Modeling Utterance-mediated Attention in Situated Language Comprehension

Ján Švantner (svantner@fmph.uniba.sk)
Igor Farkaš (farkas@fmph.uniba.sk)
Department of Applied Informatics, Comenius University
Mlynská dolina, 824 48 Bratislava, Slovakia

Matthew Crocker (crocker@coli.uni-sb.de)
Department of Computational Linguistics and Phonetics, Saarland University
66123 Saarbrücken, Germany

Abstract

Empirical evidence from studies using the visual world paradigm reveals that spoken language guides attention in a related visual scene and that scene information can influence the comprehension process. Here we model sentence comprehension using the visual context. A recurrent neural network is trained to associate the linguistic input with the visual scene and to produce the interpretation of the described event. The feedback mechanism in the form of sigma-pi connection is added to model the explicit utterance-mediated visual attention behavior revealed by the visual world paradigm. The results show that the network successfully learns sentence final interpretation and also demonstrates the hallmark anticipation behavior of predicting upcoming constituents.

Keywords: connectionist modeling; sentence comprehension; attentional mechanism; visual scene

Introduction

During the last decade, research in human language comprehension has progressed well beyond the examination of the syntactic and semantic properties of words and sentences considered in isolation. Detailed on-line evidence for how people comprehend visually-situated language has come from the visual world paradigm (see Huettig, Rommers, and Meyer (2011) for a recent review). The visual world paradigm takes advantage of the listeners' tendency to look at relevant elements of the visual scene as they are mentioned or anticipated (which is typically measured by eye-tracking). Specifically, it has been shown that spoken language can guide attention in a related visual scene and that scene information can immediately influence the comprehension process (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). Findings have revealed the rapid and incremental influence of visual referential context (Spivey, Tanenhaus, Eberhard, & Sedivy, 2002; Tanenhaus et al., 1995) and depicted events (Knoeferle, Crocker, Scheepers, & Pickering, 2005) on ambiguity resolution in online-situated utterance processing. Further research demonstrated that listeners even anticipate likely upcoming role fillers in the scene based on their linguistic and general knowledge (e.g. Kamide, Altmann, and Haywood (2003)). Knoeferle and Crocker (2006) identified several cognitive characteristics based on the above mentioned findings, claiming that situated language comprehension is incremental, anticipatory, integrative, adaptive, and coordinated, which led to the proposal of the coordinated interplay account (Cia).

The recent CiaNet model (Mayberry, Crocker, & Knoeferle, 2009) instantiates the Cia proposal and accounts for a range of empirical findings. CiaNet is a recurrent sigma-pi neural network that models the rapid use of scene information, exploiting an utterance-mediated attentional mechanism. The model was shown to achieve very good performance (both with and without scene contexts), while also exhibiting hallmark behaviors of situated comprehension, such as incremental processing, anticipation of appropriate role fillers, as well as the immediate use and priority of depicted event information through the coordinated use of utterance-mediated attention to the scene. Several other models that link language with the visual world, do exist, including those mentioned in the very recent review (Huettig et al., 2011), as well as Yu, Ballard, and Aslin (2005); Gold and Scassellati (2007). These models emphasize situated lexical learning and processing, however, and there remain very few attempts to model the compositional and incremental nature of visually situated sentence comprehension.

Inspired by above mentioned CiaNet, we investigate a more general network architecture that also learns to adapt the attention mechanism to help the network focus on (and predict upcoming) relevant constituents and in principle allows generalization to more complex scenes (the attention mechanism in CiaNet is restricted to favor one of the two concurrent events). Our model also differs from CiaNet (and other models) in that inhibition operates at both the object and event levels (rather than only at the event level) that are assumed to underlie the cognitive representation of the visual scene. In addition, our work assumes that visually grounded lexical representations are in place, focusing rather on the compositional aspects of situated sentence comprehension.

