Modelling the informativeness and timing of non-verbal cues in parent–child interaction

Kristina Nilsson Björkenstam[1], **Mats Wirén**[1] **and Robert Östling**[2]
{kristina.nilsson, mats.wiren}@ling.su.se, robert.ostling@helsinki.fi

[1]Department of Linguistics
Stockholm University
SE-106 91 Stockholm, Sweden

[2]Department of Modern Languages
University of Helsinki
PL 24 (Unionsgatan 40)
00014 Helsinki, Finland

Abstract

How do infants learn the meanings of their first words? This study investigates the informativeness and temporal dynamics of non-verbal cues that signal the speaker's referent in a model of early word–referent mapping. To measure the information provided by such cues, a supervised classifier is trained on information extracted from a multimodally annotated corpus of 18 videos of parent–child interaction with three children aged 7 to 33 months. Contradicting previous research, we find that gaze is the single most informative cue, and we show that this finding can be attributed to our fine-grained temporal annotation. We also find that offsetting the timing of the non-verbal cues reduces accuracy, especially if the offset is negative. This is in line with previous research, and suggests that synchrony between verbal and non-verbal cues is important if they are to be perceived as causally related.

1 Background and introduction

There is a growing literature on how infants use non-verbal input such as parents' hand manipulations of salient objects to infer the meanings of their first words. Meaning seems to arise as a probabilistic process where recurrent acoustic patterns gain referential value as they are linked to time-synchronous recurrent patterns in other modalities (Trueswell et al., 2016; Gogate et al., 2006; Matatyaho and Gogate, 2008; Lacerda, 2009). The details of this process, such as the informativeness and temporal dynamics of different cues in word–referent mapping, are still contested, though. In the social-pragmatic approach, joint attention and understanding of speakers' communicative intentions are the central vehicle for investigating the mapping, but the mechanisms typically appear to be deterministic (Tomasello, 2000). In contrast, the associative learning approach emphasises how cross-situational co-occurrences of words and referents increase the salience of objects, including multiple objects in ambiguous learning contexts.

A frequently used methodology for studying word–referent mapping is the Human Simulation Paradigm (HSP), originally devised by Gleitman and colleagues (Gillette et al., 1999; Piccin and Waxman, 2007; Medina et al., 2011). Here, observers try to estimate referential transparency by reconstructing intended referents from non-verbal cues as they watch a muted video of parent–child interaction. Another methodology, which is used in this paper, is to try to model the word–referent mapping directly. Such a model is based on coding of the referential events in a video, typically as perceived by an ideal observer (Geisler, 2011); in other words, someone assumed to optimally handle the perceptual task given by the learning environment as a whole, as recorded by the video. An example of this line of work is Yu and Ballard (2007). They combined social cues (in the form of prosodic affect and joint attention) with statistical learning of cross-situational co-occurrence into a unified model of word learning, showing that this model performed better than a purely statistical approach. Furthermore, Frank et al. (2009) showed that a unified model of cross-situational co-occurrence and interpretation of speakers' referential intention out-

Proceedings of the 7th Workshop on Cognitive Aspects of Computational Language Learning, pages 82–90,
Berlin, Germany, August 11, 2016. ©2016 Association for Computational Linguistics

performed other models of cross-situational word learning, including the model of Yu and Ballard (2007).

In a subsequent study which is the closest parallel to the problems dealt with in this paper, Frank et al. (2012) attempted to quantify the informativeness of eye gaze, hand positions and hand pointing (social cues), as well as referents of previous utterances (discourse continuity), using an ideal observer scenario. For each utterance, the toys present in the field of view of the child at the time of the utterance were coded. (To determine the timing, coders were listening to the audio.) The union of the sets of such objects associated with all the utterances of a video thus formed the set of possible referents. There were between 3 and 21 different objects per dyad, but the number of objects in the child's view (the ambiguity) for each utterance was on average between 1.18 and 2.93 per dyad. Then the object(s) in the context that were being looked at, held or pointed to by the parent (the social cues) were coded. In addition, the object(s) that were being looked at or held by the child (referred to as attentional cues) were coded. Finally, the parent's intended referent for each utterance — those that contained the name of an object or pronoun referring to it — were coded ("look at *the doggie*", "look at *his* eyes and ears").

