*INTERSPEECH 2017*
August 20–24, 2017, Stockholm, Sweden

# Modelling the Informativeness of Non-Verbal Cues in Parent–Child Interaction

*Mats Wirén, Kristina N. Björkenstam and Robert Östling*

Department of Linguistics, Stockholm University, Sweden
{mats.wiren, kristina.nilsson, robert}@ling.su.se

## Abstract

Non-verbal cues from speakers, such as eye gaze and hand positions, play an important role in word learning [1]. This is consistent with the notion that for meaning to be reconstructed, acoustic patterns need to be linked to time-synchronous patterns from at least one other modality [2]. In previous studies of a multimodally annotated corpus of parent–child interaction, we have shown that parents interacting with infants at the early word-learning stage (7–9 months) display a large amount of time-synchronous patterns, but that this behaviour tails off with increasing age of the children [3]. Furthermore, we have attempted to quantify the *informativeness* of the different non-verbal cues, that is, to what extent they actually help to discriminate between different possible referents, and how critical the timing of the cues is [4]. The purpose of this paper is to generalise our earlier model by quantifying informativeness resulting from non-verbal cues occurring both before and after their associated verbal references.

**Index Terms**: language acquisition, child-directed speech, word learning, cross-situational learning, social cues, non-verbal cues, synchrony

## 1. Introduction

Several computational models of word learning based on cross-situational information from sounds and perceptually salient objects have been put forward, for example, Yu and Ballard [5], but most of these models (too numerous to survey here) do not take the time-order of the associated events into account. An exception to this is Frank et al. [6], 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; b) the objects in the context being looked at, held or pointed to by the parent (the social cues); c) the objects being looked at or held by the child (referred to as attentional cues); and d) the parent's intended referent for the noun phrase in the utterance ("look at *the doggie*", "look at *his* eyes and ears"). The condition for coding an event in this way was that it had some overlap with the time-wise extension of the whole utterance.

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. A possible explanation of the noisiness of this model (suggested by Frank et al. themselves) is its coarse temporal granularity in the sense that a referent was predicted from all the events observed during an utterance, thus losing temporal coordination.

Björkenstam et al. [4] showed that it is possible to arrive at a much more precise model of the informativeness of non-verbal cues in parent–child interaction by using continuous-time resolution. This, in turn, was made possible by their fine-grained, multimodal corpus annotation [3]. As a proxy for informativeness, they used classification accuracy of verbally referred objects, with predictions being based on information from the non-verbal cues. It was assumed in this model that only non-verbal cues that occur *before* the verbal mentions have predictive value. To capture the timewise co-occurrence of the cues, a model of memory decay was used which decreased as a function of the time between the non-verbal cue and the verbal mention. In other words, the purpose of the function was to reflect the hypothesis that the non-verbal cues and the verbal mentions must be timewise synchronised, in that particular order, for them to be perceived as causally linked.

A seemingly quite common behaviour, however, is that a non-verbal cue can also occur *after* the verbal reference with which it is associated. For example, the parent may start to look at a target object before naming it, and may then display an additional non-verbal cue which strengthens the one displayed before the mention. The aim of this paper is to generalise the above model by investigating the effects on accuracy of using information from non-verbal cues both before and after their associated verbal reference.

## 2. Data

### 2.1. 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 [2]. 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.

### 2.2. Coding

All annotation of the corpus was made with the ELAN tool [7] according to the guideline of [3]. 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



http://dx.doi.org/10.21437/Interspeech.2017-1143

on the associated tier, thereby allowing us to track information from the cues very precisely.

