A Connectionist Model of Anticipation in Visual Worlds

Marshall R. Mayberry, III, Matthew W. Crocker, and Pia Knoeferle

Department of Computational Linguistics,
Saarland University, Saarbrücken, Germany
{martym, crocker, knoferle}@coli.uni-sb.de

Abstract. Recent "visual worlds" studies, wherein researchers study language in context by monitoring eye-movements in a visual scene during sentence processing, have revealed much about the interaction of diverse information sources and the time course of their influence on comprehension. In this study, five experiments that trade off scene context with a variety of linguistic factors are modelled with a Simple Recurrent Network modified to integrate a scene representation with the standard incremental input of a sentence. The results show that the model captures the qualitative behavior observed during the experiments, while retaining the ability to develop the correct interpretation in the absence of visual input.

1 Introduction

There are two prevalent theories of language acquisition. One view emphasizes syntactic and semantic bootstrapping during language acquisition that enable children to learn abstract concepts from mappings between different kinds of information sources [1,2]. Another view emerges from connectionist literature and emphasizes the learning of linguistic structure from purely distributional properties of language usage [3,4]. While the perspectives are often taken to be diametrically opposed, both can be seen as crucially relying on correlations between words and their immediate context, be it the sentence as a whole or extra-linguistic input, such as a scene.

We combine insights from both distributional and bootstrapping accounts in modelling the on-line comprehension of utterances in both the absence and presence of a visual scene. This is an important achievement in at least two regards. First, it emphasizes the complementarity between distributional and bootstrapping approaches–discovering structure across linguistic and scene contexts [5]. Further, it is an important first step in linking situated models of on-line utterance comprehension more tightly to accounts of language acquisition, thus emphasizing the continuity of language processing.

We present results from two simulations on a Simple Recurrent Network (SRN; [3]). Modification of the network to integrate input from a scene together with the characteristic incremental processing of such networks allowed us to model people's ability to adaptively use the contextual information in order to more rapidly interpret and disambiguate a sentence. The model draws on recent studies that appeal to theories of language acquisition to account for the comprehension of scene-related utterances [6,7]. Recent research within the *visual worlds* paradigm, wherein participants' gazes to a scene while listening to an utterance are monitored, provides support for this view. Findings from this paradigm support an account of scene-related utterance comprehension in

R. Dale et al. (Eds.): IJCNLP 2005, LNAI 3651, pp. 849–861, 2005.

which the rapid coordinated interaction of information from the immediate scene, and linguistic knowledge plays a major role in incremental and anticipatory comprehension.

2 Simulation 1

In Simulation 1, we simultaneously model four experiments that show the rapid influence of diverse informational sources–linguistic and world knowledge as well as scene information – on utterance comprehension. All experiments were conducted in German, a language that allows both subject-verb-object (SVO) and object-verb-subject (OVS) sentence types. In the face of word order ambiguity, case marking indicates the subject or object grammatical function, except in the case of feminine and neuter noun phrases where the article does not distinguish the nominative and accusative cases.

2.1 Anticipation Depending on Stereotypicality

The first two experiments that we modeled examined how linguistic and world knowledge or stereotypicality enabled rapid thematic role assignment in unambiguous sentences, thus determining who-does-what-to-whom in a scene.

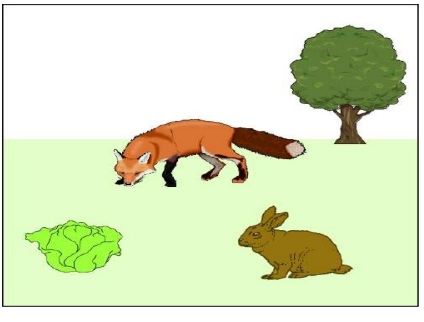

Fig. 1. Selectional Restrictions

Experiment 1: Morphosyntactic and lexical verb information. To examine the influence of case-marking and verb plausibility on thematic role assignment, [8] presented participants with utterances such as (1) or (2) that described a scene showing a hare, a cabbage, a fox, and a distractor (see Figure 1) :

(1) *Der Hase frisst gleich den Kohl.*
The hare$_{nom}$ eats shortly the cabbage$_{acc}$.
(2) *Den Hasen frisst gleich der Fuchs.*
The hare$_{acc}$ eats shortly the fox$_{nom}$.

