Automatic generation of naturalistic child–adult interaction data

Yevgen Matusevych (Y.Matusevych@uvt.nl)
Department of Culture Studies, Tilburg University
PO Box 90153, 5000 LE Tilburg, the Netherlands

Afra Alishahi (A.Alishahi@uvt.nl)
Department of Communication and Information Sciences, Tilburg University
PO Box 90153, 5000 LE Tilburg, the Netherlands

Paul Vogt (P.A.Vogt@uvt.nl)
Department of Communication and Information Sciences, Tilburg University
PO Box 90153, 5000 LE Tilburg, the Netherlands

Abstract

The input to a cognitively plausible model of language acquisition must have the same information components and statistical properties as the child-directed speech. There are collections of child-directed utterances (e.g., CHILDES), but a realistic representation of their visual and semantic context is not available. We propose three quantitative measures for analyzing the statistical properties of a manually annotated sample of child-adult interaction videos, and compare these against the scene representations automatically generated from the same child-directed utterances, showing that these two datasets are significantly different. To address this problem, we propose an interaction-based framework for generating utterances and scenes based on the co-occurrence frequencies collected from the annotated videos, and show that the resulting interaction-based dataset is comparable to naturalistic data. We use an existing model of cross-situational word learning as a case study for comparing different datasets, and show that only interaction-based data preserve the learning task complexity.

Keywords: Child language acquisition; computational modeling; child-directed speech; cross-situational word learning.

Introduction

A usage-based approach to language claims that natural languages in all their complexity can be learned merely from the input (or usage) data that is available to human learners. Computational modeling has been extensively used as a methodology for supporting this view: using a dataset that is statistically similar to child-directed input, a computational model can show that certain linguistic representations are learnable without domain-specific prior knowledge. Therefore, the input to a cognitively plausible model of language acquisition must have the same information components and statistical properties as the natural child-directed speech (CDS). A careful analysis and reconstruction of such data is a prerequisite of developing a model.

Recent decades have seen a significant growth in the variety and quantity of data collections for studying language. One major resource in this domain is CHILDES (MacWhinney, 2000), a collection of corpora containing recorded interactions of adults with children of different age and language groups. The interaction transcriptions have been used in several models of grammar induction from a large text corpus (e.g., Clark, 2001). The problem arises when a learning task demands perceptual and linguistic input. This might be due to the nature of the process under study (e.g., learning the meaning of words) or the theoretical framework on which the model is based (e.g., construction grammar). In such cases, each utterance must be paired with a representation of its visual context. Many of the databases in CHILDES contain video recordings of the interaction sessions, but these recordings are mostly not annotated and hard to use without preprocessing or manual coding. Some models in fact use a small set of manually annotated videos as input (e.g., Yu & Ballard, 2007; Frank, Tenenbaum, & Fernald, 2013), but this approach is limited in quantity and scalability.

A common strategy for dealing with this challenge is to use artificially generated input: each sentence is constructed by randomly sampling from a presumed distribution over a list of words; the visual context is similarly built by sampling from a set of symbols which represent concepts or objects (e.g., Siskind, 1996; Niyogi, 2002). To make the data more naturalistic, some models select sentences from the transcriptions of actual child–adult interactions, and build the accompanying scene artificially by assuming a semantic representation for each word in the sentence and combining them (Fazly, Alishahi, & Stevenson, 2010).

Generating the visual context automatically based on child-directed utterances (Utterance-Based Data, or UBD) eliminates the quantity concern (since manual annotation of the surrounding scene is not needed). However, the generated context is different from what the child observes in important ways. In a natural interaction scenario between a child and an adult, the surrounding scene is rather consistent or changes minimally (although the attention of the participants might move from one set of perceivable objects or actions to another). In contrast, in the automatic scene generation approach the utterance determines the scene, so the visual context can change drastically from one sentence to the next. A disproportional variation in visual context or scene can affect language learning; for example, context diversity has been shown to facilitate cross-situational word learning (Kachergis, Yu, & Shiffrin, 2009). Moreover, a UBD approach guarantees that the relevant meanings for all the words in an utterance are included in the constructed scene. Artificial noise can be added by post-processing data and ran-

domly removing some meaning elements from the scene, but the noise ratio can still be unrealistically low.

