Word Learning in the Wild: What Natural Data Can Tell Us

Barend Beekhuizen
Leiden University Centre for Linguistics
Leiden University
b.f.beekhuizen@hum.leidenuniv.nl

Afsaneh Fazly, Aida Nematzadeh, Suzanne Stevenson
Department of Computer Science
University of Toronto
{afsaneh,aida,suzanne}@cs.toronto.edu

Abstract

When a child first begins to acquire a lexicon, the sources of word-meanings must be available from the situational context. However, it has been argued that the situational availability of the meanings of relational terms, such as verbs, is lower than that of whole-object labels, such as nouns. In this paper, we present a corpus of child-directed language, paired with situational descriptions, that enables us to explore the situational availability of word-meanings using a computational learner.

Keywords: word learning; relational meaning; corpus development; computational modeling

Introduction

However the lexical acquisition process in infants develops beyond the earliest stages, the seeds of the first word meanings must be found in the immediate situational context of early linguistic interaction (Gleitman, 1990). Bootstrapping these early meanings across a variety of situations, so-called cross-situational learning (Akhtar & Montague, 1999), is one of the early cognitive tasks that children need to perform. For cross-situational learning to work, the situational contexts have to contain information that can be extracted and used to determine what the caregiver is likely referring to. However, relatively little is known about the information actually contained in situational contexts.

In this paper, we present a corpus of child-directed language in which the situational context, as found in the accompanying video material, is described in a precise, formalized manner. Not only have the basic-level categories of objects been coded, but also some of their properties and the observable relations among agents and objects. This annotated corpus enables us to explore the situational availability of these various sources of meaning using computational modeling techniques. As such, we demonstrate the use of computational models as a methodological tool to gain insight the information that children have available in their natural learning environment, and that can contribute to cross-situational learning of word meaning.

The process of cross-situational learning has been studied using a multitude of methodologies, each with its limitations. Experimental set-ups must trade off control of the stimuli with the naturalism of the interaction, and thus typically underestimate the complexity of the situations caregiver–child interactions normally take place in (as Medina, Snedeker, Trueswell, and Gleitman (2011) recently noted again). Some computational studies use child-directed language from transcribed child language corpora, which require the researchers to automatically enrich the corpora with artificial meaning representations (Fazly, Alishahi, & Stevenson, 2010).

Interest has grown in the use of multimodal material for computational studies of word-meaning acquisition, since it contains language embedded in a video of the situation of its use. There have been experiments with virtual environments (Fleischman & Roy, 2005), natural environments in which participants were asked to label objects and actions (Yu & Ballard, 2003; Roy & Pentland, 2002), and natural caregiver–child interaction (Roy et al., 2006; Frank, Goodman, & Tenenbaum, 2009). Despite the greater potential for naturalistic data, these corpora also suffer from limitations. First, some only code whole-object labels, thus restricting themselves to the meaning of one sort of words, viz. nouns (Roy et al., 2006; Frank et al., 2009). In others, the language is not child-directed (Fleischman & Roy, 2005; Yu & Ballard, 2003), or the language and situation are unrealistically temporally aligned (Yu & Ballard, 2003). In this paper, we also overcome the above limitations by developing a corpus of child-directed language paired with a precise description of the situational context. Unlike other corpora, the restriction of our corpus to a particular structured activity allows us to precisely describe situational aspects that are relevant to the meanings of various sorts of content words, although the resulting corpus is necessarily small.

One topic we explore in detail is the extent to which words with observable relational meanings (i.e., physical actions and spatial relations) can be bootstrapped from cross-situational learning. As Gentner (1978) argues, mapping words to relations is more difficult than to objects because relations can typically be construed in more ways. Gleitman (1990) shows how even observable relations are often not present at the time of uttering a word referring to them. This paper shows, using a different methodology, that relational terms are indeed harder to glean from the situational context.

A Situated Corpus of Child-Directed Language

Our goal is to construct a corpus that contains situational information that is available to a learner and that can be used in learning the meaning of a variety of content words. For developing such a corpus, there are two requirements. At a minimum, in very early word learning at least, we assume that the information that contributes to a word's meaning must be *situationally available*—that is, the information must be reflected in the situation that is perceivable at or near the time that the word is uttered. But it must also be the case that the learner can process this information and understands its relevance to the interaction with an interlocutor—i.e., the information must be *cognitively available* as well.

