A Connectionist Model of Sentence Comprehension in Visual Worlds

Marshall R. Mayberry, III (martym@coli.uni-sb.de)

Matthew W. Crocker (crocker@coli.uni-sb.de)

Pia Knoeferle (knoeferle@coli.uni-sb.de)

Department of Computational Linguistics,
Saarland University, 66041
Saarbrücken, Germany

Abstract

People process utterances incrementally, often anticipating arguments in advance of actually hearing those arguments. The presence of a visual scene during spoken sentence comprehension has enabled researchers to study the circumstances under which this type of anticipatory behavior occurs and explore the influence of the scene itself. Such "visual world" studies have revealed much about the interaction of various information sources and the time course of their influence on comprehension. In the current study, five experiments that trade off scene context with a variety of linguistic factors are modelled with a Simple Recurrent Network that has been 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.

Introduction

It is widely accepted in the psycholinguistics community that sentence processing is incremental. There is growing evidence from recent eye-tracking experiments demonstrating that anticipation also plays an important role in how people interpret sentences. These observations naturally lead to questions about the factors underlying this behavior. That is, what kinds of information does the sentence processor rely on to construct intepretations incrementally and hypothesize expected continuations, and what is the time course of such processes? The richness of language and the context in which it occurs provide many relevant information sources, including morphosyntactic and lexical information, world knowledge, as well as information from the immediate visual context.

Language is naturally *situated*. People learn their native tongue within the context of the world around them, and they use language to make reference to objects in that world, as well as relationships among those objects (e.g., Gleitman, 1990). Given the relevance of the visual context, a natural question to ask is how language processing interacts with non-linguistic input such as a scene. Research within the *visual worlds* paradigm has already begun to offer insights into this question that were not possible with more traditional approaches such as reading studies that relied on processing load to track how people comprehend language. In the visual worlds paradigm, research has shown that people's attention to objects in a scene closely tracks their mention in a spoken sentence, and that the visual scene itself can influence the interpretation of linguistic input (Tanenhaus et al., 1995). Furthermore, referential contrast among objects in a scene can, for example, facilitate object identification in structurally unambiguous sentences with temporary referential ambiguity (Sedivy et al., 1999). World knowledge, too, may speed up object identification, allowing early interpretation of a sentence where selectional restrictions on the verb apply to only one depicted object (Altmann and Kamide, 1999). Such world knowledge may interact with linguistic knowledge such as case markings to influence expectations of post-verbal arguments and visual referents (Kamide et al., 2003). Recent studies by Knoeferle et al. (2005) have also shown that when scenes include depicted events, such visual information can establish important relations between the entities, including role relations. These findings show that the careful manipulation of information sources in an experimental setting can help shed considerable light on the underlying processes involved in language comprehension.

Despite this important research, models of sentence comprehension continue to focus on modelling reading behavior. To our knowledge, none attempt to model use of immediate (non-linguistic) context. In this paper we present results from two simulations using a Simple Recurrent Network (SRN; Elman, 1990) that has been modified to integrate input from a scene with the traditional word-by-word processing in order to demonstrate the *adaptive* use of the scene (when available) during comprehension. In addition to modelling the characteristic behaviors of incrementality and anticipation, we model people's ability to use the contextual information in visual scenes to more rapidly interpret and disambiguate a sentence. In the modelling of five visual worlds experiments reported in this paper, accurate sentence interpretation hinges on proper case-role assignment to sentence participants. In particular, modelling is focussed on the following aspects of sentence processing:

- anticipation of role and filler
- adaptive use of the visual scene
- influence of depicted events on interpretation
- processing without the scene
- multiple/conflicting information sources
- relative importance of the scene

Simulation 1

In the first simulation, we sought to simultaneously model four experiments that featured revealing contrasts. These four experiments show that the human sentence processor is extremely adept at utilizing all available sources of information to rapidly interpret language. In particular, information from visual context can readily be integrated with linguistic and world knowledge to disambiguate argument roles where the information from the auditory stream is insufficient in itself.