UBD also differs from actual exchanges between children and their caregivers in that it lacks any interaction-based features. Crucially, utterances and actions directed at the learner at each point in time are independent of the learner's reaction to previous input data. In reality, the content of adult's utterance often depends on what the child just did or said (Kishimoto, Shizawa, Yasuda, Hinobayashi, & Minami, 2007; Chouinard & Clark, 2003). Interaction is suggested to be an essential mechanism of language development (e.g., MacWhinney, 2010).

In this paper, we investigate the characteristics of the visual context in a sample of child–adult interaction sessions, and compare them to those in an automatically generated one. We show that in every measure, the two contexts are considerably different, and argue that these differences might have implications for modeling child language learning. We propose a hybrid approach for generating an input corpus of utterance–scene pairs, where co-occurrence frequencies collected from a sample of manually annotated videos are used for generating utterances and visual contexts. Our framework not only takes the usage frequencies of the words and objects into account, but also includes interaction-based features such as dependence of adult's utterance on child's recent behavior. Finally, as a case study, we use an existing model of word learning (Fazly et al., 2010) to compare the complexity of the learning task using UBD vs. Interaction-Based Data (IBD, generated by our proposed framework). Our results show that using UBD for word learning unrealistically simplifies the learning task. Using IBD, in contrast, yields results that are closely comparable to the ones based on manually annotated scenes from videos of child–adult interaction.

Analyzing Utterance-based Input

We analyze the cognitive plausibility of UBD by comparing its characteristics to a carefully annotated set of video recordings of child–adult interactions. The details of this data set are described below.

Data set

As part of a larger project to study cross-cultural aspects of child-language acquisition (CASA MILA; Vogt and Mastin, 2013a, 2013b), three 13-month-old children from the Netherlands were recorded on video. The videos were recorded at the children's homes and involved interactions with one of the parents. The parents were instructed to continue their daily routines and ignore the recordings.

For each child, we selected an interaction session in a toy playing setting. The video fragments were 8 min 37 sec, 8 min 50 sec and 10 min long. We excluded some short episodes from the analysis, namely those (1) where the child or the adult was not captured properly by the camera, and (2) where the other parent was present, and the child's attention was focused on him/her. From the videos we extracted adults' and children's gaze directions, actions, objects or participants that the actions were directed at, and utterances. Using this data, we constructed a corpus of child-directed utterances, each paired with a representation of the accompanying scene.

Scene representation There is no easy way to determine which elements a child perceives as potential referents at a certain moment of time. In fact, any object, action or event from the natural environment can be occasionally referred to in speech. However, studies suggest that children use certain mechanisms and constraints such as referential and salience cues to focus on relevant aspects of the scene (e.g., Behrend, 1990; Moore, Angelopoulos, & Bennett, 1999). In particular, Yu, Smith, Shen, Pereira, and Smith (2009) show that objects in child's and parent's hands dominate the child's visual field.

In coding the interaction context in the video recordings, we consider two different interpretations for a scene:

active: all the objects that either participant (or both of them) is acting on or looking at during an utterance, in addition to the actions that (s)he performs (a similar approach was used by Frank, Goodman, and Tenenbaum (2008)).

all: the full set of visible objects, the action(s) performed during an utterance and the participants.

In addition, a third dataset was automatically generated:

UBD: Fazly et al. (2010) construct a scene by putting together the semantic symbols that correspond to the words in the accompanying utterance. Referential uncertainty is simulated by merging the representations of two consecutive scenes, and pairing them with only one of the utterances. They include noise into the data by removing the semantic symbol of one word from the scene for 20% of the input items. Since we wanted to compare our results to those of Fazly et al. (2010), we applied the exact same approach to the child-directed utterances that we extracted from the CASA MILA recordings.

Measures

To compare the datasets described above, we use three measures: scene stability, noise, and referential certainty.