After hearing "The hare$_{nom}$ eats ..." and "The hare$_{acc}$ eats ...", people made anticipatory eye-movements to the cabbage and fox respectively. This reveals that people were able to predict role fillers in a scene through linguistic/world knowledge that identified who-does-what-to-whom.

Experiment 2: Verb type information. To further investigate the role of verb information, the authors replaced the agent/patient verbs like *frisst* ("eats") with experiencer/theme verbs like *interessiert* ("interests"). This manipulation interchanged agent (experiencer) and patient (theme) roles from Experiment 1. For Figure 1 and the subject-first (3) or object-first sentence (4), participants showed gaze fixations complementary to those of Experiment 1, confirming that both case and semantic verb information are used to predict relevant role fillers.

(3) *Der Hase interessiert ganz besonders den Fuchs.*
The hare$_{nom}$ interests especially the fox$_{acc}$.

(4) *Den Hasen interessiert ganz besonders der Kohl.*
The hare$_{acc}$ interests especially the cabbage$_{nom}$.

2.2 Anticipation Depending on Depicted Events

The second set of experiments investigated whether depicted events showing who-does-what-to-whom can establish a scene character's role as agent or patient when syntactic and thematic role relations are temporarily ambiguous in the utterance.

Fig. 2. Depicted Events

Experiment 3: Verb-mediated depicted role relations. [9] presented such initially ambiguous spoken SVO (5) and OVS sentences (6) together with a scene in which a princess both paints a fencer and is washed by a pirate (Figure 2):

(5) *Die Princessin malt offensichtlich den Fechter.*
The princess$_{nom}$ paints obviously the fencer$_{acc}$.

(6) *Die Princessin wäscht offensichtlich der Pirat.*
The princess$_{acc}$ washes obviously the pirate$_{nom}$.

Linguistic disambiguation occurred on the second NP; disambiguation prior to the second NP was only possible through use of the depicted events. When the verb identified an action, the depicted role relations disambiguated towards either an agent-patient (5) or patient-agent (6) role relation, as indicated by anticipatory eye-movements to the patient (pirate) or agent (fencer) respectively for (5) and (6). This gaze-pattern showed the

rapid influence of verb-mediated depicted events on the assignment of thematic roles to a temporarily ambiguous sentence-initial noun phrase.

Experiment 4: Weak temporal adverb constraint. [9] also investigated German verb-final active (7) and passive (8) constructions. In this type of sentence, the initial subject noun phrase is role-ambiguous, and the auxiliary *wird* can have a passive or future interpretation.

(7) *Die Princessin wird sogleich den Pirat washen.*
The princess$_{nom}$ will right away wash the pirate$_{acc}$.
(8) *Die Princessin wird soeben von dem Fechter gemalt.*
The princess$_{acc}$ is just now painted by the fencer$_{nom}$.

To evoke early linguistic disambiguation, temporal adverbs biased the auxiliary *wird* toward either the future ("will") or passive ("is -ed") reading. Since the verb was sentence-final, the interplay of scene and linguistic cues (e.g., temporal adverbs) were rather more subtle. When the listener heard a future-biased adverb such as *sogleich*, after the auxiliary *wird*, he interpreted the initial NP as agent of a future active construction, as evidenced by anticipatory eye-movements to the patient in the scene. Conversely, listeners interpreted the passive-biased construction *soeben* with these roles exchanged.

2.3 Architecture

The Simple Recurrent Network is a type of neural network typically used to process temporal sequences of patterns such as words in a sentence. A common approach is for the modeller to train the network on prespecified targets, such as verbs and their arguments, that represent what the network is expected to produce upon completing a sentence. Processing is incremental, with each new input word interpreted in the context of the sentence processed so far, represented by a copy of the previous hidden layer serving as additional input or *context* to the current hidden layer. Because these types of associationist models automatically develop correlations among the data they are trained on, they will generally develop expectations about the output even before processing is completed because sufficient information occurs early in the sentence to warrant such predictions. Moreover, during the course of processing a sentence these expectations can be overridden with subsequent input, often abruptly revising an interpretation in a manner reminiscent of how humans seem to process language. Indeed, it is these characteristics of incremental processing, the automatic development of expectations, seamless integration of multiple sources of information, and nonmonotonic revision that have endeared neural network models to cognitive researchers.

