Reference to Objects in Longitudinal Parent–Child Interaction

Kristina Nilsson Björkenstam and Mats Wirén

Department of Linguistics, Computational Linguistics
Stockholm University
kristina.nilsson@ling.su.se, mats.wiren@ling.su.se

1. Background and problem

There are two opposing views on how children master their first language. The first view, usually referred to as linguistic *nativism*, holds that the child is able to acquire language because of innate domain-specific linguistic knowledge. The second view holds that language acquisition can be explained without appeal to this notion, and that the child rather uses stimuli in the environment together with domain-general cognitive abilities to learn language. While these views are seldom rigorously specified and may not even be falsifiable (Clark and Lappin, 2011), they do lead to different research programmes. This paper is part of work that attempts to test how far we can push the second view; in other words, what kinds of domain-general cognitive capabilities might explain language acquisition and how we might model this as an unsupervised learning process.

So far, modelling of language acquisition has mainly been carried out using textual (unimodal) corpora. Although this line of research has shown great progress (Solan et al., 2005), a cognitive model of language learning clearly needs to be dialogue-driven and multimodal to reflect the situation of a (normally developed) child. Parent and child interact a lot, using devices such as words, gaze and object manipulation, and the interpretations of these devices are constrained by the context in which the child is situated. As put by (Clark and Lappin, 2011, page 207): "In order to [reduce] the complexity of the grammar induction problem it is necessary to construct multimodal data bases in which nonlinguistic aspects of interaction ... are encoded in enriched representations of the PLD [primary linguistic data]".

To this end, we are annotating longitudinal video and sound recordings of parent–child dyads. Our basic hypothesis is that the device primarily used to reduce complexity in language learning is *invariance* (or *synchrony*), which means not constancy but rather "relatively stable patterns or structural regularities" (Gogate and Hollich, 2010, page 496). Specifically, we will use the video annotations to try to determine if the amount of synchrony across modalities of parent–child interaction decreases as the child grows older and learns more language and gestures.

2. Corpus

The video recordings have been made from naturalistic parent–child interactions in a recording studio at the Phonetics Laboratory at Stockholm University, using two cameras (Lacerda, 2009). The speech signals from the the child and parent were recorded in separate channels via wireless lavalier microphones. One was attached to a vest that the child wore during the session, and the other was clip-mounted on the shirt of the parent. The child and parent were thus free to move around in the studio. They were also provided with several toys, including two primary target objects, namely, the cuddly toys *Kucka* (a yellow rabbit) and *Siffu* (a black monkey). The scenario was free play, but the parent was instructed to use these toys.

From this corpus, we have so far annotated five longitudinal dyads with two children (a girl and a boy) between the ages of 7 and 31 months. The child is interacting with either his/her mother or father in each dyad. We have grouped the dyads based on the child's age at the time, resulting in four data sets at the ages 7-8 months, 12-13 months, 17-19 months, and 27-31 months. We have annotated a total of 40:59 minutes (15:18, 11:56, 11:46, and 1:59 minutes).

3. Annotation

In this paper, we are primarily interested in exploring invariance across modalities as manifested in parent–child interaction. Although the annotation covers additional features, we here discuss speech, eye gaze, and hand movements by parents and children with respect to the target objects, and the extent to which references to these are synchronized. The data were annotated using ELAN[1].

3.1 Discourse annotation

Segment. A discourse segment is an interval of a dyad in which any or both of the target objects are in focus, by virtue of having been orally referred to by the parent (using for example the name Kucka). A segment starts when one of the target objects is brought into focus by either parent or child by means of speech, eye gaze, or hand movement, and ends when focus is shifted to other objects.

At present, all non-verbal annotation (described below) is restricted to these segments.

3.2 Verbal annotation

Transcription. All utterances of the parents and children were transcribed, using two separate ELAN tiers. Transcription of the parents' speech is orthographical, with additional labels for features like laughter, onomatopoeia and disfluency. In the first two sets of recordings, the children

This research is part of the project "Modelling the emergence of linguistic structures in early childhood", funded by the Swedish Research Council as 2011-675-86010-31.

[1] Freely available at http://tla.mpi.nl/tools/tla-tools/elan/.

do not utter any words that can be orthographically transcribed. Here, the vocalizations by the children are transcribed phonetically. In the last two sets of recordings, the children's vocalizations are transcribed as a combination of orthographic words and non-word vocalizations.

Object mentions. Based on the transcriptions, the timewise extent of each verbal mention of the target object by child or parent was annotated, using a separate ELAN tier for each person (P-Speech and C-Speech, respectively). Such mentions are either names (*Kucka*, *Siffu*), nouns like *kaninen* ('the rabbit'), *apan* ('the monkey'), or pronouns like *den* ('it'), *han* ('he'), *hon* ('she').