The model

The network architecture, shown in Fig. 1, is based on a simple recurrent network (SRN) (Elman, 1990). The network reconciles an incrementally presented utterance with a representation of the current visual context to incrementally and predictively recover the situated meaning representation. The model takes situational inputs coupled with linguistic inputs and is trained to produce the representation of the target event, mentioned in the linguistic utterance. The scene representations stand for encoding the objects and events in the visual world, the linguistic representations are presented as short sentences. In each trial, the scene representation is presented at the input and the associated sentence is presented at the linguistic input, one word a time. The network task is produce a (partial) situational representation at the output. This process is mediated by the hidden layer that combines scene-related representations with the symbolic language. The target is available at the output during the entire sentence processing. The explicit feedback (from the output) is added to the network using a sigma-pi mechanism to model the process of focusing attention to relevant constituents (objects) shown in the visual scene and mentioned in the associated utterance.

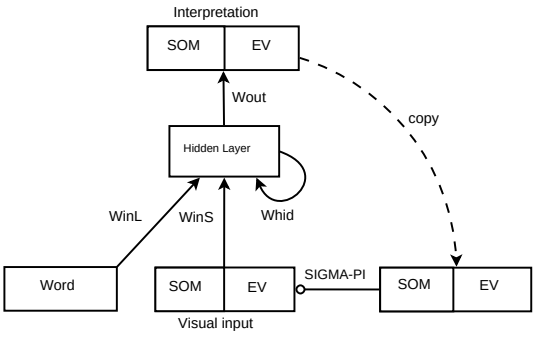

Figure 1: Model architecture with an utterance-mediated attentional mechanism. For description see the text.

Scene representations

The scene representations consist of two levels – the object level (SOM) and the event level (EV). The objects may be the constituents of events – corresponding to physical agents/patients that can be focused on – whereas the event level refers to specific ongoing actions in the concrete context (with given semantic roles, i.e. known agent and patient). The scene is assumed to consist of two events that may or may not share a constituent (e.g. an agent of one event can be a patient of another event), plus a few distractors (see Figure 2). In contrast with Mayberry et al. (2009), we can also encode more than two events because the types of representations allow that extension in principle (which would probably lead to lower accuracy of the model).

Objects Objects include human agents (e.g. toddler/woman), animate agents (e.g. dog/donkey) and one artificial agent (robot) that can be involved in various meaningful activities, with or without a patient. Agents can operate on machines[1](forklift/bulldozer), on objects (e.g. barrel/house) or food items (e.g. apple/juice). The actions include moving (e.g. walks/sits), physical manipulation (e.g. lifts/holds), socially oriented activities (e.g. greets/looks-at) and "sustenance" (eats/drinks). Agents and patients are manually assigned binary features that encode various physical and functional properties and form 40-dim. vectors $\mathbf{c}_\mathrm{A}$ and $\mathbf{c}_\mathrm{P}$, respectively. Analogically, actions are described by 16-dim. vectors [2] of binary features $\mathbf{c}_\mathrm{V}$.

We have used the standard self-organizing map (SOM) (Kohonen, 1990) to learn the localized representations of objects. The SOM is constructed in advance using only agent $\mathbf{c}_\mathrm{A}$, patient $\mathbf{c}_\mathrm{P}$ and distractor $\mathbf{c}_\mathrm{D}$ inputs, one at the time. The SOM is trained to provide a topographically organized map of objects according to their hand-designed semantic features. Each object is represented in the SOM by 3 most active units, focused around the winner (best matching unit), all other units are set to zero. The activity of unit i is calculated as $y_i = \exp(-\|\mathbf{x} - \mathbf{w}_i\|)$, where $\mathbf{w}_i$ is the unit's i weight vector and $\mathbf{x} \in \{\mathbf{c}_\mathrm{A}, \mathbf{c}_\mathrm{P}, \mathbf{c}_\mathrm{D}\}$. The activity of the three most active units is rescaled so that $y_\mathrm{bmu} = 1$. Since these object representations are mostly localist, they do not interfer with one another in the map. The SOM size was chosen to have 64 units to allow unambiguous learning of each object representation (by assigning it a separate winner). The purpose of using 3 most active units (instead of just a winner) is to allow an activation overlap between similar objects with neighboring winners (this helped the model to generalize better). Actions are excluded from SOM training; they are included only in the event-level representation $\mathbf{e}_\mathrm{in}$. The scene representation on the object level contains the superimposed representations of all relevant objects (showing that all objects are simultaneously present) plus several distractors resulting in SOM activation $\mathbf{c}_\mathrm{in}^\mathrm{all} = \mathbf{c}_\mathrm{in}^{(1)} \oplus \mathbf{c}_\mathrm{in}^{(2)} \oplus \mathbf{c}_\mathrm{D}$.

