A connectionist model for automatic generation of child-adult interaction patterns

Moinuddin M. Haque (m.m.haque@uvt.nl), Paul Vogt (P.A.Vogt@uvt.nl),
Afra Alishahi (A.Alishahi@uvt.nl), Emiel Krahmer (E.J. Krahmer@uvt.nl)
Tilburg Center for Cognition and Communication,
Tilburg University
PO Box 90153, 5000 LE Tilburg, the Netherlands

Abstract

This study introduces a neural network that models the social interactions from a video corpus. The corpus consists of recordings of naturalistic observations of social interactions among children and their environment. The videos are annotated multimodally including features like gestures. We explore how this video corpus can be utilized for modelling by training our model on a portion of the annotated data extracted from the corpus, and then by using the model to predict novel interaction sequences. We evaluate our model by comparing its automatically generated sequences to an unseen portion of the corpus data. The initial results show strong similarities between the generated interactions and those observed in the corpus.

Keywords: Neural Networks; Child Language Acquisition; Sequence Generation; Modelling Social Interactions.

Introduction

Children adapt to their environment and communication partners trough interaction. These interactions along with the linguistic information are richly augmented with social cues (such as eye gaze and gestures), and are suggested to facilitate child language developement (Tomasello & Todd, 1983; Iverson, Capirci, Longobardi & Caselli, 1999; Hollich, Hirch-Pasek & Golinkoff, 2000). Therefore, cognitive models of child language learning should take these cues into account and integrate such interaction-based features with linguistic input in the process of learning language.

One of the well-studied mechanisms for learning the meaning of words is cross situational learning (Quine 1960), which draws on word-referent co-occurrences that children observe from their environment. Many computational models have implemented (variations of) this mechanism using associative networks to predict a word form based on semantic features (e.g., Li & Farkas, 2002; Regier, 2005) or to discover statistical regularities in observations of linguistic labels and visual features or concepts (e.g., Siskind, 1996; Frank, Goodman & Tenenbaum, 2007; Fazly, Alishahi & Stevenson, 2010). Typically, such models treat learning as a unidirectional process where the learner focuses only on the linguistic cues and discards the social interactions. We refer to these models as *data-driven models of word learning*. To date, only a few models of word learning have incorporated interaction features. The data-driven models by Yu and Ballard (2007), and Frank, Tenenbaum and Fernald (2013) have incorporated information about eye gaze and pointing in a cross situational learning model. Another example of a data-driven model where multiple modalities are incorporated is the work Matusevych, Alishahi and Vogt (2013). In this model features, such as occurrence frequencies of utterances, utterance types, action types, action arguments along with participants and objects in the visual context, were used to simulate interactions in the context of playing with toys.

The language game model of Steels (2003) is an example of an *agent-based model* in which agents interact with each other, exchange utterances and can learn from each other. Typically, such models have been used to study language evolution. Various language game models have been used to investigate vocabulary development, also incorporating cross-situational learning (Smith, 2005; Steels & Loetzsch, 2008; Vogt & Haasdijk, 2010). However, these models tend to implement interactions between agents using toy languages and do not reflect naturalistic interaction patterns. Considering the restrictions of data-driven and agent-based approaches to word learning, the next natural extension is to integrate the two approaches, but the problem is finding large and rich datasets for training such models.

Large-scale corpora of child-adult conversations, such as CHILDES (MacWhinney, 2000), are available and provide us with naturalistic linguistic exchanges between children and adults. However, most of the corresponding audio and video files are not annotated with extra-linguistic (e.g., semantic information about the surrounding scene) or interaction-based cues (such as gaze and gesture) that are machine readable.