Scene stability As mentioned before, the stability of the visual scene is one of the main points of deviation between natural interaction settings and the artificially generated input. We measure scene stability as the overlap between every pair of consecutive scenes. Since in both cases (the produced scenes in UBD and the annotated ones in our data set) a scene is represented as a set of symbols, we define the overlap between each two sets as the cardinality of their intersection divided by the cardinality of their union:

$$\text{overlap}(S_i, S_{i+1}) = \frac{|S_i \cap S_{i+1}|}{|S_i \cup S_{i+1}|}$$

Noise We count a word's usage (or token) in an utterance as noisy if its semantic symbol is not included in the scene representation for that utterance. The total number of noisy words in an utterance, then, is calculated as

$$\text{noise}(U_i) = \frac{|U_i| - |U_i \cap S_i|}{|U_i|}$$

where S_i is the current scene, and U_i is the (correct) meaning representation of the current utterance. To avoid making arbitrary decisions about the meaning of abstract or function words, we limit our analysis of noise to objects and actions.

Referential certainty We define the referential certainty for a scene as

$$\text{certainty}(S_i) = \frac{|U_i \cap S_i|}{|S_i|}$$

Conceptually referential certainty shows what portion of a scene is referred to in the respective utterance. Note that this measure is the opposite of the more commonly used *referential uncertainty*, but it avoids the problem of having zero denominators in case the meaning representation of the utterance does not overlap with the scene.

Results

We calculated the above measures for three datasets: child-directed utterances extracted from CASA MILA and paired with two interpretations of the accompanying visual scene (i.e. **active** and **all**), or with **UBD**-style automatically generated scene representations. Results are shown in Table 1.

Table 1: Plausibility measures for three datasets

	all	**active**	**UBD**
Scene stability	0.916	0.436	0.112
Noise	0.414	0.426	0.099
Referential certainty	0.019	0.112	0.602

The average values provided in the table inform us that the *all* condition differs substantially from the other two in terms of scene stability ($\mu = 0.916$ vs. 0.436 and 0.112) and referential certainty ($\mu = 0.019$ vs. 0.112 and 0.602). For this reason, and taking into account the fact that the standard deviation values for the *all* condition are rather small as compared to the respective means ($\sigma_{stability} = 0.065; \sigma_{certainty} = 0.032$), we eliminate this condition from the analysis.[1] To compare the other two conditions, we ran the Mann–Whitney *U*-test for each of the three measures. We found significant differences between the annotated data (*active* condition) and the *UBD* in terms of all three measures: scene stability ($Mdn = 0.400$ vs. 0.059; $U = 5230, n_{active} = 274, n_{UBD} = 133, p < .001, r = -.583$), noise ($Mdn = 0.400$ vs. 0.000; $U = 8927, n_{active} = 278, n_{UBD} = 139, p < .001, r = -.466$) and referential certainty ($Mdn = 0.000$ vs. 0.571; $U = 3910, n_{active} = 278, n_{UBD} = 139, p < .001, r = -.690$). This demonstrates that UBD may be an easier input for the learner than the natural data.

[1] The noise values for the *active* and the *all* conditions are almost equal, since the way we interpret a scene has little impact on the amount of noise in utterances. Due to this fact, for noise we also use only the *active* condition in the further analysis.

An interaction-based framework for input generation

We propose an interaction-based framework for generating input data which resembles the verbal and non-verbal exchanges between a child and a caregiver. Our model is inspired by the language game model used to study the evolution of language (Steels, 1996; Vogt & Haasdijk, 2010). In this model, agents communicate with each other through verbal and non-verbal behavior. Language game interactions involve a context, and agents communicate about items in this context, potentially learning associations between words and items.

We simulate the input generation process as a series of interactive sessions between two agents, Adult and Child. Each session starts with constructing a visual context (i.e., a collection of objects), followed by a sequence of exchanges between the two agents, until one of them leaves or terminates the session. In each turn, Adult performs an action (*AdAct*) while producing an utterance (*AdUttr*), to which Child responds by performing another action (*ChAct*) and producing an utterance (*ChUttr*, implemented as presence or absence of a verbal reaction).[2] The main algorithm can be described as follows:

for $s \leftarrow 1$ **to** *number of interaction sessions* **do**
  $t \leftarrow 0$;
  $Context \leftarrow \text{setupContext}(s)$;
  repeat
    $t \leftarrow t + 1$;
    $Situation_t \leftarrow \text{initialize}(Context)$;
    $Situation_t \leftarrow \text{updateAdult}(AdAct_{t-1}, AdUttr_{t-1})$;
    $Situation_t \leftarrow \text{updateChild}(ChAct_{t-1}, ChUttr_{t-1})$;
    $(AdAct_t, AdUttr_t) \leftarrow \text{adultTurn}(Situation_t)$;
    $Situation_t \leftarrow \text{updateAdult}(AdAct_t, AdUttr_t)$;
    $(ChAct_t, ChUttr_t) \leftarrow \text{childTurn}(Situation_t)$;
  until $ChAct_t =$ *'leave'* **or** $AdAct_t =$ *'leave'*;
end

Each of the main steps in the algorithm are explained in more detail below.

Visual context From the sample data we extracted all the objects that were directly used by adults or children in their interactions. In each computational simulation, we randomly selected a fixed number of objects from the list and added them to the context. Since the size of the visual context may depend on the interaction domain (e.g., toy playing, book reading, etc.), we added it as a parameter to our framework.

Actions and action types We compiled two lists of actions, one for each agent. Actions might take arguments that can be an object type or the agents themselves (e.g., *take toy* or *touch child*). In order to base our computational model on more general behavioral patterns rather than on occasional events, we classified agents' actions into six *types*, based on

[2] Since children in our sample video recordings were too young to talk, we did not gather enough statistical information about their produced utterances. However, the main concern of our framework is to create realistic child-directed input, and the child-produced data is an outcome of the learning model.

the factors that motivate them. These action types are listed in Table 2.

Table 2: Action types and their motivating factors

Action type	Motivating factor	Example
Continuation	*Same* person's previous *action*	$Adult_t$: [move bag] $Adult_{t+1}$: [move box]
Reaction	*Other* person's previous *action*	$Child_t$: [put ball] $Adult_{t+1}$: [take ball]
Result	*Same* person's prev. *utterance*	$Adult_t$: *Bumba first* $Adult_{t+1}$: [take Bumba]
Reaction to utterance	*Other* person's prev. *utterance*	$Adult_t$: *The tree* $Child_{t+1}$: [take toy tree]
Initiating	None	$Adult_t$: [sit down]

Utterances and utterance types We compiled a list of utterances produced by adults. Some of these contain placeholders which, depending on the context, can be filled with the labels for the respective actions and their arguments. Similar to actions, we recognized six utterance types based on their motivating factor, as listed in Table 3.

Table 3: Utterance types and their motivating factors

Utterance type	Motivating factor	Example
Accompanying	*Same* person's *current action*	$Adult_t$: [show ball] $Adult_t$: *This is a ball*
Continuation	*Same* person's *prev. utterance*	$Adult_t$: *Dad the ball?* $Adult_{t+1}$: *Can dad the ball?*
Reaction	*Other* person's *previous action*	$Child_t$: [stand up] $Adult_{t+1}$: *Gonna walk*?
Answer	*Other* person's *prev. utterance*	$Child_t$: *babbling* $Adult_{t+1}$: *Yeah, Bumba*
Unknown	None	$Child_t$: *babbling*

Producing actions and utterances At each step t during a session, the actions and utterances produced by the agents are sampled from the frequency distributions collected from the annotated videos, each conditioned on the current situation. A situation includes all the relevant parameters, including the current and previous utterances and actions of both agents, the action arguments, and the visual context. Thanks to these parameters, the agents do not produce completely random actions and utterances, and the interaction process appears to be logical. (For more details on the estimated probabilities for each variable, see Matusevych (2012).)

Each turn in a session consists of the following steps:

1. The current situation ($Situation_t$) is set to include the visual context (*Context*), the previous actions ($AdAct_{t-1}$ and $ChAct_{t-1}$) and utterances ($AdUttr_{t-1}$ and $ChUttr_{t-1}$)
2. Adult's next action is generated:
   (a) An action type for Adult ($AdActType_t$) is randomly selected, conditioned on $Situation_t$
   (b) An action for Adult ($AdAct_t$) is randomly selected, conditioned on $AdActType_t$ and $Situation_t$
   (c) Arguments for the action are randomly selected, conditioned on $AdAct_t$
   (d) $Situation_t$ is updated to include $AdAct_t$ and its arguments
3. Adult's next utterance is generated:
   (a) An utterance type for Adult ($AdUttrType_t$) is randomly selected, conditioned on $Situation_t$
   (b) An utterance for Adult ($AdUttr_t$) is randomly selected, conditioned on $AdUttrType_t$ and $Situation_t$
   (c) $Situation_t$ is updated to include $AdUttr_t$
4. Child's next action and utterance ($ChAct_t$ and $ChUttr_t$) are generated in the same way as Adult's.