Events To obtain representations $\mathbf{e}_\mathrm{in}$ of events, an auto-associative network (AAN), modeled by a two-layer perceptron (i.e. with one hidden layer) is pretrained on vectors $[\mathbf{c}_\mathrm{A}\ \mathbf{c}_\mathrm{V}\ \mathbf{c}_\mathrm{P}]$ to form the compressed distributed representations at the hidden layer with 48 units. Patient $\mathbf{c}_\mathrm{P}$ is optional, so its components are set to zero in case of its absence. The input size dimension for training AAN off-line was 40+16+40=96 dimensions. The

[1]Machines can serve as agents of some actions, too (e.g. lift, push).

[2]Actually, they consist of only 8 binary features, but these were doubled to increase the differentiation of compressed event representations, performed by AAN module.

Figure 2: Example of the depicted scene that is assumed to consist of two events (*Boy chases dog* and *Girl looks-at boy*) and two unrelated distractors (*house, sparrow*). The two events share the constituent *boy*.

functionality of the trained AAN was checked via accuracy of compressed representations using the encoding and decoding of novel agent-action-patient triplets. The accuracy almost reached 100% for the testing data.

Once the AAN is trained, the event-level representation corresponding to the scene is taken as a superposition of two (compressed) representations of events, resulting in the vector $\mathbf{e}_{\text{in}}^{\text{all}} = \mathbf{e}_{\text{in}}^{(1)} \oplus \mathbf{e}_{\text{in}}^{(2)}$. The vector components are constrained in the interval [0,1]. Using the superposition is analogous to that of used in CIANET – it encodes simultaneous information provided to the subject as the visual input. However, in CIANET the representational medium is separated whereas in our model it is shared. Unlike localist object representations, the superposition of distributed event representations leads to an overlap between the two codes which expectedly makes the decompression task more difficult. A scene consists of two events, with 50% possibility of sharing one constituent (i.e. if the agent of one event matches the patient of another event, or if two events share the patient). Some elements of the event vector could become larger than one after superposition (i.e. if both events had the same component very active). The elements of an event vector were normalized by value of the most active element.

Linguistic inputs

The lexicon consists of 40 words, with one-to-one mapping to the objects/actions. Words are treated as symbols and are assigned one-hot codes with 40 dimensions creating an input $\mathbf{l}_{\text{in}}$. The sentences have a SV(O) form, such as 'Toddler looks-at crate' or 'Woman walks.'

Network activations

The model has two output slots – $\mathbf{e}_{\text{out}}$ is expected to predict the representation of the target event and $\mathbf{c}_{\text{out}}$ is the object-level output that, analogically, tries to activate the target objects, taking part in the described event. Together, $\mathbf{e}_{\text{out}}$ and $\mathbf{c}_{\text{out}}$ form the situational output. The model has no linguistic output.