The CASA MILA corpus, which consists of longitudinal video recordings of 40 children interacting in naturalistic environments, is a corpus that has incorporated annotations for non-verbal social cues (Vogt & Mastin, 2013a). The video frames are richly annotated based on the observed interactions between children and their caregivers, and thus provides a valuable resource for modeling child-adult interaction. However, it only covers 1.5 hours of recording for each child, which is hardly enough for training a computational model of child language development. What we need is an automatic input-generation engine that can replicate the interaction patterns and their statistical properties observed in a corpus such as CASA MILA, without the quantitative limitations of such a corpus.

The current paper presents a study to generate novel interactions based on observations from the corpus. One approach to create more data can be simply by copying the already annotated data multiple times. The limitation of this approach is that there will be no new interactions present in

the data, and the rigidity of the interactions remains. Yet, the flexibility of having new interaction patterns in the data is one of our aims. We therefore present a method for discretizing the continuous video and annotation data, and using these data to train a neural network that generates new interactions. We evaluate this method by analysing the newly generated sequences against the original annotated sequences, as well as three different baseline models.

Methods

Data

The CASA MILA corpus contains video recordings of 40 different children at home, interacting with one or more communication partners. These recordings are from three different cultures: rural and urban Mozambique, and the Netherlands. Each recording contains naturalistic observations of interactions between infants and their communication partners at their home. The corpus is longitudinal in nature, with recordings taken at children's ages of 13-, 18- and 25-months old.[1] For this study we will only use the data from the 13-months old children from the Netherlands. This was done, because these mainly contained one-to-one interactions, thus simplifying our problem.

The corpus was annotated for a variety of tiers, five of which we use for the present study: child engagement, child-directed speech, child-directed gesture, child-speech and child-gesture. In the recordings we only focus on those parts in which the child interacted with someone. The annotated features are hierarchically organized. On the top layer is the child's joint engagement level (Mastin & Vogt, 2016), as described in Table 1.

Table 1. Joint engagement levels.

Name	Description	Example
Persons engagement	Infants interact with another person by responding to the other person or by trying to start an exchange.	The infant responds with a smile to the mother's voices; infant reaches toward the mother.
Passive joint attention	Infants play with an object that is also the focus of another person's attention but they do not acknowledge the other person's attention.	The infant plays with a toy car. The mother says: "What a nice car!", but receives no response at all from the infant.
Shared joint attention	Infants share the attention to an object or event with the interlocutor, but they do not share a mutual goal in the interaction.	The mother offers the infant a toy to play with, the infant looks from the mother to the toy, but does not respond otherwise.
Coordinated joint attention	Infants share the attention to an object or event with the interlocutor, and they clearly share a mutual goal in the interaction.	The mother offers the infant a toy to play with, the infant looks to the mother and the infant takes the toy and starts playing.

At the lower layers, speech and gestures (e.g., pointing, showing or reaching) are annotated for both caregiver and child.[2] In this study, we let the machine learn to predict the joint engagement level, and when the children and communication partners produce utterances and/or gestures, but not what words or gestures are actually used.

Feature selection

Figure 1, illustrates how the original annotations are transformed to provide the learning algorithm a simplified representation of the input. Figure 1(a) shows a small fragment of an actual timeline showing the engagement level (top row), child-directed speech (second row), child-directed gesture (third row), child speech (fourth row) and child gesture (bottom row). The highlighted regions show where speech or gesture was observed.