In Simulation 1, the four experiments described above have been modelled simultaneously using a single network. The goal of modelling all experimental results by a single architecture required enhancements to the SRN, the development and presentation of the training data, as well as the training regime itself. We describe these next.

In two of the experiments, only three characters are depicted, the representation of which can be propagated directly to the network's hidden layer. In the other two experiments, the scene featured three characters involved in two events (e.g., **pirate-washes-princess** and **princess-paints-fencer**, as shown in Figure 3). The middle character was

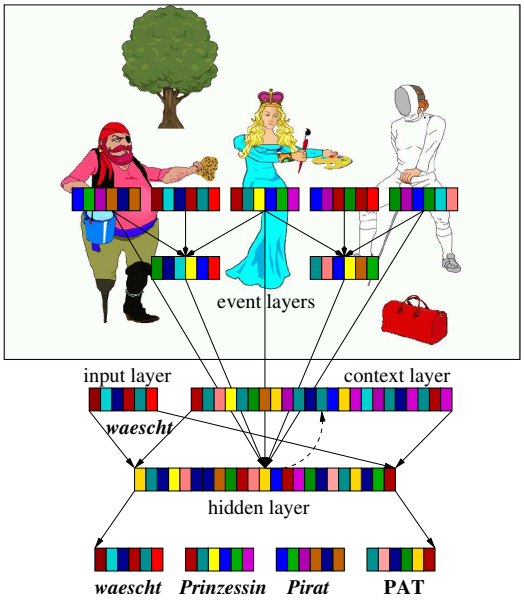

Fig. 3. Scene Integration

involved in both events, either as an agent or a patient (e.g., **princess**). Only one of the events, however, corresponded to the spoken linguistic input.

The representation of this scene information and its integration into the model's processing was the primary modification to the SRN. Connections between representations for the depicted characters and the hidden layer were provided. Encoding of the depicted events, when present, required additional links from the characters and depicted actions to **event** layers, and links from these event layers to the SRN's hidden layer. Representations for the events were developed in the event layers by compressing the scene representations of the involved characters and depicted actions through weights corresponding to the action, its agent and its patient for each event. This event representation was kept simple and only provided conceptual input to the hidden layer: who did what to whom was encoded for both events, when depicted; richer grammatical information (e.g., case and gender on articles) only came from the linguistic input.

Neural networks will usually encode any correlations in the data that help to minimize error. In order to prevent the network from encoding regularities in its weights regarding the position of the characters and events given in the scene (such as, for example, that the central character in the scene corresponds to the first NP in the presented sentence) which are not relevant to the role-assignment task, one set of weights was used for all characters, and another set of weights used for both events. This weight-sharing ensured that the network had to access the information encoded in the event layers, or determine the relevant characters itself, thus improving generalization. The representations for the characters and actions were the same for both input (scene and sentence) and output.

The input assemblies were the scene representations and the current word from the input sentence. The output assemblies were the verb, the first and second nouns, and an assembly that indicated whether the first noun was the agent or patient of the sentence (token **PAT** in Figure 3). Typically, agent and patient assemblies would be fixed in a case-role representation without such a discriminator, and the model required to learn to instantiate them correctly [10]. However, we found that the model performed much better when the task was recast as having to learn to isolate the nouns in the order in which they are introduced, and separately mark how those nouns relate to the verb. The input and output assemblies had 100 units each, the event layers contained 200 units each, and the hidden and context layers consisted of 400 units.

2.4 Input Data, Training, and Experiments

We trained the network to correctly handle sentences involving non-stereotypical events as well as stereotypical ones, both when visual context was present and when it was absent. As over half a billion sentence/scene combinations were possible for all of the experiments, we adopted a grammar-based approach to randomly generate sentences and scenes based on the materials from each experiment while holding out the actual materials to be used for testing. Because of the complementary roles that stereotypicality played in the two sets of experiments, there was virtually no lexical overlap between them. In order to accurately model the first two experiments involving selectional restrictions on verbs, two additional words were added to the lexicon for each character selected by a verb. For example, in the sentence *Der Hase frisst gleich den Kohl*, the nouns *Hase1*, *Hase2*, *Kohl1*, and *Kohl2* were used to develop training sentences. These were meant to represent, for example, words such as "rabbit" and "jackrabbit" or "carrot" and "lettuce" in the lexicon that have the same distributional properties as the orignal words "hare" and "cabbage". With these extra tokens the network could learn that *Hase*, *frisst*, and *Kohl* were correlated without ever encountering all three words in the same training sentence. The experiments involving non-stereotypicality did not pose this constraint, so training sentences were simply generated to avoid presenting experimental items.