The activation of the hidden layer of A-SRN at time t is computed as

$$\begin{aligned}\mathbf{a}_{\text{hid}}(t) &= \sigma(\mathbf{W}_{\text{inL}}.\mathbf{l}_{\text{in}}(t) + \mathbf{W}_{\text{inS}}.(\mathbf{s}_{\text{in}}(t).*\mathbf{a}_{\text{out}}(t-1)) \\ &\quad + \mathbf{W}_{\text{hid}}.\mathbf{a}_{\text{hid}}(t-1))\end{aligned}$$

where the scene representation $\mathbf{s}_{\text{in}} = [\mathbf{c}_{\text{in}}^{\text{all}}, \mathbf{e}_{\text{in}}^{\text{all}}]$, '.*' denotes component-wise multiplication of the two vectors (implementing sigma-pi connection) and σ is the standard logistic function $\sigma(x) = 1/(1+\exp(-x))$. Sigma-pi connections (Rumelhart, Hinton, & Williams, 1986) implement the modulation mechanism on a component-wise basis, i.e. for each unit (propagation of the afferent input is modulated by feedback input). To avoid propagation of misleading activation from the previous sentence, the sigma-pi activation is excluded at the beginning of each sentence, leaving only $\mathbf{s}_{\text{in}}(t)$ as the scene input.

The network output is computed as

$$\mathbf{a}_{\text{out}}(t) = [\mathbf{c}_{\text{out}}(t), \mathbf{e}_{\text{out}}(t)] = \sigma(\mathbf{W}_{\text{out}}.\mathbf{a}_{\text{hid}}(t)).$$

and feeds back with one-step delay to be multiplied with the network input.

Network training

We focused on the sigma-pi network (A-SRN) but also included SRN for comparison. For reasons explained in Results section we also tested a third model whose architecture falls between A-SRN and SRN and its input representation is calculated as (with $\gamma = 0.3$)

$$\mathbf{s}_{\text{in}}^{'}(t) = \gamma\ \mathbf{s}_{\text{in}}(t) + (1-\gamma)\ \mathbf{s}_{\text{in}}(t).*\mathbf{a}_{\text{out}}(t-1).$$

This linear combination guarantees that input representation remains preserved to a certain degree (given by γ) which is desirable in cases when output inhibition incorrectly inhibits all inputs, hence hindering the correct output of the network. This may happen after processing the first word in the sentence when the model's prediction of the target is not very accurate.

We systematically looked for optimal model parameters which were then used in testing the model and performing comparisons as described below. The hidden layer of all networks had 150 hidden units. Networks were trained with back propagation through time algorithm (Rumelhart et al., 1986) by propagating the error after each word, using the learning rate 0.01.

We generated 10,000 scenes, each of which was associated with two events. The model's attention was driven by the linguistic input to the single – major event of each situation. All generated events were consistent with the world, obeying semantic constraints. With each scene representation, a number of distractors (ranging from 0 to 3) was added to the input, taken from the pool of remaining agent/patient objects. Randomly chosen 70% of situations were used for training and the remaining 30% for testing. Data sets were distinguished by major

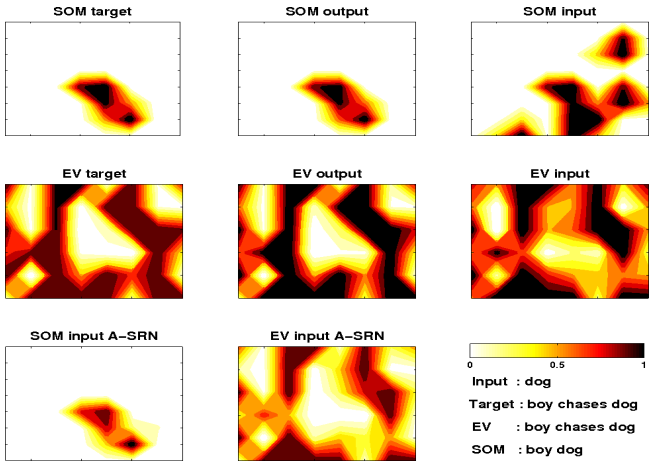

Figure 3: Example of a behavior of a trained A-SRN at the end of sentence 'Boy chases dog.' For explanation, see the text.

events used in their scenes. Model accuracy was evaluated using 5-fold cross validation.