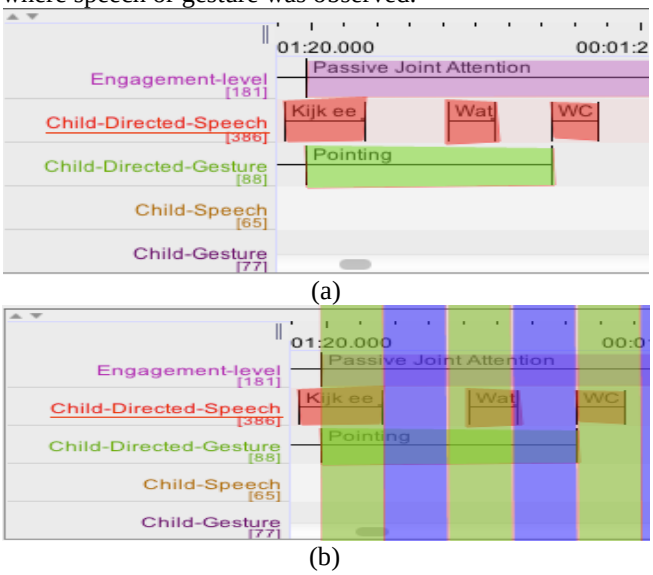

Figure 1. (a) Simplified annotated timeline with highlights. (b) Spliced timeline (not to scale).[3]

To process our data, the original time sequence is broken down into slices of 200 milliseconds duration, with the purpose of capturing the information in the annotations (Figure 1(b)). The duration of 200 milliseconds was determined after trial and error to capture significant information without having too many unchanged sufficient. These time slices allow us to represent the state of an interaction at time *t* as an 8-bit vector *x(t)*, where presence of activity is represented as 1 and absence as 0. To construct these vectors, engagement levels are represented with four bits, of which exactly one bit has the value of 1 at any given time (see Table 2, rows 1-4). Since we are only interested in predicting when someone speaks or gestures, the remaining four bits encode whether or not child-directed speech, child-directed gesture, child speech or child gesture was present at time *t*. This binary vector representation is then used to serve as input for our neural network.

[1] For more details on the recording procedures, consult Mastin and Vogt (2016) or Vogt, Mastin, and Schots (2015).

[2] Consult Vogt et al. (2015) for the transcriptions of speech, and Vogt and Mastin (2013b) for the annotation of gestures.

[3] Created with ELAN (Sloetjes & Wittenburg, 2008).

Table 2. The bit vector representation.

Position	*Represents*	*Possible values*
1	Persons engagement	0/1
2	Passive joint attention	0/1
3	Shared joint attention	0/1
4	Coordinated joint attention	0/1
5	Child-directed speech	0/1
6	Child-directed gesture	0/1
7	Child speech	0/1
8	Child gesture	0/1

Model

We developed a neural network model to generate novel interaction sequences based on the naturalistic patterns observed in the annotated corpus, and evaluated this on how well the model can recreate an unseen sequence. Treating the bit sequence as a time series with each vector as a given state, the model is trained to predict the next possible stage given a previous set of states.

We used a non-linear autoregressive neural network with external input (NARX) (Haykin, 1999; Lin, Horne, Tiňo & Giles, 1996; Gao & Er., 2005), which is a class of neural networks that is well suited for training nonlinear systems and time series. The NARX is a recurrent dynamic network with feedforward connections enclosing several layers of the network. The network is used to predict the next value of the input signal. The NARX architecture was chosen over others for its success with predicting time series.

Since the true output is available during the training of the network, one can create a 'series-parallel architecture', in which the true output is used instead of feeding back the estimated output, as a series-parallel architecture (Figure 2). During training the network is set in the series-parallel architecture. Once the network has finished training the network is changed into the closed loop, standard NARX 'parallel architecture' to make it usable for predicting the next state in the evaluation phase (Figure 3). During this prediction stage the output is fed back to the input of the feed forward neural network. The output of the NARX network is an estimate of the output of the nonlinear dynamical system that is being modelled. Making this distinction has two advantages: First, in the series-parallel architecture the input to the feedforward network is more accurate. Second, this series-parallel architecture has a purely feedforward network, and static backpropagation can be used for training.

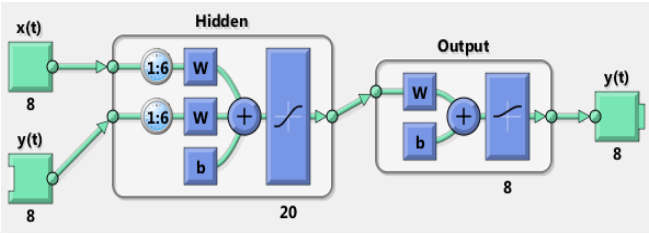

Figure 2: The Series-parallel Architecture of the NARX network in open loop.