In recent work on coding the available whole objects in

video data paired with child-directed language (Roy et al., 2006; Frank et al., 2009), generally only situational availability need be considered, because the cognitive availability of the objects is implicitly assumed. Turning to relational terms, as we do here, we must explicitly argue that the appropriate meanings are cognitively available, because of the evidence that gleaning the appropriate relational meanings from a situation is more difficult (cf., Gentner, 1978). Here we assume that, although child–caregiver interactions take place in the complex world of everyday life, cognitive availability of meanings for the child is eased (again, early on) because of the highly-structured nature of such situations, along with the joint attention caregivers and children share for their objects, relations, goals and consequences, which function to narrow down the set of meanings communicated (cf., Tomasello, 2003). Thus we focus the annotation on those meanings we argue to be cognitively available to the child, which are not all the objects and relations in the situation, but only the subset that pertains to the current activity.

The result is a corpus that provides information on both the situationally and cognitively available objects, properties of objects, and relations between objects. These annotations rely on relatively lean assumptions about the cognitive availability of this information. To the best of our knowledge, this is the first corpus that pairs observed objects, properties *and* relations with spontaneously produced language. As such annotation is costly, the corpus is necessarily small. It can, however, give us insight into the availability of the sources of lexical meaning in the situational context, and the problems a lack of availability may bring about. In that respect this small but naturalistic corpus complements earlier annotated corpora in enabling us to explore what is and is not available at the time some word is uttered.

The source of our material is a collection of 131 videotaped dyadic interactions (recorded for other purposes) between Dutch-speaking mothers and their 16-month-old daughters, containing activities such as playing games and eating. In the videos, each dyad played a game of putting variously-shaped blocks in a bucket with holes of matching shapes in the lid . A set of 32 block games (152 minutes of video) was selected for our annotation. The first author (a native speaker of Dutch) transcribed all speech according to CHAT-guidelines[1], and two assistants coded the video data for the objects, properties and relations in the situations. The transcriptions contained 7842 word tokens (480 types) in 2492 utterances. The language mostly refers to aspects of the game.

The situational coding was done according to guidelines developed by the first author. As the situation consists of just one type of activity (playing the game), the set of objects, properties and relations is relatively limited. The most common objects are the `bucket`, `lid`, `blocks`, `holes` and the two participants, `mother` and `child`. The feature `color={red,green,yellow,blue}` was coded for the blocks and the feature `shape={square,round,triangular,star}` for blocks and holes. The relations and their roles are in Table 1.

[1] Available at http://childes.psy.cmu.edu/manuals/CHAT.pdf

Table 1: Coded relations. Parentheses denote optionality. Ag = Agent, Pa = Patient, In = Instrument, Re = Recipient, So = Source, Go = Goals, Fi = Figure, Gr = Ground

type	name	roles
action	`grab,letgo,hit`	Ag, Pa, (In)
action	`point,show`	Ag, Pa, Re, (In)
action	`move,force`	Ag, Pa, So, Go, (In)
action	`position`	Ag, Pa, Gr, (In)
spatial	`in,on,off,out,at,near`	Fi, Gr
spatial	`match,mismatch`	Fi, Gr

For every three-second interval of video, all coder-observed relations, their associated objects and their properties were coded.[2] The actions (first four rows of Table 1) denote simple manual behavior, which we assume children can recognize (Baillargeon & Wang, 2002). The spatial relations reflect basic categories of containment and support (`in,on`) and their negation (`out,off`), as well as two relations denoting non-containment and non-support contact (`at`) and nearness (`near`). Understanding basic spatial relations precedes the onset of meaning acquisition and can thus be assumed to be in place (Needham & Baillargeon, 1993; Hespos & Baillargeon, 2001), although many specifics may be language-specific (Choi, 2006).[3] The `match` or `mismatch` with a hole was furthermore inferred from these relations. Spatial relations were deemed salient if a change in the relation occurred (e.g., if a `block` was the Figure of an `in`-relation in the current interval, when it was not in the previous interval).

