Modelling Synchrony of Non-Verbal Cues in Parent–Child Interaction

Mats Wirén[1], Kristina Nilsson Björkenstam[1] and Robert Östling[2]

Department of Linguistics[1]/Department of Modern Languages[2]
Stockholm University/University of Helsinki
SE-106 91 Stockholm, Sweden/00014 Helsinki, Finland
{mats.wiren, kristina.nilsson}@ling.su.se, robert.ostling@helsinki.fi

1. Introduction

A cognitive model of language acquisition ultimately needs to be grounded in social interaction and multimodality to reflect the significance of phenomena like joint attention, non-verbal cues and intention-reading in language learning (Baldwin, 1991; Baldwin, 1993; Tomasello, 2000). A behaviour which involves all of these phenomena is *synchrony*; that is, timewise co-occurrences of signals in separate modalities displayed by the parent, such as sound (speech) and vision (gaze direction, hand movements). In previous studies of a multimodally annotated corpus of parent–child interaction, we have shown that parents who interact with infants at the early word-learning stage (7–9 months) display a large amount of synchrony, but that this behaviour tails off with increasing age of the children (Björkenstam and Wirén, 2014). To shed light on the mechanisms underlying synchrony, we have also investigated the *informativeness* of the non-verbal cues, that is, to what extent they can actually help the infant discriminate between different possible referents (Björkenstam et al., 2016). A crucial aspect of this is the timing of the cues, and whether the informativeness of the cues used by the parents is unaffected by small displacements in time (that is, by less synchrony). This paper reports continued work along these lines.

2. Related work

Several computational models of word learning based on cross-situational information about sounds and perceptually salient objects have been put forward, for example, Yu and Ballard (2007), but most of these models do not take the time-order of events into account. One exception is Frank et al. (2012), who attempted to quantify the informativeness of eye gaze, hand positions and hand pointing (collectively called social cues) directed at objects as coded from video sessions of parent–child interaction. For each spoken utterance by the parent, they coded a) the toys present in the field of view of the child at the time of the utterance; b) the objects in the context that were being looked at, held or pointed to by the parent (the social cues); c) the objects that were being looked at or held by the child (referred to as attentional cues); and d) the parent's intended referent for the referring name, noun phrase or pronoun in the utterance ("look at *the doggie*", "look at *his* eyes and ears").

An analysis of the informativeness of the individual social cues showed that they were noisy, and that no such cue was able to disambiguate fully between objects on its own. (The number of objects in the child's view, hence the ambiguity, for each utterance was on average between 1.2 and 2.9 per dyad.) The cues were used frequently but correct only half or less than half of the time in the sense that they were directed at the object referred by the parent. Simulations with a supervised classifier showed that only a moderate improvement of the accuracy could be achieved by combining information from different cues. However, a possible explanation of this noisiness is the coarse temporal granularity of the model, where a referent was predicted from all the events observed during an entire utterance. Thus, any temporal coordination below the utterance level was invisible. For example, if the parent was looking first at one object and then at another object during the same utterance, the coding did not capture the timing and ordering of these events.

Björkenstam et al. (2016) showed that it is possible to obtain a much more precise picture of the informativeness of non-verbal cues in parent–child interaction by adopting continuous time resolution, which in turn was made possible by their fine-grained, multimodal corpus annotation. As a proxy for informativeness, they used classification accuracy with respect to verbally referred objects, with predictions being based on information about the non-verbal cues, and with different assumptions about the length of short-term memory. A limitation of the study, however, was that only a memory backwards in time was modelled in the sense that only non-verbal cues that occurred *before* the verbal mentions were remembered (that is, generated features used for prediction). But a frequently occurring phenomenon, not captured by this model, is that the parent looks at the child at the time of the mention, and only then looks at the object. This is typically what happens when the parent "follows in" on the child's current focus (Tomasello and Farrar, 1986).

