suggest that the infants' ability to extract a part of a continuous stream of speech and associate it with an action increases over age.

To expand this study to encompass older infants would be of interest in order to determine if there is a period where the ability increases drastically; hypothetically between 9 and 13 months of age, which would be in accordance with Echols' results [6] and might indicate a leap in cognitive development.

One could also consider a replicating this study but using less complex actions as target, so the infant doesn't have to grasp the concept of changes in spatial relationship between different agents in order to understand the event.

4. ACKNOWLEDGEMENTS

The research was financed by grants from The Bank of Sweden Tercentenary Foundation and the European Commission.

5. REFERENCES

[1] F. Lacerda, E. Klintfors, L. Gustavsson, L. Lagerkvist, E. Marklund, and U. Sundberg, "Ecological Theory of Language Acquisition" Proceedings of the 4th International Workshop on Eigenetic Robotics, pp. 147-148. Genoa, 2004.

[2] U. Sundberg, "Mother Tongue: Phonetic aspects of infant-directed speech" PERILUS XXI, Institutionen för lingvistik, Stockholms Universitet, 1998.

[3] F. Lacerda, E. Marklund, L. Lagerkvist, L. Gustavsson, E. Klintfors, and U. Sundberg, "On the linguistic implications of context-bound adult-infant interactions" Proceedings of the 4th International Workshop on Eigenetic Robotics, pp. 147-148. Genoa, 2004.

[4] L. Gustavsson, U. Sundberg, E. Klintfors, E. Marklund, L. Lagerkvist, and F. Lacerda, "Integration of audio-visual information in 8-month-old infants" Proceedings of the 4th International Workshop on Eigenetic Robotics, pp. 149-150. Genoa, 2004.

[5] E. Klintfors and F. Lacerda, "Potential relevance of audio-visual integration in mammals for computational modeling" This conference, 2006.

[6] C. H. Echols, "Attentional predispositions and linguistic sensitivity in the acquisition of object words.," Paper presented at the Biennial Meeting of the Society for Research in Child Development. New Orleans, LA, 1993.

[7] E. S. Spelke and C. J. Owsley, "Intermodal exploration and knowledge in infancy," *Infant Behavior and Development*, vol. 2, pp. 13-24, 1979.