The coding procedure was evaluated for inter- and intra-coder agreement (Carletta, 1996). All relations were coded reliably both within and between coders (Cohen's $\kappa > 0.8$), except `position` (intercoder: $\kappa = 0.51$, intracoder: $\kappa = 0.47$). When the coders disagreed, the first author decided the annotation. A sample of the resulting data is given in Table 2.

The Computational Model

We use the probabilistic alignment-based word learning model of Fazly et al. (2010), which has been shown to perform well using naturalistic data. Using a computational model, we can manipulate input, and doing so, explore the situational and cognitive availability of information, as well as how changes in the input affect learning (Experiment 2).

The model incrementally takes as input a pair of an utterance (a set of words) and a situation (a set of primitive meanings). The learning algorithm has two phases. In the **alignment** phase, the words and meanings in the input are prob-

[2] Using ELAN (Brugman & Russel, 2004).

[3] Ideally, one would encode the range of construals of a situation, including 'tightness-of-fit'. As a first attempt at relational coding of situations, we opted for convenient, yet widely known, universal notions like 'containment' and 'support'.

Table 2: A sample of the dataset. The dash-separated abbreviations denote blocks and holes and their properties, where for blocks the order is **b**-{**re**d,**gr**een,**bl**ue,**ye**llow}-{**ro**und,**st**ar,**sq**uare,**tr**iangular}, and for holes **ho**-{**ro**und,**st**ar,**sq**uare,**tr**iangular}

time	type	coding/transcription
0m0s	**situation**	<nothing happens>
	utterance	een. nou jij een.
	translation	one. now you one. "One. Now you try one."
0m3s	**situation**	position(mother, toy, on(toy, floor)) grab(child, b-ye-tr) move(child, b-ye-tr, on(b-ye-tr, floor), near(b-ye-tr, ho-ro)), mismatch(b-ye-tr, ho-ro)
	utterance	nee daar.
	translation	no there. "No, there."
0m6s	**situation**	point(mother, ho-tr, child) position(child, b-ye-tr, near(b-ye-tr, ho-ro)) mismatch(b-ye-tr, ho-ro)
	utterance	nee lieverd hier past ie niet.
	translation	no sweetie here fits he not. "No sweetie, it won't fit in here."

abilistically mapped to each other; this process is guided by the conditional probabilities of the meanings given the words ("the learned meanings"). Second, in the **update** phase the obtained alignments are used to update the word–meaning associations by adding the alignment score to the association. The word–meaning associations, next, are used to calculate the learned meanings, which are then used in the alignment phase of the next input. These probabilities are based on the association mass a meaning has for a word, relative to all other meanings associated with that word. For a formal explanation, we refer the reader to Fazly et al. (2010).

Experiment 1: Exploring the Corpus

Using the computational model and the corpus, we aim to gain insight into questions such as: what kind of and how much information is derivable from the situational contexts? And is the information equally valuable for different kinds of words (relational words like verbs and prepositions, and non-relational words like adjectives and nouns)?

Running the Model

A set of each utterance's lemmatized word forms is used as the linguistic input. As the model takes a set of primitive meanings as the other part of its input, we considered all content elements from the structured meaning annotation of the interval containing the start of the utterance as the set of situation primitives. An example of an input item is:

Utterance: {*nee lieverd hier passen hij niet*}

Situation: { `point`, `mother`, `hole`, `triangular`, `child`, `position`, `block`, `yellow`, `near`, `round`, `mismatch` }

We set the two smoothing parameters of the model to reflect the size of the lexicon, as in Fazly et al. (2010).

Evaluation

We need to understand how the model learns various types of words that refer to aspects of the situational context. To this end we need some sort of gold standard, as well as some measure of how well the model approximates this standard.