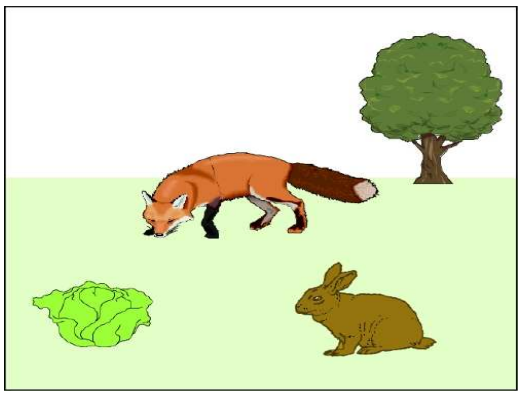

Figure 1: **Selectional Restrictions.**

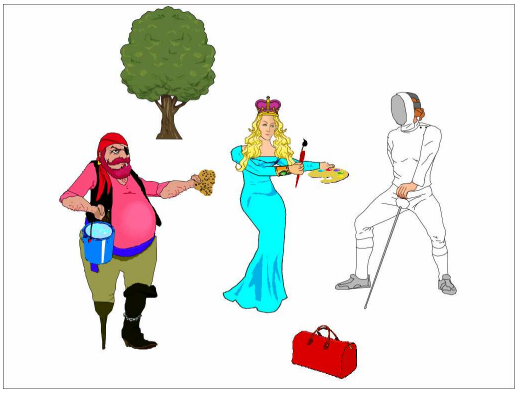

Figure 2: **Depicted Events.**

All experiments were conducted in German, a language that allows both subject-verb-object (SVO) and object-verb-subject (OVS) sentence types, so that word order cannot be reliably used to determine role assignments. Rather, case marking in German is used to indicate grammatical function such as subject or object, except in the case of feminine and neuter nouns where the article does not carry any distinguishing marking for the nominative and accusative cases.

Anticipation in unambiguous sentences

The first two experiments modelled involved unambiguous sentences in which case-marking and verb selectional restrictions in the linguistic input, together with depicted characters in a visual scene, allowed rapid assignment of the roles played by those characters.

Experiment 1: Morphosyntactic and lexical verb information. Kamide et al. (2003) presented subjects with a scene showing, for example, a hare, a cabbage, a fox, and a distractor (see Figure 1) together with either a spoken German SVO sentence (1) or with an OVS sentence (2):

(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$_{nom}$.

The subject and object case-marking on the article of the first noun phrase together with verb meaning and world knowledge allowed anticipation of the correct post-verbal referent. People made anticipatory eye-movements to the cabbage after hearing "The hare (subj) eats ..." and to the after "The hare (obj) eats ...". Thus, when the utterance is unambiguous, and linguistic/world knowledge restricts the domain of potential referents in a scene, the comprehension system may predict mention of post-verbal referents.

Experiment 2: Verb type information. To further investigate the role of verb information, the authors used the same visual scenes in a follow-up study, but replaced the agent/patient verbs like *frisst* ("eats") with experiencer/theme verbs like *interessiert* ("interests"). The agent/experiencer and patient/theme roles from Experiment 1 were swapped. Given the same scene in Figure 1 but the subject-first sentence (3) or object-first sentence (4), participants showed gaze fixations complementary to those in the first experiment, confirming that both syntactic case information and semantic verb information are used to predict subsequent referents.

(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$_{nom}$.

Anticipation in ambiguous sentences

The second set of experiments investigated temporarily ambiguous German sentences. Findings showed that depicted events–just like world and linguistic knowledge in unambiguous sentences–can establish a scene character's role as agent or patient in the face of linguistic structural ambiguity.

Experiment 3: Verb-mediated depicted role relations. Knoeferle et al. (2005) investigated comprehension of spoken sentences with local structural and thematic role ambiguity. An example of the German SVO/OVS ambiguity is the SVO sentence (5) versus the OVS sentence (6):

(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$_{nom}$.

Together with the auditorily presented sentence a scene was shown in which a princess both paints a fencer and is washed by a pirate (see Figure 2). *Linguistic* disambiguation occurred on the second noun phrase (NP); in the absence of stereotypical verb-argument relationships, disambiguation prior to the second NP was only possible through use of the depicted events and their associated depicted role relations. When the verb identified an action, the depicted role relations disambiguated towards either an SVO agent-patient (5) or OVS patient-agent role (6) 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 a thematic role to a temporarily ambiguous sentence-initial noun phrase.

Experiment 4: Soft temporal adverb constraint. Knoeferle et al. also investigated German verb-final active/passive constructions. In both the active future-tense (7) and the passive sentence (8), 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$_{acc}$ was 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 an agent of a future construction, as evidenced by anticipatory eye-movements to the patient in the scene. Conversely, listeners interpreted the passive-biased construction with these roles exchanged.

Architecture

Neural networks are a natural choice for modelling multiple modality eye-tracking experiments such as those just described because seamless integration of disparate information sources has long been a hallmark of these types of architectures, due to their operation through the soft constraints of computation over weights between numerous interconnected units employing simple summation and compression.

The Simple Recurrent Network is a type of neural network that allows the processing of temporal sequences of patterns like words in a sentence. The modeller trains 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 to the current hidden layer. Because these types of associationist models automatically develop correlations among the data they are trained on, they will typically develop expectations about the output even before processing is completed. 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 this study, the four experiments described above have been modelled simultaneously using the same 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. These will be described in turn below.

In two of the experiments, only three characters are depicted, 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 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 most important modification to the SRN involved the representation of this scene information and integration into the model's processing (see Figure 3). The encoding of the scene was complicated by the need to represent the depicted events, which involved actions, in addition to just the characters themselves in the scene. Accordingly, the model has links from the characters to the hidden layer, links from the characters and depicted actions to **event** layers, and links from these event layers to the hidden layer of the SRN. 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, the agent of the action, and the patient of the action. This event representation was kept simple to provide conceptual input to the hidden layer, divorced from the linguistic information the network was to learn from the sentence input. That is, who did what to whom was encoded for the events, when depicted; grammatical information came from the linguistic input.