The result, based on regarding each cue as a predictor for the object reference, was that pointing was a powerful predictor with a precision of 0.78. However, pointing was not frequently used; in other words, it had low recall in the sense that it was seldom used when an object was referred to (and instead other means were used). Eye gaze and hand position, on the other hand, had low prediction accuracies, with F-scores around 0.45. The result was that the social cues appeared to be noisy and that, generally speaking, no such cue on its own would allow an observer to resolve the referential ambiguities. Simulations with a supervised classifier indicated that the prediction accuracy could be somewhat improved by combining information from any two different cues, but that the third did not add anything.

As discussed by Frank et al. (2012), however, it is possible that some discriminatory power was lost because of the coarse temporal granularity of the model, where any temporal coordination below the utterance level was invisible. For example, if the parent was looking first at one object and later at another object during the same utterance, the coding did not capture the timing and ordering of these events. More generally, if there is a systematic timing relation between verbal and non-verbal cues that can support the learner's choice of referent, then we would want to distinguish it. A second limitation of the model was that all kinds of hand movements and gestures were coded as either of two discrete cues, namely, hand position and hand pointing.

This paper attempts to provide answers to two research questions arising out of this line of work: First, is it possible to obtain a more precise measure of the relative informativeness of the different social cues by adopting a more fine-grained model? Secondly, can we see any effects on informativeness in this model if we offset the timing of the non-verbal cues? In other words, is the timing actually used by the parents in some sense optimal with respect to the synchrony of verbal and non-verbal cues, or is the informativeness robust to (small) displacements of the cues forward or backward in time? To measure the information provided by social and attentional cues, we use a supervised classification method, and different assumptions about the length of short-term memory.

2 Data

This section describes our corpus and the annotation used to code the parents' and children's referential behaviour.

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 (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 of them (the target objects) were present in all dyads.

2.2 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), producing annotation cells on tiers time-aligned

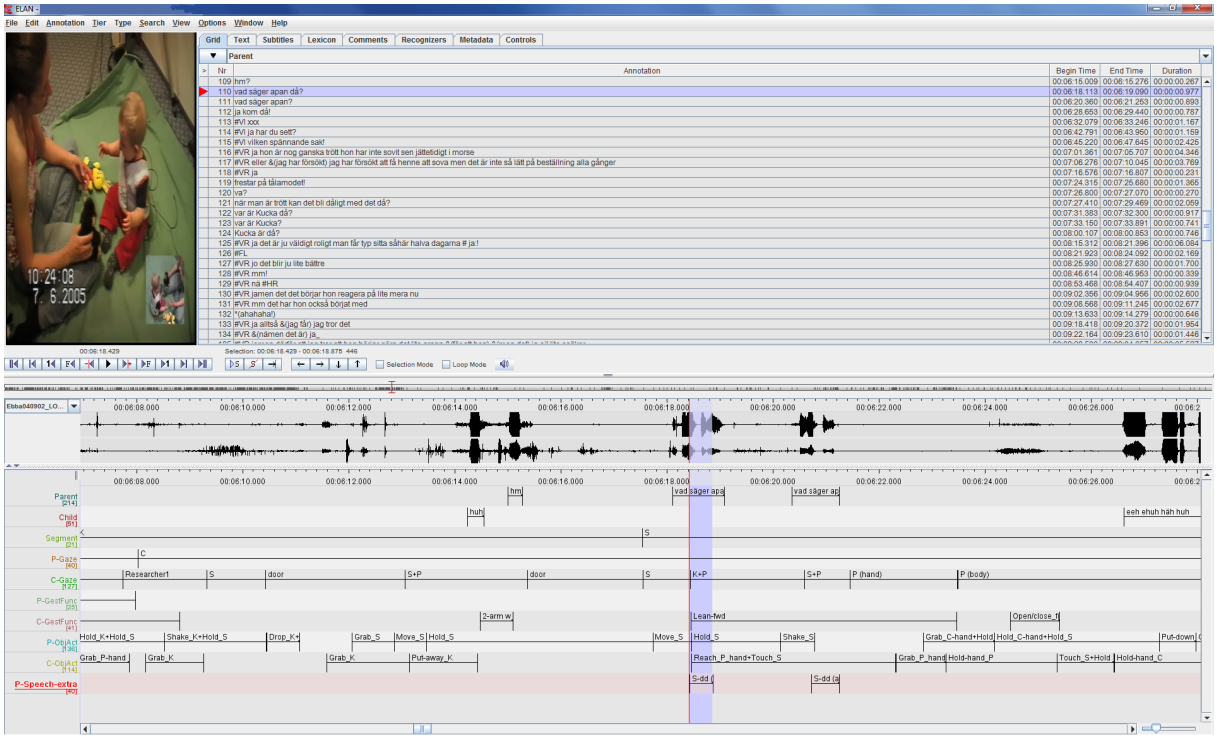

Figure 1: Screencap of ELAN annotation.

with the audio and video files (see Figure 1). The basic approach was to code each type of verbal and non-verbal referential event as well as the parent and child in separate tiers, thereby allowing for analysis separately and in different combinations.