Table 3: A sample of the lexicon of target words

type	examples
action	*duwen* = `force`, *halen* = {`move`, `off`, `out`}
spatial	*in* = `in`, *af* = `off`, *dicht* = {`lid`, `on`, `bucket`}
object	*gat* = `hole`, *emmer* = `bucket`
property	*rood* = `red`, *ster* = `star`

Many words in the utterances have no semantic representation in the coded situations (articles, modals, discourse particles). As we cannot expect the model to learn anything about them, we do not consider them in our evaluation. This leaves us with a small subset of lemmas ($n = 41$) that do refer to possible aspects of the situation. These are verbs of manipulation (e.g., *pakken* 'grab') and placing (e.g., *stoppen* 'put into'), spatial relations (e.g., *op*, 'on'), object labels (e.g., *blok* 'block') and properties (*vierkant* 'square'). As some words have multiple meanings (*stoppen* meaning `put` and `in`), we have to determine which *set* of meanings should be associated with each word. Table 3 gives a sample of words and their relevant gold-standard (**true**) meanings.

We evaluate the learned meanings using two measures. First, we look at the summed meaning probabilities over the set of true meanings (Summed Conditional Probability or *SCP*). This measure tells us what proportion of the probability mass is correctly assigned.

$$SCP = \sum_{f \in \text{true meanings}(w)} p(f|w) \quad (1)$$

Second, we look at how high the true meanings are ranked among all learned meanings, and do so using Average Precision (*AP*), calculated as follows:

$$AP = \sum_{k=1}^{n} P(k)\Delta r(k) \quad (2)$$

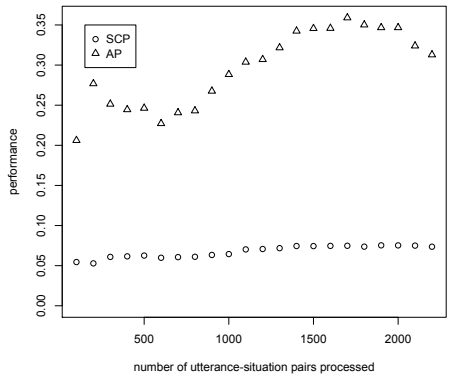

Figure 1: Development of the lexicon's mean *SCP* and *AP*

where k is the rank, n the total number of ranks, $P(k)$ is the number of true meanings found up to and including k, divided by the number of meanings found up to and including k, and $\Delta r(k)$ is the change in recall between $k-1$ and k, which is the number of true meanings found at k divided by the total number of true meanings (which is zero in case no true meanings are found at rank k). This tells us whether the true meanings are more or less prominent than the irrelevant ones.

Results

Table 4 presents the global results, binned per meaning type (properties, objects labels, spatial relations, and actions). We can see that the meanings of non-relational word meanings are ranked higher than those of relational word meanings (compare $AP = 0.81$ and $AP = 0.25$ for properties and object labels, with $AP = 0.19$ and $AP = 0.15$ for spatial relations and action labels), although *SCP* does not differ much between the categories. In general, the probability distributions of the learned meanings do not have very strong peaks: the highest ranking meanings rarely have a learned meaning probability of more than 0.20. Nevertheless, with 78 primitive meanings, the model does learn well beyond a baseline of $\frac{1}{78} = 0.013$.

Looking at the development of the *SCP* and *AP* values over time (Fig. 1), we see strikingly little development in the *SCP*, whereas the *AP* rises for a time, then shows a slight decline.

Splitting the developmental curves out over some of the words (Fig. 2), we see that the words are learned rather heterogeneously. Looking at the *AP* first, some words are acquired instantly, with $AP = 1$ (i.e., the correct meaning ranking first) from early on (*groen* and *rond*), others gradually

Table 4: Results of Experiment 1

	property	object	spatial	action	**total**
SCP	0.10	0.05	0.09	0.07	**0.08**
AP	0.81	0.25	0.19	0.15	**0.31**