The aim of this paper is to generalise the earlier model to the more realistic scenario in which there is both a forward and a backward memory, and to determine what effects this has on the informativeness of the non-verbal cues.

3. Data

3.1 Corpus

Our primary data consist of audio and video recordings (using two cameras) from parent–child interaction in a recording studio at the Phonetics Laboratory at Stockholm University (Lacerda, 2009). The corpus consists of 18 parent–child dyads, totalling 7:29 hours, with three children each

participating longitudinally in six dyads between the ages of seven and 33 months. The mean duration of a dyad is 24:58 minutes. The scenario was free play where the set of toys varied over time, but where two target objects (cuddly toys) were present in all dyads and thus very frequently referred to.

3.2 Coding

All annotation of the corpus was made with the ELAN tool (Wittenburg et al., 2006) according to the guideline of Björkenstam and Wirén (2014). Annotations in ELAN are created on multiple tiers that are time-aligned to the audio and video, with separate tiers for the parent and child, as well as for events that include different verbal and non-verbal cues. The latter are coded in cells spanning the corresponding timelines on the associated tier, thereby allowing us to track information from the cues very precisely.

First, for each dyad, the discourse segments in which a target object was in focus were coded by creating cells that spanned the corresponding timelines in a designated tier, annotated with the name of the focused object. "Focus" here means that at least one of the participants' attention was directed at a target object, and that, in the course of the segment, at least one verbal reference to the object was made by the parent. (Thus, there is not necessarily joint attention to the target object in the whole of such a segment.) Such a segment was considered to end when the focus was shifted permanently to another (target or non-target) object.

These segments were then coded for verbal and non-verbal referential cues, involving speech, eye gaze and manipulation of an object by hand. The coding used cells spanning the timelines corresponding to the respective events in a separate tier for each type, and with separate tiers for the parent and child, thus resulting in six ELAN tiers overall.

The coding of speech involved all references to objects and persons present in the room by means of a name, definite description or pronoun. Each such reference was coded in an annotation cell spanning the timeline corresponding to the duration of the expression, with addition of its orthographic transcription and the speaker's intended referent. The coding of gaze similarly consisted of a cell spanning the timeline of the act, with a specification of the agent and object looked at (see Table 1). The coding of manual object manipulation additionally distinguished between different types of object manipulation acts (see Table 2).

An additional coding which is relevant here distinguishes (re)introductions of objects where the parent is responding verbally to the child's attention (*follow-in*) or getting the child's attention (*bring-in*) by means of speech and/or object manipulation (Tomasello and Farrar, 1986).

4. Method

Following Frank et al. (2012), we use classification accuracy as a proxy for the variable we are really interested in, namely, the informativeness of different cues. Highly informative cues provide relatively unambiguous information about the referent, and a classifier should then be able to identify the referent with a high level of accuracy. The classifier is only given information about the non-verbal cues

Table 1: Tuples extracted from coding of gaze. `P` = parent, `C` = child, `Siffu` = target object 1, `Kucka` = target object 2

Element	Values
Predicate	`gaze`
Agent	`P`, `C`
Patient	`Siffu`, `Kucka`, `C`, `bag-lid`, `bag`, `P`, ...

Table 2: Tuples extracted from coding of hand manipulation of object. `P` = parent, `C` = child, `Siffu` = target object 1, `Kucka` = target object 2

Element	Values
Predicate	`hold`, `reach`, `move`, `show`, ...
Agent	`P`, `C`
Patient	`Siffu`, `Kucka`, `C`, `bag-lid`, `bag`, `P`, ...

and the time of the parent's referring utterance.We used supervised classification in the form of multinomial logistic regression, equivalently formulated as maximum entropy modelling (Berger et al., 1996). We performed multinomial classification between the possible referents at time t coinciding with the start of a mention by the parent, using predictors that depended on the type of event as well as the time passed since the event finished.