Some standard simplifications to the words have been made to facilitate modelling. For example, multi-word adverbs such as *fast immer* were treated as one word through hyphenation so that sentence length within a given experimental set up is maintained. Nominal case markings such as *-n* in *Hasen* were removed to avoid sparse data as these markings are idiosyncratic, and the case markings on the determiners are more informative overall. More importantly, morphemes such as the infinitive marker *-en* and past participle *ge-* were removed, because, for example, the verb forms *malt*, *malen*, and *gemalt*, would all be treated as unrelated tokens, again contributing unnecessarily to the problem with sparse data. The result is that one verb form is used, and to perform accurately, the network must rely on its position in the sentence (either second or sentence-final), as well as whether the word *von* occurs to indicate a participial reading rather than infinitival. All 326 words in the lexicon for the first four experiments were given random representations.

We trained the network by repeatedly presenting the model with 1000 randomly generated sentences from each experiment (constituting one epoch) and testing every

100 epochs against the held-out test materials for each of the five experiments. Scenes were provided half of the time to provide an unbiased approximation to linguistic experience. The network was initialized with weights between -0.01 and 0.01. The learning rate was initially set to 0.05 and gradually reduced to 0.002 over the course of 15000 epochs. Four splits took a little less than two weeks to complete on 1.6Ghz PCs.

2.5 Results

Figure 4 reports the percentage of targets at the network's output layer that the model correctly matches, both as measured at the adverb and at the end of the sentence. The model clearly demonstrates the qualitative behavior observed in all four experiments in that it is able to access the information either from the encoded scene or stereotypicality and combine it with the incrementally presented sentence to anticipate forthcoming arguments.

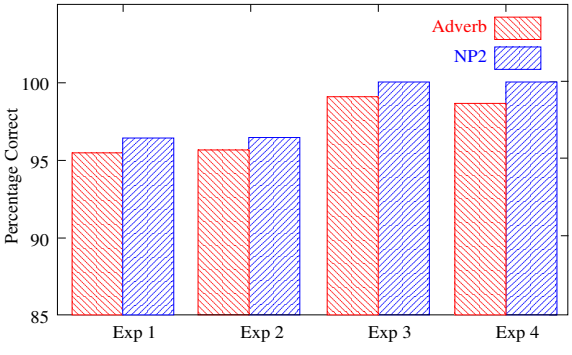

Fig. 4. Results

For the two studies using stereotypical information (experiments 1 and 2), the network achieved just over 96% at sentence end, and anticipation accuracy was just over 95% at the adverb. Because these sentences are unambiguous, the model is able to correctly identify the role of the upcoming argument, but makes errors in token identification, confusing words that are within the selectionally restricted set, such as, for example, *Kohl* and *Kohl2*. Thus, the model has not quite mastered the stereotypical knowledge, particularly as it relates to the presence of the scene.

In the other two experiments using non-stereotypical characters and depicted events (experiments 3 and 4), accuracy was 100% at the end of the sentence. More importantly, the model achieved over 98% early disambiguation on experiment 3, where the sentences were simple, active SVO and OVS. Early disambiguation on experiment 4 was somewhat harder because the adverb is the disambiguating point in the sentence as opposed to the verb in the other three experiments. As nonlinear dynamical systems, neural networks sometimes require an extra step to settle after a decision point is reached due to the attractor dynamics of the weights. For both experiments, most errors occurred on role-assignment due to the initially-ambiguous first noun phrase.

The difference in performance between the first two experiments and second two experiments can be attributed to the event layer that was only available in experiments 3 and 4. Closer inspection of the model's behavior during processing revealed that finer discrimination was encoded in the links between the event layers and hidden layer than that encoded in the weights between the characters and the hidden layer.

3 Simulation 2

The previous set of experiments demonstrated the rapid use of either linguistic knowledge or depicted events to anticipate forthcoming arguments in a sentence. A further important question is the relative importance of these two informational sources when they conflict. We first review an experimental study by [6] designed to address this issue and then report relevant modelling results.