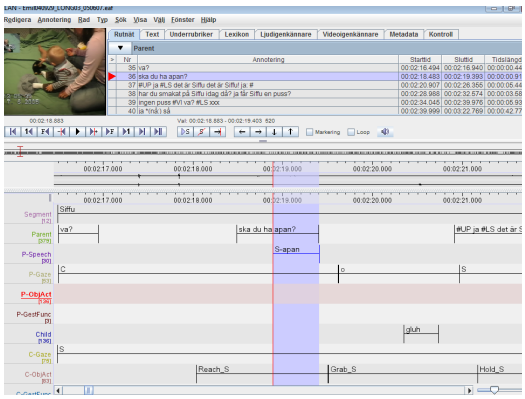

Figure 1: *Screen cap of the annotation in ELAN.*

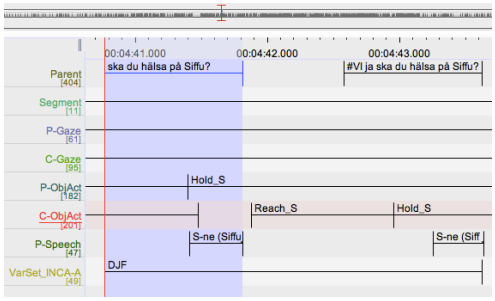


Figure 2: *A close-up of some of the annotation tiers in ELAN: Annotation cells in the* `Parent` *tier contain transcriptions of the parents' utterances. Annotation cells in the* `Segment` *tier contain intervals in which a target object is in focus. Annotation cells in the* `P-Gaze` *and* `C-Gaze` *tiers contain information about the objects looked at by the parent and child, respectively. Annotation cells in the* `P-ObjAct` *and* `C-ObjAct` *tiers contain information about the object-related actions displayed by the hands of the parent and child, respectively, as well as the objects involved. Annotation cells in the* `P-Speech` *tier identify, for each verbal mention, the object referred to by the parent. Note that the extension of each annotation cell codes the time interval of the associated event in continuous time.*

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, comprising altogether 100 minutes or 22% of the corpus, and including 648 mentions to target objects, were then coded for verbal and non-verbal referential cues, involving speech, eye gaze and hand position relative to an object. 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 event, with a specification of the agent and object looked at. The coding of hand additionally distinguished between different types of object manipulation acts.

## 3. Method

Following Frank et al. [6], we use classification accuracy as a proxy for the variable we are really interested in, namely, the informativeness of the 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 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 [8].

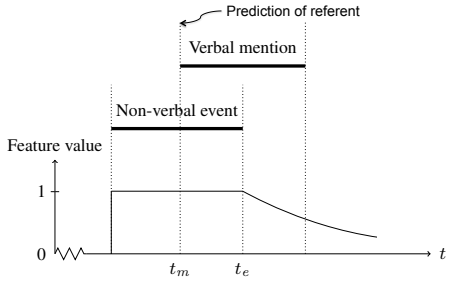


Figure 3: *Short-term memory as seen from spoken mention, where $t_m$ is the time at which the mention starts and $t_e$ is the time at which the non-verbal event ends. Features for on-going non-verbal events have value 1. After the end of the non-verbal event, the value is determined by the decay function.*

As features for the classifier, we extracted information from the coding which we represent as tuples. For gaze, we extracted triples consisting of $\langle \texttt{gaze}, \mathit{agent}, \mathit{patient} \rangle$, and for hand triples in the format $\langle \mathit{predicate}, \mathit{agent}, \mathit{patient} \rangle$. 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 3. The memory equation has

Table 1: *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 [6].*

| Type of cue used | P | C | P + C |
|---|---|---|---|
| **Continuous-time resolution** | | | |
| Hand | 72.9 | 71.8 | 82.5 |
| Gaze | 75.8 | 80.8 | 84.2 |
| Hand + gaze | 81.7 | 83.6 | 88.7 |
| **Utterance-level time resolution** | | | |
| Hand | 61.5 | 64.1 | 66.6 |
| Gaze | 61.4 | 59.8 | 62.3 |
| Hand + gaze | 64.4 | 65.0 | 69.5 |

the form $f(t) = e^{-kt}$. Here, $k$ is a constant 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 defined to have a value of 0; in other words, only non-verbal cues that do not start after a verbal mention can predict the referent of this mention.

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 avoiding session-specific context to influence the model.

## 4. Accuracy and timing

Björkenstam et al. [4] used the model described above to train classifiers on cues for gaze and/or hand for the input from each agent as well as from both of them, using the two target objects as referents. Table 1 shows the classification 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, the figures are clearly above the baseline, indicating that the non-verbal cues provide a lot of information for disambiguation. Furthermore, gaze is more accurate than hand, and the single most informative cue is the child's gaze. A similar result for child gaze was obtained by Johnson et al. [9]. We can also see that the prediction accuracy is higher when the information sources are combined, as expected. The lower half of Table 1 shows the results of emulating the model of Frank et al. [6], that is, associating all features with the utterances with which they overlap (without temporal coordination or memory decay). The result is a sharp decline in prediction accuracy, only slightly above the baseline. Also, gaze is then less accurate than hand. Both of these results are consistent with those of Frank et al., and the conclusion drawn by Björkenstam et al. [4] is that continuous-time resolution is needed to properly quantify the informativeness of the cues.