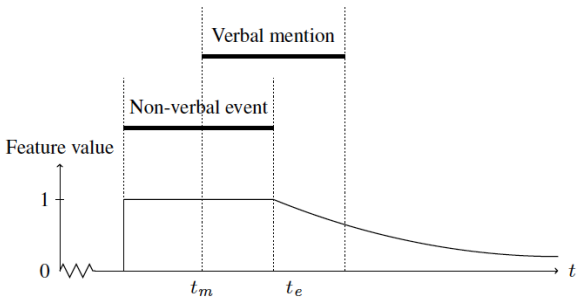

Figure 1: Backward memory as seen from spoken mention, where t_m is the time at which the mention starts and t_e is the time at which the non-verbal event ends. Features for on-going non-verbal events have value 1. After the end of the non-verbal event, the value is determined by the decay function.

As features for the classifier, we extracted information from the coding which we represent as tuples. Thus, for gaze, we extracted triples consisting of $\langle \texttt{gaze}, \textit{agent}, \textit{patient} \rangle$, as shown in Table 1, and for object manipulation triples in the format $\langle \textit{predicate}, \textit{agent}, \textit{patient} \rangle$ as shown in Table 2, for example, $\langle \texttt{pick-up}, \texttt{C}, \texttt{car} \rangle$. Each combination of values in a tuple that encodes a non-verbal event corresponds to a feature in the model. To compute the value of this feature at time t, we used an exponential decay function to simulate short-term memory, as illustrated in Figure 1. The memory equation has the form $f(t) = e^{-kt}$. Here, k is a constant

Table 3: Accuracy (in percent) of model prediction given type of cue. Columns show from which agents information is incorporated into the model (P = parent, C = child, P + C = both). The upper half shows results from our model as described. The lower half uses the same data but only utterance-level binary features, thus emulating the model of Frank et al. (2012).

Type of cue used	P	C	P + C
Fine-grained temporal information			
Hand	72.9	71.8	82.5
Gaze	75.8	80.8	84.2
Hand + gaze	81.7	83.6	88.7
Utterance-level temporal information			
Hand	61.5	64.1	66.6
Gaze	61.4	59.8	62.3
Hand + gaze	64.4	65.0	69.5

that determines the half-life of the memory, and t is defined by $t = t_m - t_e$, where t_m is the time at which the mention *starts*, and t_e is the time at which the non-verbal event *ends*, or $t = 0$ in case these two overlap. Features for on-going non-verbal events are defined to have a value of 1; when a non-verbal event ends, the value of the feature is determined by the decay function. In case the non-verbal event and mention overlap, the event will have a value of 1, according to the memory equation. Future non-verbal events (that have not yet occurred) are here defined to have a value of 0. Put differently, there is no forward memory.

We trained and evaluated models using leave-one-out cross validation on the recording session level, so that we fitted as many models as there are recording sessions (18). Each model was fitted using data from all but one session, then used to predict the referents of the remaining session. This method allowed us to use as much as possible of the available data, while at the same time avoiding session-specific context to influence the model.

5. Informativeness and timing

Björkenstam et al. (2016) used the model described above to obtain measures of the informativeness of the non-verbal cues from the parent and child separately and jointly. Table 3 shows the accuracy of the model's predictions given different cue combinations and agents. The half-life of the short-term memory decay in this experiment was 3 seconds. The baseline was given by the most frequently referred object (target object 1, `Siffu`), which was used in 58% of the cases.

As seen in the table, gaze is more informative than hand manipulation of objects, and, perhaps somewhat surprisingly, the most informative cue is the child's gaze. We can also see that the prediction accuracy is higher when the information sources are combined, as expected. The lower half of Table 3 shows the results of emulating the model of Frank et al. (2012), that is, associating all features with the utterances with which they overlap (with no use of menory decay). The result is a sharp decline in prediction accuracy, only slightly above the baseline. Also, gaze is then less informative than hand manipulation. Both of these results are consistent with the those of Frank et al., and the conclusion drawn by Björkenstam et al. (2016) is that continuous time resolution is needed for a proper analysis of the informativeness of the cues.