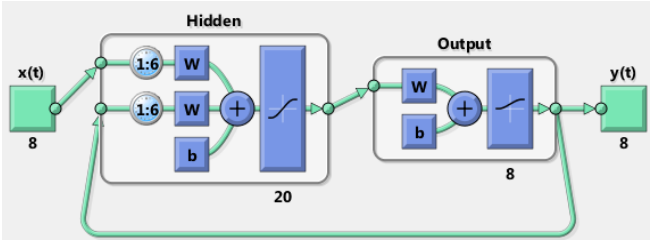

Figure 3: The Parallel Architecture of the NARX network in closed loop.

During training, the network receives two input vectors, *x* (input time series) and *y* (output time series) each represented as a vector with 8 bits. The input time series corresponds to the bit sequence we are training at time *t*. The output series corresponds to the observed output at t+1. To introduce a temporal memory the network delay is set to 6, meaning that each new vector is predicted based on the 6 previous input vectors. This parameter was derived empirically. The hidden layer has 20 neurons. For the transfer function in the hidden layer a hyperbolic tangent sigmoid transfer function was used. At the output layer a log-sigmoid transfer function is used to get the final result in the format of a matrix where each row corresponds to the target annotations.

The network is trained with a function that updates the weights and bias values according to Levenberg-Marquardt optimization (Levenberg, 1944). It minimizes a combination of squared errors and weights, and then determines the correct combination so as to produce a network that generalizes well. The process is called Bayesian regularization. The error calculation is done using mean squared normalized error.

Experimental setup

From the corpus we constructed 2 different data sets. We will discuss each separately as study 1 and study 2. For both studies, data from 12 children in the 13-month age group of the Netherlands dataset was used.

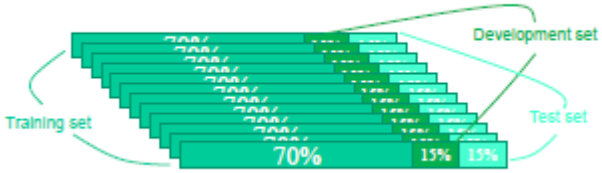

Figure 4: Division of data for study 1.

Study 1. Each of the children's data is broken into 3 distinct parts: a training section, a development section, and a test section, with a size of 70%, 15%, and 15% for each section respectively. The individual parts of the sections are joined together to form aggregated sets. The training set is used for training of the neural network. The development set is used

to fine tune the parameters of the network. The test set is used to evaluate the performance of the network. Figure 4 shows the division of data for study 1.

Study 2. The development set and training set are formed by selecting data from 11 of the 12 children. This data is broken into two sections per child. The development sections contained 15% of the data and, aggregated, these formed the development set. The training sections contained 85% of the data and formed the training set. Data from the 12th child is taken as the test set. Figure 5 shows the division of data for study 2.

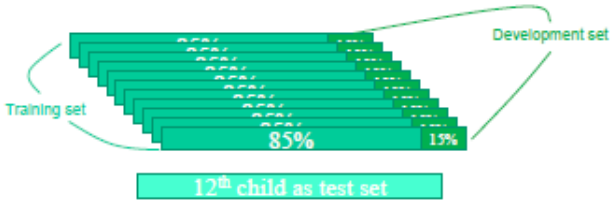

Figure 5: Division of data for study 2

Evaluation

Baselines. To evaluate the accuracy of the neural network generated sequences (NNG), we implemented three baseline models for predicting sequences. In each of these baselines the size of the generated sequence equals the length of the test set. The three baseline models are defined as:

- Complete random generation (CRG). This creates a sequence by putting together random slices taken from a set of all distinct slices observed in the training set.
- Attribute-based generation (ABG). This sequence is created by calculating the probability distributions for each of the attributes annotated in the individual slices (e.g., child gesture, mother gesture or engagement level). In this model it is assumed that the attributes are independent of each other and the predictions are based on their distribution probability only.
- Transition based generation (TBG). This is formed by calculating the transition probability for each of the unique slices, as observed in the training set. Once these probabilities are known a new discrete sequence is generated using the transition probabilities.