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.[1] "Focus" here means that at least one of the participants' attention was directed to a target object,[2] and that, in the course of the segment, at least one verbal reference to the object was made by the parent. 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, manual gesture, and manipulation of an object by (one or two) hands. 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 eight ELAN tiers overall. We took care in trying to recover information from each cue as objectively as possible. Accordingly, an important methodological consideration was that each tier was coded independently of the others in such a way that all the other tiers were hidden for the annotator.

Table 2: Values of Cohen's Kappa (required overlap 0.6)

Annotation tier	Kappa
Parent's object manipulation	0.71
Child's object manipulation	0.75
Parent's eye gaze	0.60
Child's eye gaze	0.69

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. There were altogether 45 types of objects referred to verbally in the videos, but the distribution of these events was heavily skewed, mostly because of the prominent role of the two

[1] In some segments, both of the target objects were in focus and were then annotated with both names.

[2] Thus, there is not necessarily joint attention to the target object in the whole of such a segment.

Table 1: Number of occurrences of ten most frequent objects referred to verbally, ditto hand manipulations of objects, and ditto objects referred to non-verbally (using gaze or hand), all in decreasing order. P = parent, C = child, Siffu = target object 1, Kucka = target object 2.

Objects referred to verbally	Occur.	Hand manipulation	Occur.	Objects referred to non-verbally	Occur.
Siffu	377	hold	797	Siffu	1229
Kucka	275	reach	539	Kucka	1103
C	166	move	321	C	184
subS	29	show	262	bag-lid	173
subK	24	touch	217	bag	146
P	22	grab	165	P	66
dress-white	14	pick-up	143	dress-white	61
bib	11	explore	120	bottle	55
car	11	enact	114	dress-pink	42
wire	8	shake	95	brush	36

target objects. The most frequently referred objects are shown in the two leftmost columns of Table 1. As seen in the table, only three objects were referred to more than 30 times: target object 1 (called Siffu), target object 2 (Kucka) and child.

As for non-verbal references, the coding of gaze similarly consisted of a cell spanning the timeline of the act, with a specification of the object looked at. If two objects were joined together in the field of view of an agent, the object looked at was coded as the larger of them. For example, if the parent was looking at the child holding a car, we would code this as the child being the subject of the gaze.

In the coding of manual object manipulation, we wanted to capture the large variation in how the parent and child were handling the objects. We thus distinguished 79 types of object manipulation acts, which again turned out to occur in a skewed distribution as shown in the two middle columns in Table 1. Altogether, there were 85 different objects referred to non-verbally (using gaze or hand), of which the most frequent ones are shown in the two rightmost columns in Table 1. Manual gesture occurred very infrequently (and only for the purpose of deictic pointing), and was not used in the subsequent analysis.

The use of timelines in ELAN allows for a high temporal resolution, permitting us to track the information from the cues very precisely. The high resolution also brings technical challenges, however; while Frank et al. (2012) could assume a discrete-time setting and simply use a model predicting referents from all the events observed during an entire utterance, we need a continuous-time model to fully exploit the information from our coding.

Table 3: 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, ...

The reliability of the coding scheme was evaluated by comparing the output by two annotators on two representative dyads, using the built-in ELAN function for calculating Cohen's Kappa (see Table 2). Reliability was high for children's eye gaze as well as object manipulation by parent and child (around 0.7), but slightly lower for parent eye gaze (0.6).

3 Method

While the child has access to a vast amount of information from different senses (including touch, taste, smell, etc.), as well as memories from before the recording session, the goal of our simulated learner is to predict which object is being referred to given nothing but the information from the different cues. We assume, however, that our learner knows how to segment continuous speech

Table 4: 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, ...

into utterances and words, that it can perceive and represent objects in the physical context, and that it is sensitive to the interlocutor's gaze. We furthermore assume that the learner simulates the *beginnings* of lexical acquisition in the sense that the only information provided by the speech is *that* some object in the context is being referred to verbally, but nothing related to the meaning of the words.

To provide a measure of the information inherent in the cues, we use a supervised classification method. Following Frank et al. (2012), we thus 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 reasonable classifier should then be able to identify the referent with a high level of accuracy.