A sample interaction session

We illustrate the interaction process using the following example (see Table 4).

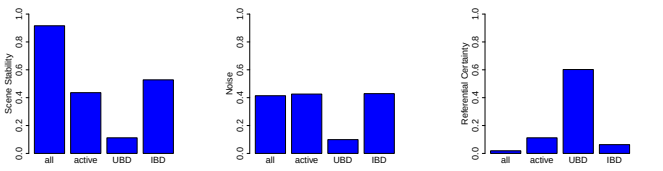
Table 4: A fragment of a generated interaction

Context: puzzle, piece-clown, bin, ball, piece-frog			
Turn	Agent	Action	Utterance
1.	Adult	play puzzle	—
1.	Child	play piece-clown	*babbling*
2.	Adult	point puzzle	*It fits here.*
2.	Child	touch bin	*babbling*
3.	Adult	play puzzle	*Yes?*

The example can be interpreted as following. Adult starts the interaction by playing with a puzzle toy without saying anything. Child plays with a clown-shaped puzzle piece and babbles. Adult points at the puzzle saying *It fits here*. However, Child's attention is distracted by the bin, which he touches. He continues babbling. Adult continues playing with the puzzle toy, asking *Yes?*. His question can be interpreted either as a support for his previous utterance or as an attempt to clarify the child's utterance. The interaction goes on in this manner until one of the agents leaves.

Comparing IBD and UBD

We used the interaction-based framework for generating a dataset. While in UBD scenes were constructed from utterances, in IBD each scene included salient elements, namely, the objects that the agents had in their hands, the agents' most recent actions and their arguments—in a manner similar to the *active* condition in the *Analyzing Utterance-based Input* section above. Using the same three measures—scene stability, noise, and referential certainty—we compare IBD to UBD and manually annotated CASA MILA data (both conditions).



Scene stability Noise Referential certainty

Figure 1: Plausibility measures for four datasets

As can be seen in the charts (Figure 1), for each of the three measures the input data generated by our framework is much closer to the manually annotated data from the interaction videos than UBD. Again, for the reasons specified in the *Analysis* section, we did not use the *all* condition in the further analysis. For the other three conditions, the Kruskal–Wallis H-test showed the significant difference in terms of stability ($H(2) = 213.822, p < .001$), noise ($H(2) = 95.725, p < .001$) and certainty ($H(2) = 289.410, p < .001$). To examine the pairwise differences between the three groups, we used Mann–Whitney tests, taking into account Bonferroni correction (which resulted in .025 level of significance). The difference between *active* and IBD was not significant in terms of noise ($Mdn = 0.400$ vs. 0.333; $U = 37897.5, n_{active} = 278, n_{IBD} = 278, p > .025$), and significant with only small effect size in terms of scene stability ($Mdn = 0.400$ vs. 0.500; $U = 30012, n_{active} = 274, n_{IBD} = 274, p < .001, r = -.174$) and certainty ($Mdn = 0.000$ vs. 0.000; $U = 35002.5, n_{active} = 278, n_{IBD} = 278, p < .025, r = -.100$). However, the difference between UBD and IBD was significant with a large effect size for each measure: scene stability ($Mdn = 0.059$ vs. 0.500; $U = 2586, n_{UBD} = 133, n_{IBD} = 274, p < .001, r = -.699$), noise ($Mdn = 0.000$ vs. 0.333; $U = 10008.5, n_{UBD} = 139, n_{IBD} = 278, p < .001, r = -.426$) and certainty ($Mdn = 0.571$ vs. 0.000; $U = 2451.5, n_{UBD} = 139, n_{IBD} = 278, p < .001, r = -.760$). These results confirm that data generated by the proposed framework is more suitable for training and evaluating cognitive models than UBD. We further investigate this claim by using these different data sets in an existing model of word learning.