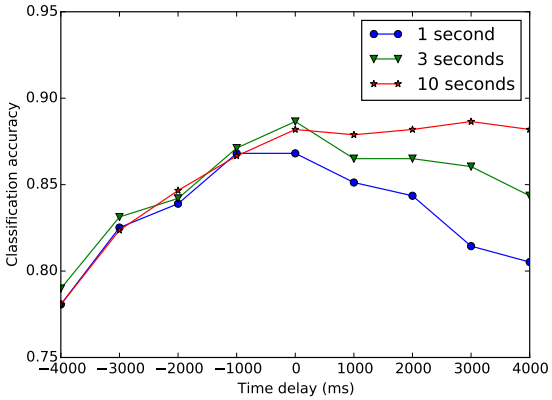


Figure 4: *Classification accuracy (*y*-axis) as a function of verbal mention offset whole seconds from actual word occurrence in parent speech up/down to* $\pm 4$ *seconds (*x*-axis), given a short-term memory of 1, 3, and 10 seconds, respectively. Time = 0 coincides with the start of the mentions by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

In a further experiment, Björkenstam et al. [4] trained a classifier on input from both agents combined, where the timing of the predictions relative to the onset of speech had been moved by whole seconds up/down to $\pm 4$ seconds. This is comparable to displacing the speech relative to the non-verbal event with the same amounts 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 4. 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 the 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 work using an altogether different methodology, the Human Simulation Paradigm [10, p. 128], in which observers try to estimate referential transparency by reconstructing intended referents from non-verbal cues as they watch a muted video of parent–child interaction.

## 5. Symmetric decay

Given that the decay function in the above model was taken to simulate short-term memory, it only made sense to predict the referents of verbal mentions from *previously* occurring non-verbal cues (since only things in the past can be remembered). It is evident from our data, however, that non-verbal cues can also occur after the verbal mentions with which they are associated. For example, the parent may look at a target object while naming it, and may subsequently also touch it, thereby providing an additional non-verbal cue to the object, strengthening the one displayed before the mention.

To capture this kind of behaviour, we were interested in modelling not just the effects of non-verbal cues prior to the

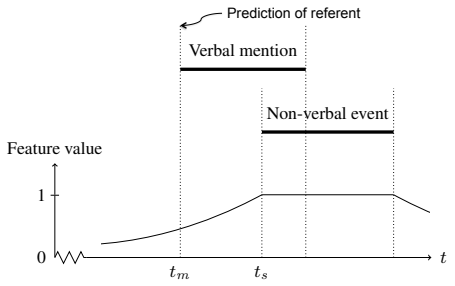


Figure 5: *Symmetric decay from the non-verbal cue, where $t_m$ is the time at which the mention starts and (in addition to what is shown in Figure 3), $t_s$ is the time at which the non-verbal event starts. Features for on-going non-verbal events have value 1. Before the beginning and after the end of the non-verbal event, the value is determined by the decay function.*

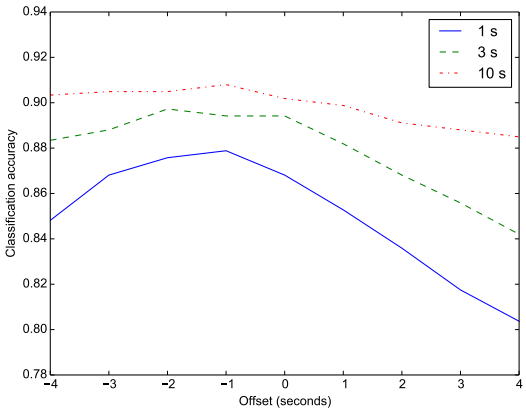


Figure 6: *Classification accuracy analogous with that in Figure 4, but with symmetric decay of the non-verbal cues, forward and backward as seen from the spoken mention at time = 0.*

mentions, but rather the effects of using symmetric decay of the values of the non-verbal cues regardless of whether they occurred before or after the spoken mentions. To this end, we modified the equation from Section 3 to be symmetric in the sense of mirroring the decay forward and backward from the end points of the non-verbal cue, as seen in Figure 5. Using this function, we trained classifiers for the non-verbal input, again from both agents combined, using half-lives for the decay function of 1, 3 and 10 seconds, respectively.