It would also be possible to use the perplexity or, equivalently, likelihood of the test data in order to compare different models. This would capture the (un)certainty of each model, rather than just its ability to predict the correct referent. While intuitively appealing, this would increase the influence of uninteresting model parameters (such as regularization strength) on the result, so for this reason we stick to the more easily interpretable measure of plain classification accuracy.

As features for the classifier, we extracted information from the coding which we represent as tuples. Thus, for gaze, we extract triples consisting of $\langle \texttt{gaze}, \textit{agent}, \textit{patient} \rangle$, as shown in Table 3. For object manipulation we extracted triples in the format $\langle \textit{predicate}, \textit{agent}, \textit{patient} \rangle$, for example, $\langle \texttt{pick-up}, \texttt{C}, \texttt{car} \rangle$. As mentioned in Section 2.2, there were 79 different values for *predicate* and 85 different values for *patient*; the most frequent ones of these are shown in Table 4.[3] We also keep track of the timing information for each mention and each gaze- or hand-related cue.

The particular task that our model solves is a multinomial classification between the possible referents at time t, which we choose to coincide with the start of a mention by the parent. For this, we use a multinomial logistic regression (Maximum Entropy) model with predictors that depend on the type of event as well as the time passed since the event finished.

Each combination of values in a tuple that encodes a non-verbal event, such as $\langle \texttt{gaze}, \texttt{P}, \texttt{car} \rangle$ or $\langle \texttt{pick-up}, \texttt{C}, \texttt{car} \rangle$, corresponds to a feature in the model. To compute the value of this feature at time t, we use an exponential decay function to simulate short-term memory. The memory equation has the form $f(t) = e^{-kt}$, where k is a constant that determines the length (half-life) of the memory, and t is defined by

$$t = t^{start}_{mention} - t^{end}_{event}$$

where $t^{start}_{mention}$ is the time at which the mention starts and t^{end}_{event} is the time at which the non-verbal event ends, or $t = 0$ in case these two overlap. Ongoing non-verbal events are defined to have a value of 1, but as soon as the non-verbal event ends, the decay begins. In case the non-verbal event and mention overlap, the event will have a value of 1, according to the memory equation. Future events (that is, events that have not yet occurred) are defined to have a value of 0.[4]

As mentioned in Section 2.2, the distributions of predicates and objects ware skewed. To avoid having a lot of unusual features in the model, we therefore used one threshold for inclusion of verbal mentions, which we set to 100, and one threshold for the use by the classifier of unique triples representing object manipulations, which we set to 10. The rationale for the lower threshold is that the classifier is robust to some noise, but only if there is a sufficient number of instances for the predicting variable (verbal mentions), hence the higher threshold in that case. Consequently, only the three most frequently mentioned objects were used in the classification.

[3] Sometimes one predicate was associated with several patients, for example, $\langle \texttt{gaze}, \texttt{C}, \langle \texttt{car}, \texttt{Siffu} \rangle \rangle$. In this case, two features were generated with the same timestamps: $\langle \texttt{gaze}, \texttt{C}, \texttt{car} \rangle$ and $\langle \texttt{gaze}, \texttt{C}, \texttt{Siffu} \rangle$.

[4] If we would like to put more emphasis on changes of state, it is possible to include decay during an event as well to down-weigh the information from this event once the novelty wears off.

Table 5: Results of experiment 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 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

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

4 Experiments

This section describes how we used our model in three experiments to try to measure the informativeness and timing of non-verbal cues.

Experiment 1: Informativeness of non-verbal cues

First, we were interested in obtaining measures of the informativeness of the non-verbal cues from both the parent and child as seen from a third-person observer (in effect, looking at their joint interaction), as well as from the agents as seen separately. To this end, we trained classifiers on cues including gaze and hand manipulation for the input from each agent as well as from both of them. For this experiment, we used the two target objects as referents. We did not include the child, because the objective here was to use external information sources as seen from the parent and child, and we did not include any other objects for lack of data. The half-life of the short-term memory decay used here was 3 seconds. The baseline is given by the most frequently referred one, target object 1 (Siffu), which was used in 58% of the cases. An uninformed model could thus achieve an accuracy of 58% by always predicting Siffu.

Table 6: Results of experiment 2. Accuracy (in percent) of model prediction per referent.