In a further experiment, Björkenstam et al. (2016) trained a classifier on input where the timing of the predictions relative to the onset of speech had been moved by whole seconds up/down to ± 4 seconds. This is comparable to displacing the speech relative to the non-verbal event with the same amount of time. They also explored how different memory decays influenced classification accuracy by comparing classifiers with a memory half-life of 1, 3 and 10 seconds, respectively. The effects of the timing displacement on accuracy is shown in Figure 2. The 0 second verbal mention offset is the baseline, with an accuracy of about 86% for the 1 second memory model, and around 88% for the 3 and 10 second memory models. Accuracy dropped when verbal mention offset was displaced. Offsetting the verbal mention ahead in time by as little as two seconds resulted in accuracy scores of 82% for the 1 second model, and 84% for the 3 and 10 second memory models. Delaying the verbal mention by 2 seconds had a less detrimental effect, in particular for the 10 second model. Interestingly, the asymmetry resulting from displaced timing is consistent with experimental results in another paradigm (Trueswell et al., 2016, p. 128), where observers try to estimate referential transparency by reconstructing intended referents from non-verbal cues as they watch a muted video of parent–child interaction.

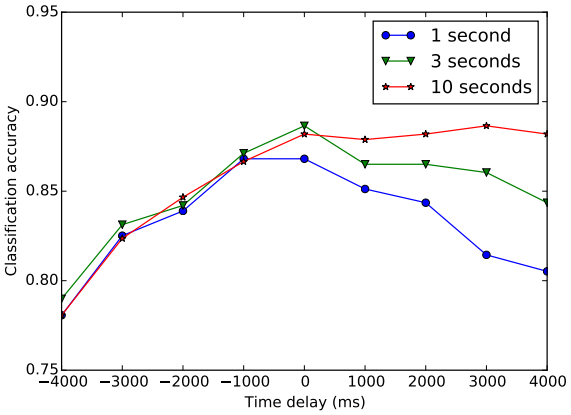

Figure 2: Classification accuracy (y-axis) as a function of verbal mention offset whole seconds from actual word occurrence in parent speech up/down to ± 4 seconds (x-axis), given a short-term memory of 1, 3, and 10 seconds, respectively. Time = 0 coincides with the start of the mentions by the parent. The memory only goes backwards as seen from the spoken mention.

6. Backward and forward memory

The model described above only handles cases where the non-verbal cue occurs (or starts) *before* the verbal mention. However, to cover situations where the non-verbal cue occurs *after* the verbal mention, a memory that looks forward in time as seen from the mention is also needed. To this end,

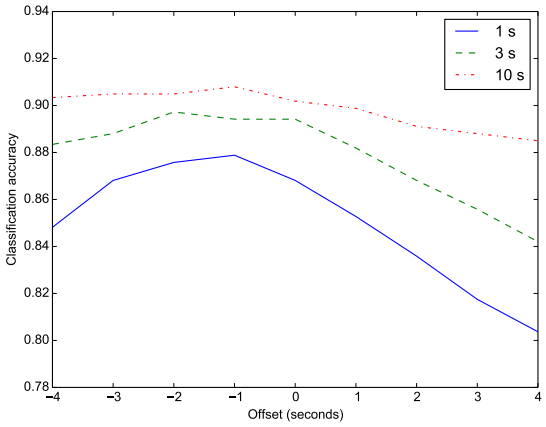

Figure 3: Classification accuracy similar to Figure 2, but with a memory going symmetrically forward and backward as seen from the spoken mention at time = 0.

we added a forward memory to the model, and trained classifiers in which the forward memory was symmetric with the backward memory in terms of half-lifes (1, 3 and 10 seconds).