These three generations present us with a comparative baseline to test the effectiveness of the neural network generated sequences.

Transitions. To capture the patterns in interactions we will evaluate the transitions in interaction states. Transitions occur when at least one bit changes between two time steps *t* and *t+1*. For example, when a bit sequence "00010001" (representing child and mother having coordinated joint attention where the child is gesturing) is followed by "00011001" (representing child and mother having coordinated joint attention where the child is gesturing and mother is speaking).

Measures. To measure the performance of the neural network, we evaluated the sequences generated at a micro and a macro level. At the micro level, the generated sequence was aligned with the test set to check for matches. A match is said to occur when slices match each other exactly. We measured *Accuracy* as the percentage of the test set that had a perfect match.

For the macro level analysis we compared the generated *distributions* with the transition distribution observed in the test set. To do so we calculated the number of times a transition took place to create a probability distribution corresponding to each sequence. These probability distributions were then compared with the test set's transition probability distribution using the *Hellinger distance*, which was used to quantify the similarity between two probability distributions (Hellinger, 1909). The Hellinger distance forms a bounded metric on the space of probability distributions over a given probability space. Mathematically this is calculated by taking the square root of the distance between two vectors. The closer the Hellinger distance is to 0, the more similar two distributions are. The maximum distance of 1 is obtained when there is no overlap between the two distributions.

Results

Study 1

Accuracy. Figure 6 provides the accuracy of the different sequence generation models. As we can see, replicating the exact time series is difficult. The accuracy of the CRG (4%) is expectedly very low.

The low accuracy for ABG (11%) shows that the assumption that the attributes are independent of each other is too simplistic and that there are meaningful dependencies between them, which are useful for sequence generation.

The accuracy gain through NNG (21%) over the TBG (19%) is small. The TBG is very faithful to the training data and therefore makes few mistakes, but it cannot generalize beyond what it is has seen in the training data. To get a better understanding of the difference in the sequences generated by these two models we look at the Hellinger distance.

Time series distribution. Figure 7 shows the Hellinger distances calculated for each of the generated sequences. The macro level analysis displays a much better trend than accuracy measures. When looking at the macro-level analysis, the NNG has a Hellinger value of 0.30, which is considerable improvement over the baseline models. The Hellinger values for the baseline models are close to 1, meaning the compared distributions are dissimilar. In short the sequences generated by the NNG have more overall similarity with the test sets than those generated by the other baseline methods.

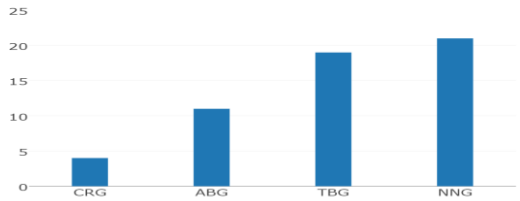

Figure 6. Accuracy measures for different generations.

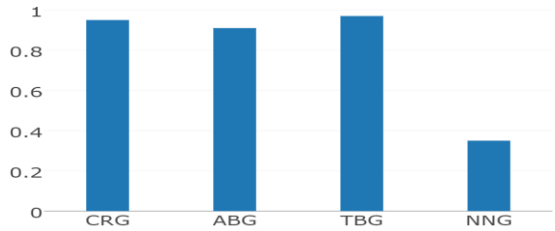

Figure 7: Hellinger Distance for different generations.

To continue our analysis, we provide a visualization of the transitions and their probabilities as observed. Figure 8 shows the probability distributions for the NNG sequence and the test set. We can see that the probabilities of the highly frequent transitions are closely replicated by the network, but it fails to replicate the low frequency transitions.