The illustration of a trained A-SRN during processing at the sentence 'Boy chases dog' is shown in Figure 3, and corresponds to the scene shown in Figure 2. SOM-related graphs contain 8×8 units, EV-related graphs contain 48-dim. vectors, reshaped to 8×6 matrix for convenience. [3] On the right, SOM input is the composition of various objects (including distractors), EV input is the superposition of two events. Both inputs are presented to the network at the sentence beginning. On the left, both targets comprise only information about the target event (and the pertaining objects). At the bottom, both inputs become overridden by the feedback attentional mechanism that filters out irrelevant objects and non-target event information. Visual inspection of the network outputs (in the middle) reveals that they match well with both corresponding targets.

Performance evaluation

In order to evaluate the output accuracy, we need to interpret the model output. Since it consists of two different components (SOM and EV), we evaluate both. For testing the accuracy of $\mathbf{e}_{\text{out}}$ we decode this output part (using the hidden-output weight matrix of AAN) and count the percentage of correct decodings in the test set. Regarding $\mathbf{c}_{\text{out}}$, we compare the output with all possible combinations of SOM representations of objects, i.e. $\mathbf{c}_{\text{tgt}}$. Analogically, we count the percentage of matches (for both agents and patients). All measures can be evaluated after each word presented, to capture the progress during sentence parsing. We looked at the output accuracy at the end of sentences, and also on network's anticipatory behavior, that is, its prediction of upcoming constituents during sentence processing (i.e. predicting an action when reading a subject word, and predicting a patient when reading a subject and/or verb).

[3] The plots are interpolated, so they look smoother.

Quantitative measures used

We first explain all measures used in Tables 1–2 and in the text. All measures are quantified by values between 0 and 1, reflecting the accuracy of the measure. EV quantifies output accuracy of $\mathbf{e}_{\text{out}}$ decoding at the end of sentence. If both decoded objects and the action match the targets, the event representation is considered successful. Decoding in SOM (of agent/patient pairs, or only agents) is considered successful if both match the targets. In addition, we looked at several prediction measures (calculated before sentence end), that are related to the concrete constituents of an event (action, patient). These measures were evaluated with respect to various degrees of consistency. The predicted action/patient is considered correctly decoded: (a) with respect to the target if it matches it, (b) with respect to the world if it exists in the training corpus in the given context, (c) with respect to the current scene if it is present in it (albeit not focused on).

Results

Results in all tables refer to the testing data (accuracy on training data was consistently somewhat higher). We looked at three things when evaluating model performance, the motivation is explained below. First, we compared the accuracy of three models at the end of sentence; second, we manipulated the availability of the scene information during training and investigated its effect on model behavior; third, we looked at predictive properties of the model, i.e. the anticipation of upcoming constituents before the sentence end.

Model comparison At first, we focused on network output at the end of sentences. The results are displayed in Table 1.

Table 1: Model performance with respect to the target event, evaluated at the end of sentence.

Model	EV	SOM
SRN	0.985	0.986
A-SRN	0.899	0.949
A-SRN+	0.949	0.976

The SRN without any feedback mechanism performed very well. It mastered the task using its implicit mechanism by associating the scene information with the language at the hidden layer. A-SRN learned to generate the correct output hence demonstrating its ability to yield the correct interpretation of the event in the scene, mediated by the linguistic utterance. The accuracy of A-SRN is also very high for both parts of the

output representation, albeit slightly inferior to SRN. However, it does explicitly model the attentional mechanism which SRN does not. We examined the behavior of a trained A-SRN and found out that it might be the suboptimality of the attention mechanism that sometimes inhibits (via sigma-pi connection) the target objects at the input (and possibly also the components in the target event), hence reducing the output accuracy towards the end of sentence. To test this hypothesis, we introduced the third model, A-SRN+, as explained above, and its performance was observed to be indeed somewhat superior to A-SRN.

Table 2: Model performance with respect to the target event, evaluated at the end of sentence, with partially (50%) and completely removed scene information during training. Results show the performance on testing data with available scene information.