The results of the experiment are shown in Figure 6. Compared to Figure 4, the accuracy is only marginally better. Furthermore, the asymmetry is slightly less pronounced, especially for the longer half-lives of the decay function. The most obvious difference, however, is that the optimal timing of the parents' spoken mentions occurs about a second earlier than the actual timing in the data for the 1 second memory, and even earlier for the longer half-lives. We conjecture that these differences are largely attributable to the following factors.

First, in the beginning of a segment, when a focus shift has just occurred, one phenomenon which may result in additional non-verbal cues from the parent after the verbal mention is the parent exhibiting *follow in* [11]. This occurs when the child takes the initiative by moving its attention to a new target object, and the parent adapts his/her focus to that of the child, typically by first referring verbally to the object and then looking at it, and/or reaching to it, etc. It was shown in Björkenstam et al. [3] that 40% of the segments in our data begin with *follow in*, so the children clearly take a lot of initiatives like this, with the parents adapting their focus.

A second factor which may result in additional non-verbal cues from both the parent and child after the verbal mention, and which is relevant in the rest of a segment, is discourse continuity. This is notion that the parent is talking about the same thing in the current utterance as in the previous utterance. Discourse continuity occurs since, by definition, the focus remains on the same target object (or both of them) throughout a segment (though this does not exclude non-verbal cues from the parent and/or child also to non-target objects in the segment). For example, if the child occupies itself with a target object which has just been referred to by the parent, the child will surely be adding to the non-verbal cues to this object.

To confirm these conjectures, and to understand the precise reasons for the fact that the optimal timing of the parents' utterances occurs earlier than in the data, more analysis is needed, however.

## 6. Conclusion

The aim of this paper was to generalise our model of the informativeness of non-verbal cues in parent–child interaction, and the effects on this of displaced timing of the non-verbal cues, by using information from non-verbal cues both backward and forward in time. The key difference of the results compared to the earlier model, where non-verbal cues occurring after verbal references were not used for predicting the associated referents, was that the optimal timing for the parents' spoken mentions occurred earlier than in the data. We attribute this difference to instances of *follow in* and to the discourse continuity manifested in the segments that we investigate. A more in-depth analysis of the instances in the data and how they contribute to referential transparency is needed to confirm this, however. We believe that our continuous-time annotation and model provide an excellent basis for such an analysis.

## 7. 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.

## 8. References

[1] D. K. MacDonald, Yurovsky and M. C. Frank, "Social cues modulate the representations underlying cross-situational learning," *Cognitive Psychology*, vol. 94, pp. 67–84, 2017.

[2] F. Lacerda, "On the emergence of early linguistic functions: A biologic and interactional perspective," in *Brain Talk: Discourse with and in the brain*, ser. Birgit Rausing Language Program Conference in Linguistics. Media-Tryck, 2009, no. 1, pp. 207–230.

[3] K. N. Björkenstam and M. Wirén, "Multimodal annotation of

synchrony in longitudinal parentchild interaction," in *MMC 2014 Multimodal Corpora: Combining applied and basic research targets: Workshop at The 9th edition of the Language Resources and Evaluation Conference*, J. Edlund, D. Heylen, and P. Paggio, Eds. ELRA, 2014.

[4] K. N. Björkenstam, M. Wirén, and R. Östling, "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*, 2016, pp. 82–90.

[5] C. Yu and D. Ballard, "A unified model of early world learning: Integrating statistical and social cues," *Neurocomputing*, vol. 70, pp. 2149–2165, 2007.

[6] M. Frank, J. Tenenbaum, and A. Fernald, "Social and discourse contributions to the determination of reference in cross-situational learning," *Language Learning and Development*, pp. 1–24, 2012.

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

[8] A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra, "A maximum entropy approach to natural language processing," *Computational Linguistics*, vol. 22, no. 1, pp. 39–71, Mar. 1996. [Online]. Available: http://dl.acm.org/citation.cfm?id=234285.234289

[9] M. Johnson, K. Demuth, and M. C. Frank, "Exploiting social information in grounded language learning via grammatical reduction." in *Proc. 50th Annual Meeting of the Association for Computational Linguistics*, 2012, pp. 883–891.

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

[11] M. Tomasello and M. J. Farrar, "Joint attention and early language," *Child Development*, vol. 57, no. 6, pp. 1454–1463, 1986. [Online]. Available: http://www.jstor.org/stable/1130423