Figure 8: Transition probabilities in the test set (blue bars) and the NNG (yellow bars). The x-axis represents the different transitions, and the y-axis shows the probability.

Study 2

In the second study, we investigated how the model would behave when faced with data from a child it has not been trained on. Accuracy of the generated sequence for the NNG was approximately 5%, showing a much lower performance compared to study 1.

Doing the macro level analysis we found the Hellinger distance to be close to 1, which means that the generated sequences distribution is very different from the one observed in the test set. Although there are some matches for the slices produced, yielding accuracy greater than zero, the distribution of transitions differs.

Discussion

In this paper we took a novel approach to generating new multimodal interaction sequences based on corpus data, using neural networks.

The interaction sequences produced by the different generation techniques shows that the neural network generated patterns have higher accuracy and more similar transition distribution than the baseline methods. Although the TBG and NNG data have similar levels of accuracy, when it comes to the transitional distribution, we observe a large difference, showing that NNG generates interactions more similarity with the training data.

While the NNG yields the highest accuracy, it is still far from perfect. The reason for this is that the interactions the network is trained on represents a non-linear complex dynamical system, whose exact time series is extremely hard to replicate. Although neural networks have shown to be universal approximators, they have difficulty modelling time series. While NARX networks perform better than others when it comes to time series modelling, they still have problems learning long term dependencies due to vanishing gradients (Diaconescu, 2008). The behavior of the network is highly dependent on the size of the input sequence that represents the temporal memory of the model. However, the NARX model lacks a decent procedure for optimizing this size.

The Hellinger distance gives us a better performance, as it looks at the distributions of the series and not exact locations. For the NNG, the transitions observed in the generated sequence are more similar to the test set than the baseline sequences, but the distance obtained with the present method is still insufficiently close to zero. Highly frequent transitions in the interactions are fairly well replicated, but many transitions observed in the test set are not. Moreover, the network generates sequences that have not been observed. The reason for these discrepancies are likely due to the possibility that the test set contains transitions that have not occurred in the training data and vice versa. Further analysis is required to verify whether this is, indeed, the case.

For study 2, the Hellinger distance approaches 1, which means that the distributions of transitions are not well generalized for replicating interactions of an unseen child. One reason for this is that there are substantial individual differences between the interactions the children engage in (Vogt et al., 2015). So, although they interact using the same gestures in conjunction with speech, the frequencies with which gestures are used vary considerably, as do the sequences of interactions (Vogt & Mastin, 2013b). Since the network has a memory that helps generating sequences based on previous observations, testing the network on an unseen test set is likely the cause for the marked difference in the transition probabilities.

In order to reduce the complexity of the learning task, we chose to represent the speech and gesture in the annotated corpus as bitstrings indicating the presence or absence of speech or gesture. However, doing this reduced the amount of information contained in the training data. While maintaining all words and different gestures is likely too complex with the amount of data, using more informative categories of speech or gesture might improved the results.

To conclude, our initial attempt to replicate patterns of interactions between young children and their caregivers observed in a video corpus yields mixed results. Our method can replicate interactions of aggregated children reasonably well, but it cannot generalize to a previously unseen child, nor does it fail to replicate exact sequences in the time series well. In later stages we intend to introduce more features into the training along with exploring other modelling paradigms, such as incorporating the corpus data more directly in an agent-based model that can then be trained to interact following the patterns observed in the data.

Acknowledgements

This study was funded by the Netherlands Organization for Scientific Research (NWO) with a grant in the Natural Artificial Intelligence program.

References

Diaconescu, E. (2008). The use of NARX neural networks to predict chaotic time series. *Wseas Transactions on Computer Research*, *3*(3), 182-191.

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

Frank, M. C., Tenenbaum, J. B., & Fernald, A. (2013). Social and discourse contributions to the determination of reference in cross-situational word learning. *Language Learning and Development*, *9*(1), 1-24.