3.3 Non-verbal annotation

During annotation of non-verbal information, the sound was turned off. Each annotation task was performed independently of the other, with the results of previous annotations hidden from the annotator.

Eye gaze. On the non-verbal level, we annotated eye gaze by marking whether the child (parent) was looking at the parent (child), Siffu, Kucka, or on any other object. From this annotation, we extracted information on whether child or parent was looking at any of the target objects in synchrony with the speech (P-Gaze and C-Gaze).

Object-related actions and gestures. Object-related actions, i.e., hand movements involving objects, are annotated as `verb_Obj`, where the possible `Obj` values are Siffu, Kucka, Child, Parent, and Other object. In our data, typical object-related actions by the parent is to pick up an object, hold up the object towards the child, and then offer the object to the child. Such a sequence of actions involving Siffu would be annotated as Pick-up_S, Show_S, Offer_S. Many of these actions (for example holding objects up) can be categorized as manipulative forms of deixis (McNeill, 1992, page 327). We include and describe all actions involving objects, for example banging a toy against the floor or dressing a doll. We also annotate object-related gestures, i.e., pointing.

From this annotation, we extracted information on whether child or parent was handling the target objects in synchrony with the speech (P-Hand and C-Hand), for example by reaching for, holding, moving, or offering the object.

4. Discussion

The (long) paper gives two contributions: First, it attempts to provide a format for annotation of video/sound recordings of parent–child dyads for modelling of language learning. Secondly, it uses parts of this annotation to determine if there are changes in the amount of synchrony used by the parents as a function of the age of the child. The results are given in Table 1.

For each mention of either of the target objects by the parent (P-Speech), we determined if the parent was synchronously looking at the object (P-Gaze) and/or moving/touching the object with the hand (P-Hand). By "synchronously" we mean overlap with a time interval beginning 0.5 seconds before and ending simultaneously with the parent's spoken mention of the object. We also determined the corresponding events (and using the same definition of "synchronously", except for C-Speech which is defined as the closest preceding utterance by the child) for the child. Overlaps were given 1 point, and non-overlaps were given 0 points. The figures in Table 1 are means, and thus reflect the proportions of the data points exhibiting synchrony with the spoken modality of the parent. Note that the number of segments (and their durations) vary between the recordings, since the scenario of the child-parent interaction is free play.

Table 1: Proportions of synchrony of parent (P) and child (C) with the spoken modality of the parent (P-Speech) with respect to target objects for different modalities (Gaze, Speech, Hand)

Age of child (mnths)	7-8	12-13	17-19	27-31
P-Gaze	0.42	0.58	0.59	0.40
P-Speech	1	1	1	1
P-Hand	0.77	0.57	0.23	0.30
C-Gaze	0.81	0.71	0.51	0.45
C-Speech	0	0	0.15	0.10
C-Hand	0.35	0.30	0.62	0.50
Number of segments	27	37	19	5
Number of data points	107	106	119	20
Duration (m:s)	15:18	11:56	11:46	01:59

We find the row P-Hand particularly interesting. Here, synchrony decreases as a function of the age of the child, as predicted by our hypothesis. The decrease, overall as well as between each age group, is statistically significant according to a z-test of sample proportions (elaborated in the full paper). P-Gaze is less informative, since it depends on the relative position of the child and parent (for example, sometimes the child was sitting in the lap of the parent, which means the child and parent cannot see each other's faces). By annotating more dyads, we expect both to improve on the annotation and to be able to report even more significant results.

5. References

Alexander Clark and Shalom Lappin. 2011. *Linguistic Nativism and the Poverty of the Stimulus*. Wiley-Blackwell.

Lakshmi J. Gogate and George Hollich. 2010. Invariance detection within an interactive system: A perceptual gateway to language development. *Psychological Review*, 117(2):496–516.

F. Lacerda. 2009. On the emergence of early linguistic functions: A biological and interactional perspective. In K. Alter, M. Horne, M. Lindgren, M. Roll, and J. von Koss Torkildsen, editors, *Brain Talk: Discourse with and in the brain*, pages 207–230. Media-Tryck, Lund.

David McNeill. 1992. *Hand and Mind. What gestures reveal about thought*. The University of Chicago Press, Chicago.

Zach Solan, David Horn, Eytan Ruppin, and Shimon Edelman. 2005. Unsupervised learning of natural languages. *PNAS*, 102(33):11629–11634, August.