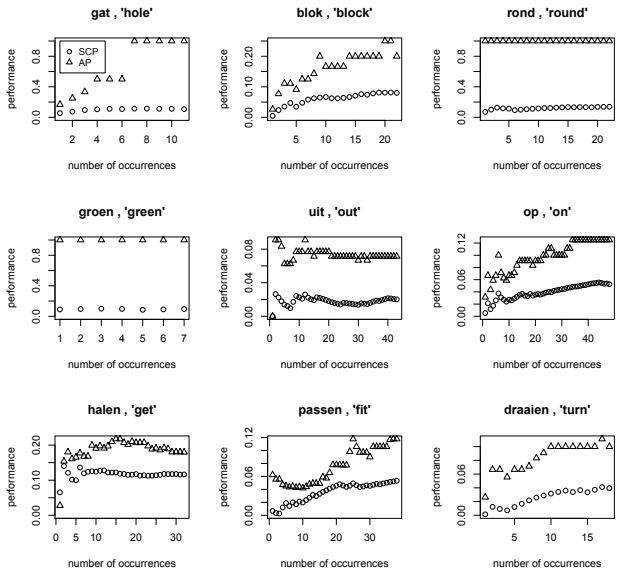

Figure 2: Development of *SCP* & *AP* over time for 9 words

approach $AP = 1$ (*gat*), while for most words, the true meanings remain low ranked. There is, however, a development towards a higher *AP* for many of these words, except for *halen* and *uit*. The *SCP* remains low in all cases, even when the true meaning is ranked first (as in *groen* and *rond*), although note that for several words there is some improvement in *SCP* over time. Recall that the model has only seen 2492 utterances at this point, and that more data may increase the *SCP* further.

Discussion

In this experiment, the model does not learn most words well. One potential reason is the small data set, representing only three hours of interaction. We observe that many developmental curves seem not to have reached their asymptotes yet, suggesting that further learning could occur with more data. We also, admittedly, have the model discard valuable information from the data. Both the linguistic structure (syntax) and the semantic structure (predicate-argument relations) are currently ignored by the model but could be useful in creating the mapping.

In addition, the highly structured and restricted nature of the data, which we expected to help by focusing the learning, may actually be hindering performance. We observe that some words have a very high ranking for their true meanings (high *AP*), yet have low learned probability mass (low *SCP*). (For example, see the words *groen* and *rond* in Figure 2.) On the one hand, the structured and restricted nature of the blocks game entails that a word's true meaning often consistently appears with it. On the other hand, however, the limited nature of the interactions in the data also entails that many irrelevant meanings consistently appear with the word. For example, the object of a `grab` action is almost always a `block`, so that the learner cannot rule out `block` as a possible

meaning of *pakken* 'grab'. The lack of situational variability in the input is thus an obstacle to cross-situational learning, because it requires a consistent co-occurrence of true meanings with a word *coupled with* variability in the presence of irrelevant meanings to help rule them out.

A first solution that comes to mind is a corpus representing a wider variety of activities, with less situational uniformity, for true cross-situational learning. The corpus from which we drew our dyads here does have a number of other types of situations we can include in future annotation. Second, even with relatively homogeneous situations, we expect the learner's attentional mechanisms to help filter out irrelevant meanings. Adding attentional mechanisms, such as the ones in Nematzadeh, Fazly, and Stevenson (2012), is a next step

A final issue we observed with the data is that the true meanings for words in an utterance are sometimes not present within the situational interval paired with the utterance. This problem is very salient for relational meanings, which are often displaced in time from the utterance that refers to them (e.g., *Go grab that one!* or *Don't take the lid off now!*). This might explain why spatial relation terms and verbs display weaker associations with their true meanings than do words for objects and their properties. In the case of positive imperatives, we do find that the actions are often carried out slightly later than the utterance. In Experiment 2, we explore whether this problem of temporal displacement can be mitigated.

Experiment 2: Widening the Temporal Scope

Our hypothesis is that presenting the model with situational meanings only from the time of the utterance impedes the learning of relational terms. Here we explore expanding the temporal scope of the situational input to the model.

Motivation and Set-up

Suppose that in word learning, the learner is not narrowly focussed on the situation at exact moment of the utterance, but also considers some of the situational context taking place around that moment. That is: not only the situation at the very moment of the utterance is cognitively available to a learner, but also some of the surrounding situations. To make this notion precise, we assume that the learner may consider as relevant to an utterance U_i any meanings in the situational context starting from the interval of the previous utterance U_{i-1} up to and including the interval of the next utterance U_{i+1}. (That is, we assume that the relevance of situations overlaps previous and subsequent utterances.) We thus evaluate the model on three possible "windows" W of situational context for utterance U_i: all video intervals up to and including the previous and next utterance in the corpus ($W = U_{i-1} : U_{i+1}$); only the interval of U_{i-1} up to the current interval ($W = U_{i-1} : U_i$), or the current interval up to U_{i+1} ($W = U_i : U_{i+1}$).