Model	EV-50	SOM-50	EV-0	SOM-0
SRN	0.995	0.989	0.504	0.627
A-SRN	0.989	0.988	0.769	0.823
A-SRN+	0.992	0.990	0.671	0.688

Restricting the situational input We restricted the availability of the visual input during training, either by randomly choosing 50% of sentences (in each training epoch), or completely. The purpose of this manipulation was twofold: to simulate the lack of visual input (for example, to simulate mere listening about the given event) but also to force the network to rely more on the linguistic pathway in predicting the output.

The simulation results shown in Table 2 reveal that partial turning-off situational inputs during training positively affects model accuracy, especially that of A-SRN. Interestingly, we also observe (not shown in the table) that A-SRN yields a better performance also on testing data patterns with corresponding situational inputs, compared to the training mode with 100% availability of scene information (Table 1). However, the complete removal of the situational input had a negative effect in both models, deteriorating the results on the test set with the scene information. Because of the top-down attentional mechanism in A-SRN, this model could handle this type of testing much better, possibly taking advantage of the initial output representation evoked by the (sole) linguistic input and fed back as the situational input that eventually contributed to higher accuracy at the end of sentence.

Anticipation of upcoming constituents We examined the predictive ability for all three models, which turned out to be quite similar. Output accuracy was examined with respect to various degrees of consistency: the target (the strictest condition), the world knowledge (output is not correct but possible), and the depicted scene (output is in the scene but should not be attended to).

Prediction of the patient can be assessed at two steps. At reading a subject, it is around 0.5 w.r.t. the target but grows over 0.8 w.r.t. both world knowledge and the depicted scene. Prediction of a patient while reading a verb grows to 0.65 w.r.t. target, to 0.95 w.r.t. the world knowledge and to 0.85 w.r.t. the depicted scene in all models.

Prediction at the level of agent/patient objects (in SOM) is slightly less accurate. Upon processing the first word, the accuracy of predicting both objects remains at $\sim$0.45 (with greater accuracy in agent prediction), and only grows to $\sim$0.6 when processing the verb. (However, at the end of sentence, the SOM output is very accurate, as already reported in Table 1).

For models with omitted object inputs, the prediction ability decreases because of the missing visual scene information. When no situation inputs are presented during training, the model cannot rely on this type of information, thus ignoring it also for the test set when the visual information is available. Additionally, prediction in the dataset without the visual input was not achieved by any model.

In sum, the presented simulations reveal that all three models achieve very high levels of accuracy with respect to meaning interpretation at the end of sentence, with small differences between them. In addition, all models demonstrate a certain level of anticipatory behavior, measured by predicting the representations of upcoming constituents before the sentence end. Only the A-SRN(+) models, however, have the explicit attentional mechanisms necessary to account for behavioral findings from the visual world experiments, and model performance is indeed largely consistent with the findings of Knoeferle and colleagues.

Discussion

We modeled the process of situated language processing as revealed by studies within the visual world paradigm. We introduced a novel recurrent neural network model with an explicit attentional mechanism (A-SRN), and we compared it with a SRN and another model (A-SRN+) to appreciate the role of the feedback in sentence comprehension task. All models can almost perfectly learn to generate at the end of sentence the representation that is interpreted as sentence meaning in the visual context. Having read the sentence, each network correctly selects the relevant scene event and its corresponding constituents (agent/patient). All networks also demonstrate some predictive behavior reflected by the ability to anticipate upcoming constituents, as mediated by the utterance. The SRN performs expectedly very well, but crucially we have shown that adding an explicit atten-

tional mechanism (in A-SRN) results in a minimal loss in performance. From the cognitive perspective, A-SRN's attentional mechanism helps the network focus on the relevant scene event, incorporates into the model the visual attention system on an abstract level, and reveals similar anticipatory shifts in visual attention that have been found using the visual world paradigm (Knoeferle et al., 2005; Knoeferle & Crocker, 2006). In addition, the availability of the attentional mechanism helps the A-SRN to perform better on testing data with and without the scene information when trained on input with 50% restricted scene information (reaching almost ceiling performance), compared to the training mode with complete availability of scene information.