Fig. 5. Scene vs Stored Knowledge

Scene vs Stored Knowledge. One goal of the study by [6] was to verify that stored knowledge about non-depicted events and information from depicted, but non-stereotypical, events each enable rapid thematic interpretation. Case-marking on the first NP always identified the pilot as a patient. After hearing the verb in (9) more inspections to the only food-serving agent (detective) than to the other agent showed the influence of depicted events. In contrast, when people heard the verb in condition two (10), a higher proportion of anticipatory eye-movements to the only stereotypical agent (wizard) than to the other agent revealed the influence of stereotypical knowledge (see Figure 5).

(9) *Den Piloten verköstigt gleich der Detektiv.*
The pilot$_{acc}$ serves-food-to shortly the detective$_{nom}$.

(10) *Den Piloten verzaubert gleich der Zauberer.*
The pilot$_{acc}$ jinxes shortly the wizard$_{nom}$.

Second, the study determined the *relative importance* of depicted events and verb-based thematic role knowledge when these information sources competed. In conditions three and four ((11) & (12)) participants heard an utterance in which the verb identified both a

stereotypical (detective) and a depicted agent (wizard). In this case, people preferred to rely on the immediate event depictions over stereotypical knowledge, and looked more often at the wizard, the agent of the depicted event, than at the other, stereotypical agent of the spying-action (the detective).

(11) *Den Piloten bespitzelt gleich der Zauberer.*
The pilot$_{acc}$ spies-on shortly the wizard$_{nom}$.
(12) *Den Piloten bespitzelt gleich der Detektiv.*
The pilot$_{acc}$ spies-on shortly the detective$_{nom}$.

3.1 Architecture, Data, Training, and Results

In simulation 1, we modelled experiments that depended on stereotypicality or depicted events, but not both. The experiment modelled in simulation 2, however, was specifically designed to investigate how these two information sources interacted. Accordingly, the network needed to learn to use either information from the scene or stereotypicality when available, and, moreover, favor the scene when the two sources conflicted, as observed in the empirical results. Recall that the network is trained only on the final interpretation of a sentence. Thus, capturing the observed behavior required manipulation of the frequencies of the four conditions described above during training. In order to train the network to develop stereotypical agents for verbs, the frequency that a verb occurs with its stereotypical agent, such as *Detektiv* and *bespitzelt* from example (12) above, had to be greater than for a non-stereotypical agent. However, the frequency should not be so great as to override the influence from the scene.

The solution we adopted is motivated by theories of language acquisition that take into account the importance of early linguistic experience in a visual environment (see the General Discussion). We found a small range of frequencies that permitted the network to develop an early reliance on the information from the scene while it gradually learned the stereotypical associations. Figure 6 shows the effect this training regime had over 6000 epochs on the ability of the network to accurately anticipate the missing argu-

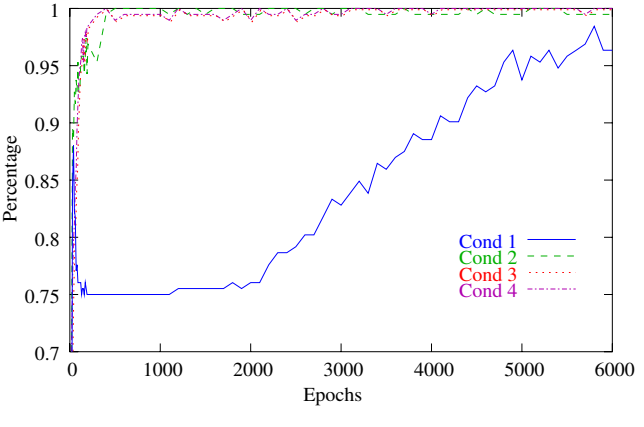

Fig. 6. Acquisition of Stereotypicality

ment in each of the four conditions described above when the ratio of non-stereotypical to stereotypical sentences was 8:1. The network quickly learns to use the scene for conditions 2-4 (examples 10-12), where the action in the linguistic input stream is also depicted, allowing the network to determine the relevant event and deduce the missing argument. (Because the graph shows the accuracy of the network at anticipating the upcoming argument at the adverb, the lines for conditions 3 and 4 are, in fact, identical.) But condition 1 (sentence 9) requires only stereotypical knowledge. The accuracy of condition 1 remains close to 75% (correctly producing the verb, first NP, and role discriminator, but not the second NP) until around epoch 1800 or so and then gradually improves as the network learns the appropriate stereotypical associations.