Frank, M., Goodman, N. D., & Tenenbaum, J. B. (2007, December). A Bayesian Framework for Cross-Situational Word-Learning. In *NIPS* (pp. 457-464).

Gao, Y., & Er, M. J. (2005). NARMAX time series model prediction: feedforward and recurrent fuzzy neural network approaches. *Fuzzy sets and systems*, *150*(2), 331-350.

Haykin, S. (1999). *Neural Networks, Second Edition*. Pearson Education.

Hellinger, E. (1909). Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen. *Journal für die reine und angewandte Mathematik*, *136*, 210-271.

Hollich, G. J., Hirsh-Pasek, K., Golinkoff, R. M., Brand, R. J., Brown, E., Chung, H. L., ... & Bloom, L. (2000). Breaking the language barrier: An emergentist coalition model for the origins of word learning. *Child Language*, 43(2): 235-264.

Iverson, J. M., Capirci, O., Longobardi, E., & Caselli, M. C. (1999). Gesturing in mother-child interactions. *Cognitive Development*, *14*(1), 57-75.

Levenberg, K. (1944). A method for the solution of certain non–linear problems in least squares. 164-168.

Li, P., Farkas, I., & MacWhinney, B. (2004). Early lexical development in a self-organizing neural network. *IEEE Transactions on Neural networks*, *17*(8), 1345-1362.

Lin, T., Horne, B. G., Tiňo, P., & Giles, C. L. (1996). Learning long-term dependencies in NARX recurrent neural networks. *IEEE Transactions on Neural Networks*, *7*(6), 1329-1338.

MacWhinney, B. (2000). The CHILDES project: Tools for analyzing talk: Volume I: Transcription format and programs, volume II: The database. *Computational Linguistics*, *26*(4), 657-657.

Mastin, J. D., & Vogt, P. (2016). Infant engagement and early vocabulary development: a naturalistic observation study of Mozambican infants from 1;1 to 2;1. *Journal of Child Language*, FirstView, 1-30.

Matusevych, Y., Alishahi, A., & Vogt, P. (2013). Automatic generation of naturalistic child–adult interaction data. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society* (pp. 2996-3001).

Quine, W. V. O. (1960). *Word and object*: Cambridge University Press.

Regier, T. (2005). The emergence of words: Attentional learning in form and meaning. *Cognitive Science*, *29*(6), 819-865.

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

Sloetjes, H., & Wittenburg, P. (2008, May). Annotation by Category: ELAN and ISO DCR. In *LREC*.

Smith, A. D. (2005). Mutual exclusivity: Communicative success despite conceptual divergence. *Language Origins: perspectives on evolution*, 372-388.

Steels, L. (2003). Evolving grounded communication for robots. *Trends in Cognitive Sciences*, *7*(7), 308-312.

Steels, L., & Loetzsch, M. (2008). Perspective alignment in spatial language. *Spatial Language and Dialogue*, 70-89.

Tomasello, M., & Todd, J. (1983). Joint attention and lexical acquisition style. *First Language*, *4*(12), 197-211.

Vogt, P. & Haasdijk, E. (2010) Modelling social learning of language and skills. *Artificial Life, 16(4): 289-310*

Vogt, P., & Mastin, J. D. (2013a). Anchoring social symbol grounding in children's interactions. *KI-Künstliche Intelligenz*, *27*(2), 145-151.

Vogt, P., & Mastin, J. D. (2013b). Rural and urban differences in language socialization and early vocabulary development in Mozambique. In *Proceedings of the 35th annual meeting of the Cognitive Science Society*(pp. 3787-3792). Austin, TX: The Cognitive Science Society.

Vogt, P., Mastin, J. D., & Schots, D. M. (2015). Communicative intentions of child-directed speech in three different learning environments: Observations from the Netherlands, and rural and urban Mozambique. *First Language*, *35*, 341-358.

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