Case study: learning word meaning

We used the cross-situational word learning model of Fazly et al. (2010) as a case study for our proposed input generation framework. Our goal is to show that the complexity of the learning task depends on the properties of the input data, and less realistically generated input can considerably simplify the task.

Description of the model

The model of Fazly et al. (2010) incrementally learns the meaning of each word (e.g., *play*) as a probability distribution over all the possible meaning components, each represented as a unique symbol (e.g., PLAY, BALL). At each moment in time, the model receives a new input item, consisting of an utterance and its (ambiguous) semantic representation, which is an unordered set of symbols. The model uses its previous knowledge of word meanings to align each word in the current utterance with the most likely symbols in the current scene representation. It then uses these alignments to update the meaning of each word by accumulating such cross-situational evidence over time.

Model performance on different types of input

We compared the performance of the word learning model on four different data sets:

1. the manually annotated portion of CASA MILA (*active*);
2. UBD generated from the same data set (UBD-CASA MILA);
3. original UBD used by Fazly et al. (2010) and generated from the Manchester corpus in the CHILDES database (UBD-Manchester);
4. IBD generated by our framework as a result of simulations with 19 objects in the environment (which was the average context size in the analyzed CASA MILA dataset).[3]

For measuring the learning success at each moment, we used *effective ratio* calculated as the number of words that the learner has acquired at that time, divided by the number of words that she heard so far. The growth of the effective ratio over time is presented in Figure 2. Note that the size of *UBD (CASA MILA)* set is two times smaller than that of the original *CASA MILA* dataset, because only every other natural utterance could be included into UBD. For *UBD (Manchester)* and *IBD* the graphs show values averaged over 10 word learning simulations.

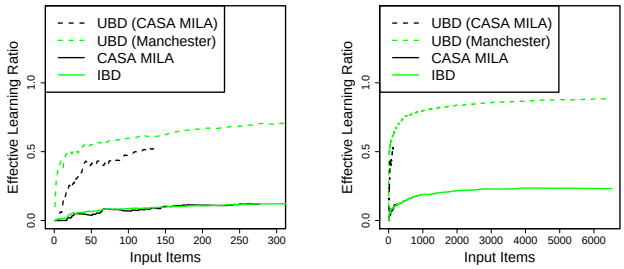


Figure 2: Overall model performance on four different datasets for 300 (left) and 6500 (right) utterances

The graph on the left shows the effective ratio over the course of 300 input items, which is slightly more than the size of the CASA MILA dataset. It is clear from these results that the performance of the word learning model is very similar when it is trained on data collected from CASA MILA and on data generated by our framework (IBD). In contrast, the model performs much better when it is trained on any of the UBD sets. This difference again suggests that UBD is not representative of what young language learners have access to, and a more realistic approach to data generation must be applied. The graph on the right shows the same measure over the course of 6500 utterances (the size of UBD-Manchester). The same pattern can be seen: there is a considerably large gap between the learning curves in UBD and IBD cases. It is also clear that in the latter case, the size of the input to the learner does not have to be constrained by the amount of data available in an existing collection.

Conclusion and discussion

We manually annotated a small dataset of video recordings of child–adult interactions and collected various types of co-

[3] Since one of the main parameters of the framework was the context size, we also investigated whether the learning process would vary with the number of objects in the environment, but our manipulations did not result in changing the overall learning pattern in terms of effective ratio.

occurrence frequencies of utterances, utterance types, accompanying actions and action types, action arguments and participants, and other objects available in the visual context. Using three quantitative measures, we compared the characteristics of these utterances and their surrounding scenes with the product of the most realistic existing approach to automatic generation of scene representations (Fazly et al., 2010). Our analyses show significant differences between the two datasets, and using an existing model of word learning as a case study further demonstrates that automatically generated utterance-based data simplifies the learning task to an unrealistic scale. However, manual annotation as an alternative approach (e.g., Yu & Ballard, 2007; Frank et al., 2013) is not scalable due to the limited quantity of the data available. The hybrid approach that we propose eliminates these problems: we present an input generation framework which can produce an infinite stream of child–adult interaction data containing both linguistic and visual information, whose statistical properties are closely comparable to those of manually annotated data.