Results from several separate runs with different training parameters (such as learning rate and stereotypicality ratio) show that the network does indeed model the observed experimental behavior. The best results thus far exceed 99% accuracy in correctly anticipating the proper roles and 100% accuracy at the end of sentence.

As in simulation 1, the training corpus was generated by exhaustively combining participants and actions for all experimental conditions while holding out all test sentences. However, we found that we were able to use a larger learning rate, 0.1, than 0.05 as in the first simulation.

Analysis of the network after successful training suggests why this training policy works. Early in training, before stereotypicality has been encoded in the network's weights, patterns are developed in the hidden layer once the verb is read in from the input stream that enable the network to accurately decode that verb in the output layer. Not surprisingly, the network uses these same patterns to encode the stereotypical agent; the only constraint for the network is to ensure that the scene can still override this stereotypicality when the depicted event so dictates.

4 General Discussion and Future Work

The model demonstrates that reliance on correlations from distributional information in the linguistic input and the scene during training of the model enabled successful modelling of on-line utterance comprehension both in the presence and absence of rich visual contexts. The model that we present acquires stereotypical knowledge from distributional properties of language during training. The mapping from words to the scene representations is established through cooccurrence of scene-related utterances and depicted events during training. The network that emerges from this training regime successfully models five *visual worlds* eye-tracking experiments in two simulations. A first simulation of four experiments models the influence of either thematic and syntactic knowledge in the utterance [8], or of depicted events showing who-does-what-to-whom on incremental thematic role assignment [9]. Crucially in modelling the fifth experiment we are able to account for the greater relative priority of depicted events when event depictions and event knowledge conflict with each other.

The simple accuracy results belie the complexity of the task in both simulations. For experiments 3 and 4, the network has to demonstrate *anticipation* of upcoming roles when the scene is present, showing that it can indeed access the proper role and filler from the compressed representation of the event associated with the verb processed in

the linguistic stream when available. This task is rendered more difficult because the appropriate event must be extracted from the superimposition of the two events in the scene, which is what is propagated into the model's hidden layer. In addition, it must also still be able to process all sentences correctly when the scene is not present.

Simulation 2 is more challenging still. The experiment shows that information from the scene takes precedence when there is a conflict with stereotypical knowledge; otherwise, each source of knowledge is used when it is available. In the training regime used in this simulation, the dominance of the scene is established early because it is much more frequent than the more particular stereotypical knowledge. As training progresses, stereotypical knowledge is gradually learned because it is sufficiently frequent for the network to capture the relevant associations. As the network weights gradually saturate, it becomes more difficult to retune them. But encoding stereotypical knowledge requires far fewer weight adjustments, so the network is able to learn that task later during training.

According to the "Coordinated Interplay" account in [7,6,11], the rapid integration of scene and utterance information and the observed preferred reliance of the comprehension system on the visual context over stored knowledge might best be explained by appealing to bootstrapping accounts of language acquisition. The development of a child's world knowledge occurs in a visual environment, which accordingly plays a prominent role during language acquisition. The fact that the child can draw on two informational sources (utterance and scene) enables it to infer information that it has not yet acquired from what it already knows. Bootstrapping accounts for the fact that a child can correlate event structure from the world around it with descriptions of events. When a child perceives an event, the structural information it extracts from it can determine how the child interprets a sentence that describes the event in question. The incremental interpretation of a sentence can in turn direct the child's attention to relevant entities and events in the environment. Events are only present for a limited time when utterances refer to such events during child language acquisition. This time-limited presence might determine the tight coordination with which attention in the scene interacts with utterance comprehension and information extracted from the scene during adult language comprehension. This contextual development may have shaped both our cognitive architecture (i.e., providing for rapid, seamless integration of scene and linguistic information), and comprehension mechanisms (e.g., people rapidly avail themselves of information from the immediate scene when the utterance identifies it).

The model presented in this paper extends current models of on-line utterance comprehension when utterances relate to a scene [12] in several ways. Existing models account for processes of establishing reference in scene-sentence integration when scenes contain only objects. Our network accounts for processes of establishing reference, and furthermore models the rapid assignment of thematic roles based on linguistic and world knowledge, as well as scene events. In this way, it achieves rapid scene-utterance integration for increasingly rich visual contexts, including the construction of propositional representations on the basis of scene events. It models the integration of utterances and relatively rich scenes (that contain actions and events) in addition to objects. Furthermore, the model–in line with experimental findings – successfully accounts for the relative priority of depicted events in thematic interpretation. It importantly achieves

this through a modification of the training regime that prioritizes scene information. This confirms suggestions from [7] that a rapid interplay between utterance comprehension and the immediate scene context during acquisition is one potential cause for the relative priority of depicted events during on-line comprehension.