Results

Using the same parameter settings and evaluation metrics as in Experiment 1, we obtain the results in Table 5 ($W = U_i : U_i$ is the window-setting used in Experiment 1). The window-setting that only draws situational context from the intervals between the previous utterance and the current one ($W = U_{i-1} : U_i$) does not improve over $W = U_i : U_i$. As hypothesized, however, due to utterances that refer to future actions, the results show that having a window that includes meanings from the intervals up to the next utterance enables the model to learn the object, spatial and action words better (at least according to our *AP* measure). The trade-off is a negligible decline in the learning of property words.

Table 5: Results of Experiment 2

W		prop.	object	spatial	action	**total**
$U_i : U_i$	*SCP*	0.10	0.05	0.09	0.07	**0.08**
	AP	0.81	0.25	0.19	0.15	**0.31**
$U_{i-1} : U_i$	*SCP*	0.10	0.04	0.09	0.07	**0.07**
	AP	0.80	0.17	0.20	0.14	**0.31**
$U_i : U_{i+1}$	*SCP*	0.11	0.06	0.11	0.08	**0.08**
	AP	0.79	0.45	0.24	0.18	**0.40**
U_{i-1}:U_{i+1}	*SCP*	0.08	0.05	0.10	0.08	**0.07**
	AP	0.79	0.41	0.22	0.20	**0.39**

Discussion

Some important information for acquiring the meaning of relational words can be found in the situations unfolding after the utterance has been produced. Clearly, this needs to be interpreted within the context of playing a game, in which the relevant topics of communication (the game goals) often lie in the future w.r.t. the moment of communication. While expanding the situational window adds some irrelevant as well as true meanings, the balance struck by this pragmatically-defined windowing approach seems to help the model acquire the meaning of relational terms (as well as objects!) somewhat better, with little negative impact on property words. Note that the improvement from adding the post-utterance meanings is found mainly in the *AP* metric: the *SCP* values remain similar across the simulations. Even though the probability mass of the true meanings is not changed much, they are now more often better than the irrelevant meanings. This means that the probability values are close to each other and a very small change may improve the rankings visibly.

General Discussion and Future Directions

In this research, we have developed a corpus of caregiver–child interactions in which video is annotated with transcribed utterances and a precise description of the depicted situational context. Unlike other recent multimodal corpora, our annotation of the situational context includes meaning elements that correspond not only to objects and their properties, but to relations as well. Thus the meaning annotations support the learning of various word types, including nouns, adjectives, prepositions/particles, and verbs. Our initial work has explored how we can use this corpus with a computational

model of cross-situational word learning to explore what information must be available to the child from the situation to support word learning, and to examine the relative ease or difficulty of learning various types of words in early acquisition.

Despite the small size of the target lexicon, the model did not perform robustly in the learning task, revealing a number of potential areas of improvement for both the corpus and the model itself. First, due to the cost of annotation, the size of the corpus (only 8,000 word tokens) almost certainly limits the learning. Nonetheless, even this small corpus can be a complementary source of information to larger corpora that are semantically less naturalistic, or contain only object labels. Second, the corpus seems to lack sufficient cross-situational variability for many words to be learned. In more general child–caregiver interactions, a word occurs across a wider variety of contexts (eating scenes, bed-time procedures and so on), enabling a child to rule out as possible meanings those aspects of the context that are irrelevant to the word. Third, regardless of the uniformity or variability of the data, a realistic model of word learning needs to incorporate an attentional mechanism that helps it focus on those aspects of the situation that are likely to be referred to.