A-SRN differs crucially from CIANET (Mayberry et al., 2009) that served as our motivation, in its potential to deal with complex visual scenes containing more than two events. Preliminary simulations reveal that in case of three concurrent events, the performance degrades only slightly. With respect to world complexity, we expect that the benefits of the A-SRN model (i.e. anticipation of objects in the scene) may in fact increase as the knowledge of the network scales up, that is, when there's a larger difference between what the network learns during training, and what is actually depicted when processing a given sentence.

We think that mechanistic understanding of attention is important in various cognitive tasks. Four processes are thought to be fundamental to attention: working memory, top-down sensitivity control, competitive selection, and automatic bottom-up filtering for salient stimuli (Knudsen, 2007). According to this view, the control of attention involves the first three processes operating in a recurrent loop. Of these, our proposal for an attentional mechanism can be viewed as introducing a top-down sensivity control that regulates the strength of different signals that compete to access to working memory. In A-SRN, these different signals are all physical objects in the scene, along with all events. Hence, the sensitivity control is postulated to operate at two levels: a more concrete level of objects and more abstract level of events (in terms of underlying semantic representations). In Knudsen (2007), the working memory employs space-specific bias signals that improve the localization and representation of stimuli.

These space-specific bias signals could implement the feedback mechanism in A-SRN in case of its improved version, in which the 'what' and 'where' visual processing streams are separated. Current models only have the 'what' part, whereas in the extension one output module would code object identity (in current models handled by SOM) and another module would code spatial location. This architectural extension would clearly increase the cognitive plausibility of the model, and naturally, also the complexity of the mapping to be learned.

Acknowledgments

This work was supported by the Slovak Grant Agency for Science, #1/0439/11 (I.F., J.Š.) and in part by Humboldt foundation (I.F.) and the Cluster of Excellence "Multi-model Computing and Interaction" (M.W.C.) funded by German Science Foundation (DFG).

References

Elman, J. (1990). Finding structure in time. *Cognitive Science*, *14*, 179-211.

Gold, K., & Scassellati, B. (2007). A robot that uses existing vocabulary to infer non-visual word meanings from observation. In *Proceedings of the 22nd conference on artificial intelligence (aaai-07).* Vancouver, Canada.

Huettig, F., Rommers, J., & Meyer, A. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica*.

Kamide, Y., Altmann, G., & Haywood, S. (2003). Prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, *49*, 133-156.

Knoeferle, P., & Crocker, M. (2006). The coordinated interplay of scene, utterance, and world knowledge: Evidence from eye-tracking. *Cognitive Science*, *30*, 481-529.

Knoeferle, P., Crocker, M., Scheepers, C., & Pickering, M. (2005). The inuence of the immediate visual context on incremental thematic role-assignment: Evidence from eye-movements in depicted events. *Cognition*, *95*, 95-127.

Knudsen, E. (2007). Fundamental components of attention. *Annual Review of Neuroscience*, *30*(1), 57-78.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*(9), 1464-1480.

Mayberry, M., Crocker, M., & Knoeferle, P. (2009). Learning to attend: A connectionist model of situated language comprehension. *Cognitive Science*, *33*, 449-496.

Rumelhart, D., Hinton, G., & Williams, R. (1986). Learning internal representations by error propagation. In *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1, p. 318-362). Cambridge, MA: MIT Press.

Spivey, M., Tanenhaus, M., Eberhard, K., & Sedivy, J. (2002). Eye-movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. *Cognitive Psychology*, *45*, 447-481.

Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*, 1632-1634.

Yu, C., Ballard, D., & Aslin, R. (2005). The role of embodied intention in early lexical acquisition. *Cognitive Science*, *29*, 961-1005.