	Precision	Recall	F-score
C	31.0	13.3	18.6
Kucka	69.0	74.5	71.7
Siffu	73.6	87.8	80.0

Table 5 shows the accuracy of the model's predictions given different cue combinations and information sources (agents). Overall, the differences in predictive accuracy between the various cue combinations are fairly small, but we can note some things. First, gaze turns out to be more informative than hand manipulation of objects. Secondly, a comparison of the P and C columns shows that roughly the same amount of information is provided by both agents, indicating a high degree of convergence in their interaction.

For comparison, we also include at the end of table 5 the corresponding accuracies obtained using the paradigm of Frank et al. (2012), that is, discarding our fine-grained temporal information and using only utterance-level binary features. The result is a sharp decline in prediction accuracy. It is noteworthy that gaze comes out as less informative than hand manipulation under these circumstances, which is consistent with the results reported by Frank et al. The relative importance of cues thus seems to depend strongly on the resolution of the temporal information available to the model.

Finally, we can see that the prediction accuracy is higher when the information sources are combined, as we would expect. The P + C column shows that the prediction accuracy of a third person view classifier (trained on both parent and child input) is consistently higher than the accuracy of the classifiers trained on input from P and C, respectively.

Experiment 2: Informativeness of non-verbal cues to known referents

In the second experiment, we were interested in determining if there were differences in informativess of non-verbal cues that depended on the

object referred to. This question may bear upon problems related to givenness and accessibility in the domain. In each dyad, the child is a second-person referent, and the target objects are third-person referents. For example, according to Ariel (1999), second-person referents are consistently highly accessible, whereas third-person referents are highly accessible only when they constitute the discourse topic. Our model thus permits us to investigate whether there are differences in the informativity of non-verbal cues with respect to second- and third-person referents. Since the number of references to the child was exceeded only by the target objects, we therefore included this as a third object.

For this experiment, we thus trained classifiers on cues including gaze and hand manipulation for the input from both agents combined. Table 6 shows that predicting the child is much more difficult than the external (target) objects. Using gaze and action information from both participants, we achieve F-scores of 71.6% and 80.0% for the two toys, but only 18.6% for the child.

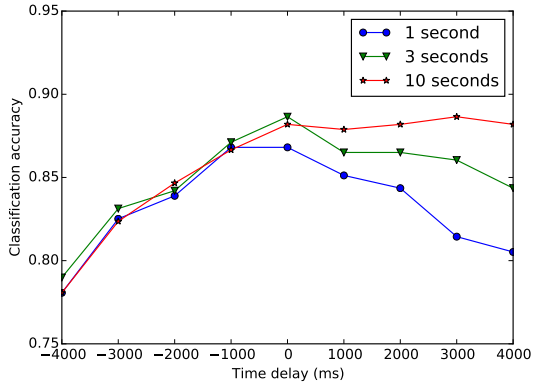

Figure 2: Results of experiment 3. Classification accuracy (y-axis) as a function of verbal mention, offset whole seconds from actual word occurrence in parent speech up/down to ± 4 seconds (x-axis), given a short-term memory of 1, 3, and 10 seconds, respectively. Time = 0 coincides with the start of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

Experiment 3: Timing of non-verbal cues

Our final experiment concerned the timing of non-verbal cues. Previous research has highlighted the time-synchronicity of non-verbal cues with verbal utterances (Matatyaho and Gogate, 2008; Lacerda, 2009). Furthermore, there has been work in the HSP paradigm on determining the effects to referential transparency by displacing these cues (Trueswell et al., 2016). Using our fine-grained representation of time, we wanted to investigate the effects in our model to see if would arrive at similar effects as Trueswell et al.

Our hypothesis was that non-verbal cues are synchronised with speech, and that displacing the verbal mention from its actual temporal position in the input would lead to a drop in classifier performance. We tested this by training 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. We also explored how short-term memory decay influenced classification accuracy by comparing three classifiers with a memory half-life of 1, 3 and 10 seconds, respectively.

The effects of the timing displacement on accuracy appear 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. Moving the verbal mention offset 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.

5 Discussion

The goal of this study was to develop a model for fine-grained measuring of the informativeness and effects of displaced timing of non-verbal cues in parent–child interaction. To this end, we used a corpus of videos of child-directed interaction in a free-play setting involving several objects, but where most of the interaction was centred on two target objects. We coded the segments of the interaction that were focused on these objects with verbal and non-verbal references, using speech, gaze and hand manipulation of objects for this study. To obtain a measure of the informativeness of different cues, we used classification accuracy of the different referents.