Even with this restricted corpus, we find that relational words (verbs, prepositions) are particularly problematic to learn compared to words for objects and properties, in line with a wealth of psycholinguistic observation to this effect (Gleitman, 1990; Gentner, 1978). Because the situational context to which a relational term refers is often displaced, expanding the temporal window of situational context for each utterance led to an improvement in the learning of relational terms, but surprisingly led to even greater improvement in the learning of words for objects.

Perhaps, following Gleitman, more structured learning is necessary for acquiring the meaning of relational words, but the exact source and nature of this structured learning, and its integration with methods of cross-situational learning, is an exciting open issue. Important to look into, and perhaps problematic, is the high proportion of closed-class items in child-directed utterances (e.g., pronouns, aspectual and modal auxiliaries, and particles) that have received little attention in word-learning models, but may play a crucial role in using the structure of an utterance to help determine the meaning of unknown lexical items. More research into the degree to which this information, as found in actual child-directed language, can help is a question in want of an answer, and modeling techniques combined with good data can help us approach it.

Acknowledgments

We gratefully acknowledge the funding of BB through NWO of the Netherlands (grant 322.70.001) and AF, AN and SS through NSERC of Canada, and the Faculty of Arts & Science, University of Toronto. We would like to thank Marinus van IJzendoorn and Marian Bakermans-Kranenburg for making the video data available, and Arie Verhagen and two anonymous reviewers for helpful comments.

References

Akhtar, N., & Montague, L. (1999). Early Lexical Acquisition: The Role of Cross-Situational Learning. *First Language*, *19*(57), 347–358.

Baillargeon, R., & Wang, S.-H. (2002). Event Categorization in Infancy. *Trends in Cognitive Sciences*, *6*(2), 85–93.

Brugman, H., & Russel, A. (2004). Annotating Multimedia/ Multi-modal resources with ELAN. In *Proceedings LREC*.

Carletta, J. (1996). Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, *22*(2), 249–254.

Choi, S. (2006). Preverbal Spatial Cognition and Language-Specific Input: Categories of Containment and Support. In K. Hirsh-Pasek & R. M. Golinkoff (Eds.), *Action Meets Word. How Children Learn Verbs* (pp. 191–207). Oxford, UK: Oxford University Press.

Fazly, A., Alishahi, A., & Stevenson, S. (2010). A Probabilistic Computational Model of Cross-Situational Word Learning. *Cognitive Science*, *34*(6), 1017–1063.

Fleischman, M., & Roy, D. K. (2005). Why Verbs are Harder to Learn than Nouns. Initial Insights from a Computational Model of Intention Recognition in Situated Word Learning. In *Proceedings CogSci.*

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using Speakers Referential Intentions to Model Early Cross-Situational Word Learning. *Psychological Science*, *20*(5), 578–585.

Gentner, D. (1978). On Relational Meaning: The Acquisition of Verb Meaning. *Child Development*, *49*, 988–998.

Gleitman, L. (1990). Sources of Verb Meanings. *Language Acquisition*, *1*(1), 3–55.

Hespos, S. J., & Baillargeon, R. (2001). Reasoning about Containment Events in Very Young Infants. *Cognition*, *78*(3), 207–45.

Medina, T. N., Snedeker, J., Trueswell, J. C., & Gleitman, L. R. (2011). How Words Can and Cannot be Learned by Observation. *PNAS*, *108*(22), 9014–9.

Needham, A., & Baillargeon, R. (1993). Intuitions about Support in 4.5-Month-Old Infants. *Cognition*, *47*, 121–148.

Nematzadeh, A., Fazly, A., & Stevenson, S. (2012). A Computational Model of Memory, Attention, and Word Learning. In *Proceedings CMCL.*

Roy, D. K., Patel, R., Decamp, P., Kubat, R., Fleischman, M., Roy, B., et al. (2006). The Human Speechome Project Stepping into the Shoes of Children. In *Proceedings CogSci.*

Roy, D. K., & Pentland, A. P. (2002). Learning Words from Sights and Sounds: A Computational Model. *Cognitive Science*, *26*, 113–146.

Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Cambridge, MA: Harvard University Press.

Yu, C., & Ballard, D. H. (2003). A Multimodal Learning Interface for Grounding Spoken Language in Sensory Perceptions. *Proceedings ICMI*.