Any data annotation or generation scheme inevitably incorporates assumptions about important components and information cues in language learning, which can be seen as built-in biases brought to the learning task. However, computational models need data and will benefit from any attempt to make this data more naturalistic.

An extension of the proposed framework can potentially provide certain interaction features such as the participants' focus of visual attention and head movement. Such extra features can allow computational models to systematically investigate the impact of interaction factors in language learning.

The dataset that we analyzed was limited in size and the interaction domain (toy playing). We add a parameter to our framework to account for potential variation in the size of the visual context. But humans' linguistic behavior (e.g., the structural and pragmatic characteristics of utterances) may also depend on the domain to some extent (e.g., Choi, 2000). Therefore, a larger and more diverse collection of interaction videos will provide a more realistic base for estimating the input generation probabilities in our framework. The larger CASA MILA corpus of interaction data that is currently under development (Vogt & Mastin, 2013a) is one suitable candidate for such expansion.

References

Behrend, D. A. (1990). Constraints and development: A reply to Nelson (1988). *Cog. Development*, *5*(3), 313–330.

Choi, S. (2000). Caregiver input in English and Korean: Use of nouns and verbs in book-reading and toy-play contexts. *J. of Child Lang.*, *27*(1), 69–96.

Chouinard, M. M., & Clark, E. V. (2003). Adult reformulations of child errors as negative evidence. *J. of Child Lang.*, *30*(3), 637–669.

Clark, A. (2001). Unsupervised induction of stochastic context-free grammars with distributional clustering. In *Proc. of CoNLL'2000* (pp. 105–112).

Fazly, A., Alishahi, A., & Stevenson, S. (2010). A probabilistic computational model of cross-situational word learning. *Cognitive Science*, *34*(6), 1017–1063.

Frank, M., Goodman, N., & Tenenbaum, J. (2008). A bayesian framework for cross-situational word-learning. In *Proc. of NIPS'2008* (pp. 457–464).

Frank, M., Tenenbaum, J. B., & Fernald, A. (2013). Social and discourse contributions to the determination of reference in cross-situational word learning. *Language, Learning and Development*.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2009). Frequency and contextual diversity effects in cross-situational word learning. In *Proc. of CogSci'2009* (pp. 2220–2225).

Kishimoto, T., Shizawa, Y., Yasuda, J., Hinobayashi, T., & Minami, T. (2007). Do pointing gestures by infants provoke comments from adults? *Infant Behavior and Development*, *30*(4), 562–567.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Lawrence Erlbaum Associates.

MacWhinney, B. (2010). Computational models of child language learning: An introduction. *J. of Child Language*, *37*(3), 477–485.

Matusevych, Y. (2012). *Modeling child–adult interaction: A computational study of word learning in context* Master's thesis, Tilburg University, the Netherlands.

Moore, C., Angelopoulos, M., & Bennett, P. (1999). Word learning in the context of referential and salience cues. *Developmental Psychology*, *35*(1), 60–68.

Niyogi, S. (2002). Bayesian learning at the syntax-semantics interface. In *Proc. of CogSci'2002* (pp. 697–702).

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*(1-2), 39–91.

Steels, L. (1996). Emergent adaptive lexicons. In *From Animals to Animats 4: Proc. of the Fourth International Conference on Simulation of Adaptive Behavior* (pp. 562–567).

Vogt, P., & Haasdijk, E. (2010). Modeling social learning of language and skills. *Artificial Life*, *16*(4), 289–309.

Vogt, P., & Mastin, J. D. (2013a). Anchoring social symbol grounding in children's interactions. *Künstliche Intelligenz*, *27*, 145–151.

Vogt, P., & Mastin, J. D. (2013b). Rural and urban differences in language socialization and vocabulary development in Mozambique. *Proc. of CogSci'2013*.

Yu, C., & Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, *70*(13-15), 2149–2165.

Yu, C., Smith, L. B., Shen, H., Pereira, A. F., & Smith, T. (2009). Active information selection: visual attention through the hands. *IEEE Transactions on Autonomous Mental Development*, *1*(2), 141–151.