The results of the experiment are shown in Figure 3. Here, the optimal timing for the parent's spoken mention occurs about a second earlier for the 1 second memory than the actual timing in the data, extending even further for the longer memories. This indicates that a common behaviour in our data is that parents react on something that has already happened, and "follow in" in response to an initiative of the child.

7. Discussion

The aim of this paper was to arrive at a fine-grained model of the informativeness of non-verbal cues in parent–child interaction and the effects of displaced timing of the non-verbal cues. To this end, we generalised our earlier model to include a memory that extends both backward and forward in time. The rationale for this generalisation is to be able to cover situations where any of the agents (parent or child) takes the initiative. What we have seen is that the optimal timing for the parent's spoken mention occurs earlier than in the data, in contrast to the model with only backward memory. We attribute this to the fact that *follow-in* is almost as common as *bring-in* in our data, which also seems to accord with the result that the child's gaze is the most informative cue. A similar result for child gaze was obtained by Johnson et al. (2012). Also, our earlier study (Björkenstam and Wirén, 2014) in general showed a high degree of synchrony for child gaze, but a decreasing tendency as the children get older. To get a further handle on cause and effect in synchrony, however, a possible next step would be to correlate our annotation of *follow-in* and *bring-in* with cases of non-verbal cues occurring before or after the verbal mention.

Acknowledgements

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. We would like to thank (in chronological order) Anna Ericsson, Joel Petersson Ivre, Johan Sjons, Lisa Tengstrand, and Annika Schwittek for annotation work.

References

Dare A. Baldwin. 1991. Infants' contribution to the achievement of joint reference. *Child Development*, 62(5):875–890.

Dare A. Baldwin. 1993. Infants' ability to consult the speaker for clues to word reference. *Journal of Child Language*, 20(2):395–418.

Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71, March.

Kristina N. Björkenstam, Mats Wirén, and Robert Östling. 2016. Modelling the informativeness and timing of non-verbal cues in parent–child interaction. In *Proceedings of the 7th Workshop on Cognitive Aspects of Computational Language Learning*, pages 82–90.

Kristina N. Björkenstam and Mats Wirén. 2014. Multimodal annotation of synchrony in longitudinal parent–child interaction. In J. Edlund, D. Heylen, and P. Paggio, editors, *MMC 2014 Multimodal Corpora: Combining applied and basic research targets: Workshop at The 9th edition of the Language Resources and Evaluation Conference*. ELRA.

M.C. Frank, J.B. Tenenbaum, and A. Fernald. 2012. Social and discourse contributions to the determination of reference in cross-situational learning. *Language Learning and Development*, pages 1–24.

Mark Johnson, Katherine Demuth, and Michael C. Frank. 2012. Exploiting social information in grounded language learning via grammatical reduction. In *Proc. 50th Annual Meeting of the Association for Computational Linguistics*, pages 883–891.

Francisco Lacerda. 2009. On the emergence of early linguistic functions: A biologic and interactional perspective. In *Brain Talk: Discourse with and in the brain*, number 1 in Birgit Rausing Language Program Conference in Linguistics, pages 207–230. Media-Tryck.

Michael Tomasello and Michael Jeffrey Farrar. 1986. Joint attention and early language. *Child Development*, 57(6):1454–1463.

M. Tomasello. 2000. The social-pragmatic theory of word learning. *Pragmatics*, 10(4):401–413.

J.C. Trueswell, Y. Lin, B. Armstrong III, E.A. Cartmill, S. Goldin-Meadow, and L.R. Gleitman. 2016. Perceiving referential intent: Dynamics of reference in natural parent-child interactions. *Cognition*, 148:117–135.

P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes. 2006. ELAN: A Professional Framework for Multimodality Research. In *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation*. ELRA.

C. Yu and D.H. Ballard. 2007. A unified model of early world learning: Integrating statistical and social cues. *Neurocomputing*, 70:2149–2165.