Connectionist models such as the SRN have been used to model aspects of cognitive development, including the time-course of emergent behaviors [13], making them highly suitable for simulating developmental stages in child language acquisition (e.g., first learning names of objects in the immediate scene, and later proceeding to the acquisition of stereotypical knowledge). The finding that modelling this aspect of development provides an efficient way to naturally reproduce the observed adult comprehension behavior promises to offer deeper insight into how adult performance is at least partially a consequence of the acquisition process.

Future research will focus on combining all of the experiments in one model, and expand the range of sentence types and fillers to which the network is exposed. The architecture itself is being redesigned to scale up to much more complex linguistic constructions and have greater coverage while retaining the cognitively plausible behavior described in this study [14].

Acknowledgements

The first two authors were supported by SFB 378 (project "ALPHA"), and the third author by a PhD studentship (GRK 715), all awarded by the German Research Foundation (DFG).

References

1. Steven Pinker. How could a child use verb syntax to learn verb semantics? In Lila Gleitman and Barbara Landau, editors, *The acquisition of the lexicon*, pages 377–410. MIT Press, Cambridge, MA, 1994.
2. Cynthia Fisher, D. G. Hall, S. Rakowitz, and Lila Gleitman. When it is better to receive than to give: Syntactic and conceptual constraints on vocabulary growth. In Lila Gleitman and Barbara Landau, editors, *The acquisition of the lexicon*, pages 333–375. MIT Press, Cambridge, MA, 1994.
3. Jeffrey L. Elman. Finding structure in time. *Cognitive Science*, 14:179–211, 1990.
4. Martin Redington, Nick Chater, and Steven Finch. Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, 22:425–469, 1998.
5. Deb Roy and Alex Pentland. Learning words from sights and sounds: A computational model. *Cognitive Science*, 26(1):113–146, 2002.
6. Pia Knoeferle and Matthew W. Crocker. Stored knowledge versus depicted events: what guides auditory sentence comprehension. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society*. Mahawah, NJ: Erlbaum, 2004. 714–719.
7. Pia Knoeferle and Matthew W. Crocker. The coordinated interplay of scene, utterance, and world knowledge: evidence from eye-tracking. submitted.
8. Yuki Kamide, Christoph Scheepers, and Gerry T. M. Altmann. Integration of syntactic and semantic information in predictive processing: Cross-linguistic evidence from German and English. *Journal of Psycholinguistic Research*, 32(1):37–55, 2003.

9. Pia Knoeferle, Matthew W. Crocker, Christoph Scheepers, and Martin J. Pickering. The influence of the immediate visual context on incremental thematic role-assignment: evidence from eye-movements in depicted events. *Cognition*, 95:95–127, 2005.
10. Risto Miikkulainen. Natural language processing with subsymbolic neural networks. In Antony Browne, editor, *Neural Network Perspectives on Cognition and Adaptive Robotics*, pages 120–139. Institute of Physics Publishing, Bristol, UK; Philadelphia, PA, 1997.
11. Pia Knoeferle and Matthew W. Crocker. The coordinated processing of scene and utterance: evidence from eye-tracking in depicted events. In *Proceedings of International Conference on Cognitive Science*, Allahabad, India, 2004.
12. Deb Roy and Niloy Mukherjee. Towards situated speech understanding: Visual context priming of language models. *Computer Speech and Language*, 19(2):227–248, 2005.
13. Jeffrey L. Elman, Elizabeth A. Bates, Mark H. Johnson, Annette Karmiloff-Smith, Domenico Parisi, and Kim Plunkett. *Rethinking Innateness: A Connectionist Perspective on Development*. MIT Press, Cambridge, MA, 1996.
14. Marshall R. Mayberry and Matthew W. Crocker. Generating semantic graphs through self-organization. In *Proceedings of the AAAI Symposium on Compositional Connectionism in Cognitive Science*, pages 40–49, Washington, D.C., 2004.