The main difference with respect to the model of Frank et al. (2012) concerns the representation of time. Frank et al. use a discrete-time setting in

which a referent is predicted from all the events observed during an entire utterance. In contrast, our model uses a continuous-time representation working off the coding along ELAN timelines. A further difference is that our model includes a simulation of short-term memory decay, where the value of a feature is 1 if it occurs at the time of the mention (the noun phrase), and then decreases exponentially.

Another kind of difference concerns the way in which we represent non-verbal cues. Frank et al. also investigated cues associated with speech, gaze and hand, but for the latter they only used binary features consisting of one discrete cue for hand position and hand pointing, respectively. Our coding is more feature-rich, distinguishing 79 types of hand manipulation.

On the other hand, Frank et al. have a broader perspective in the sense that they also model discourse continuity; in other words, the fact that in the absence of contradicting information, it is most likely that what is being talked about now is the same thing as what was talked about a moment ago. We also do not take prosody into account, as is done by Yu and Ballard (2007).

Our first experiment concerned the relative informativeness of non-verbal cues for word-referent mapping. We found that gaze is the most informative cue, which is inconsistent with the study of Frank et al. In particular, child gaze was highly informative. We interpret this as evidence of the parent's ability to recognise the focus of the child's attention, and to create and maintain joint attention. Additional support for our hypothesis is given by the fact that non-verbal cues, and gaze in particular, became much less informative when we emulated Frank et al.'s experimental setup by discarding temporal information for our classifier.

The third person view classifier, trained on both parent and child input, achieved the highest accuracy. Although we do not have any direct coding of joint attention, it seems that to some degree the third person view classifier captured instances of joint attention through the coding of gaze and object manipulation.

In our second experiment, we compared the informativeness of non-verbal cues to mentions of a second person referent (the child) with mentions of third person referents (the target objects). We found that this task is more complex than classification of mentions of third person referents. These results raise the question whether non-verbal cues are used less when the speaker assumes that the referent of a word is known to the listener. In this case, the parent knows that the child already knows his/her name, and thus references to the child may be used mainly as means of getting the attention of the child.

In our third experiment, we tested the hypothesis that non-verbal cues are synchronous with speech by displacing the verbal mention from its temporal position in the input. We expected a drop in classifier performance, and found that especially negative offsets resulted in lower accuracy. We found an assymmetry in the effect of timing that is similar to experimental results on timing by Trueswell et al. (2016, p. 128), who note that "the greatest changes in cues to referential intent occur just before, rather than after, word onset [...]; moving the beep [that is, word onset] early effectively causes these events to happen too late to be perceived as causally related to the linguistic event".

6 Conclusions

Our findings show that gaze is the single most important non-verbal cue for predicting external object referents, thereby contradicting the study of Frank et al. (2012). We attribute the difference to our addition of fine-grained temporal information, as we can compare our results to those of Frank et al. by simulating their time resolution. Another result is that that non-verbal cues seem much more informative for predicting third-person than second-person references. Finally, we have demonstrated the importance of synchrony by showing that displacing the verbal mention in time degrades prediction accuracy, particularly when the offset is negative. This is consistent with the findings of Trueswell et al. (2016, Figure 2, and compare our Figure 2) who instead of a statistical classifier working off the annotation used human observers of the video.

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, and the three

anonymous reviewers for valuable comments.

References

Mira Ariel. 1999. The development of person agreement markers: From pronouns to higher accessibility markers. In M. Barlow and S. Kemmer, editors, *Usage-based Models of Language*, pages 197–260. Stanford, California: CSLI Publications.

K.N. Björkenstam and M. 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, N.D. Goodman, and J.B. Tenenbaum. 2009. Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, 20(5):578–585.

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.

Wilson S. Geisler. 2011. Contributions of ideal observer theory to vision research. *Vision Research*, 51(7):771–781. Vision Research 50th Anniversary Issue: Part 1.

J. Gillette, H. Gleitman, L. Gleitman, and A. Lederer. 1999. Human simulations of vocabulary learning. *Cognition*, 73:135—176.

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

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.

D.J. Matatyaho and L.J. Gogate. 2008. Type of maternal object motion during synchronous naming predicts preverbal infants' learning of word-object relations. *Infancy*, 13:172–184.

T.N. Medina, J. Snedeker, J.C. Trueswell, and L. Gleitman. 2011. How words can and cannot be learned by observation. *PNAS*, 108(22):9014–9019.

T.B. Piccin and S.R. Waxman. 2007. Why nouns trump verbs in word learning: New evidence from children and adults in the human simulation paradigm. *Language Learning and Development*, 3(4):295–323.

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.