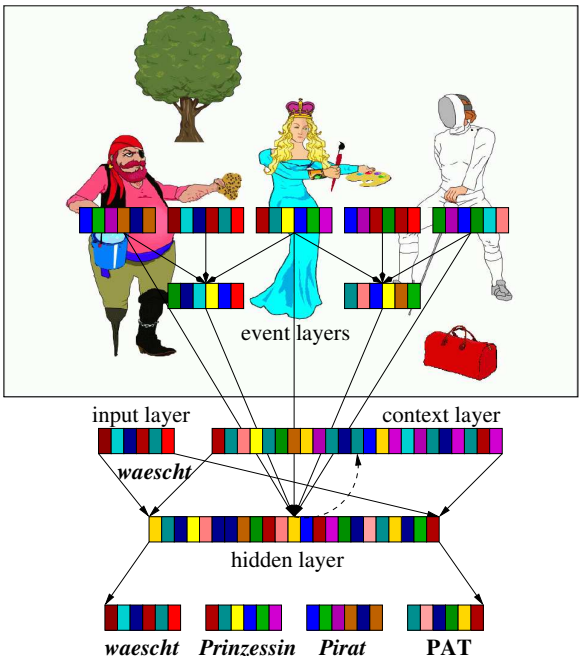

Figure 3: **Scene Integration.**

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 SRN consisted of input and output assemblies of 100 units each. 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. 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 (Miikkulainen, 1997), but 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 hidden and context layers consisted of 400 units. The network was initialized with weights between -0.01 and 0.01.

Input Data, Training, and Experiments

The network had to be trained to correctly handle sentences involving non-stereotypical events as well as stereotypical ones, both when visual context was present and absent. To this end, we adopted a grammar-based approach to exhaustively generate a set of sentences based on the experimental materials while holding out the actual materials to be used for testing. 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*[1] were used to develop training sentences so that 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 generated simply 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* are treated as one word through hyphenation so that sentence length within a given experimental set up is maintained. Nominal case markings such as the final *-n* in *Hasen* are removed to avoid an unnecessary sparse data problem 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-* are treated as separate words, again to avoid the sparse data problem, where, for example, the verb forms *malt*, *malen*, and *gemalt*, would all be treated as unrelated tokens. All 326 words in the lexicon for the first four experiments were given random representations.

Training the network involved repeatedly presenting the model with 1000 randomly generated sentences from each experiment (constituting one epoch) and testing every 10 epochs against the held-out test materials for each of the five experiments. Half the sentences were trained with scenes and half were trained without in order to approximate linguistic experience. The learning rate was initially set to 0.05 and gradually reduced to 0.002 over the course of 10000 epochs. Ten splits were run on 1.6Ghz PCs and took approximately two weeks to complete.

Results

Figure 4 reports the percentage of targets at the network's output layer 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 scene information and combine it with the incrementally presented sentence to anticipate forthcoming arguments.

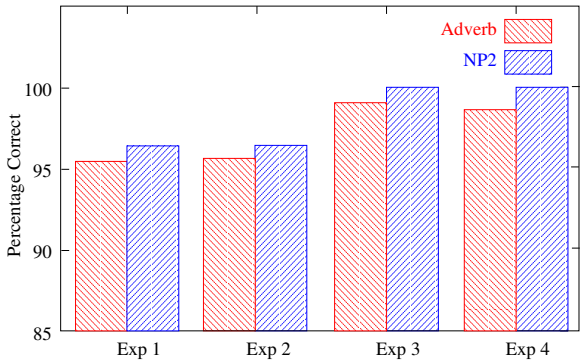

Figure 4: **Results.**

For the two experiments (1 and 2) using stereotypical information, the network achieved just over 96% at sentence end, and anticipation accuracy was just over 95% at the adverb. Analysis shows that the network 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.

For 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.

On closer inspection of the model's behavior during processing, it is apparent that the event layers provide enough additional information beyond that encoded in the weights between the characters and the hidden layer that the model is able to make finer discriminations in experiments 3 and 4, enhancing its performance.

Simulation 2

In this section, we present preliminary results from modelling a more challenging experiment that examines how non-stereotypical depicted events interact with stereotypicality in the form of (internal) world knowledge to influence incremental thematic interpretation.

Coordinated Interplay Account. Based on the findings from the four experiments in Simulation 1, Knoeferle and Crocker (2004b) presented a study that examined two issues. First, it replicated the finding that stored knowledge about events that were not depicted and information from depicted, but non-stereotypical, events each enable rapid thematic interpretation. An example scene showed a wizard, a pilot, and a detective serving food (see Figure 5). Item sentences had an object-verb-subject order. When people heard (9), case-marking on the first NP identified the pilot as a patient. The subsequent verb uniquely identified the detective as the only food-serving agent, revealed by more inspections to the agent

[1]These are meant to represent, for example, words such as "carrot" and "lettuce" in the lexicon that have the same distributional properties as "cabbage".

Figure 5: **Coordinated Interplay Account.**

of the depicted event (detective) than to the other agent. In contrast, when people heard the verb in sentence (10), stereotypical knowledge about agents identified the wizard as the only relevant agent, as indicated by a higher proportion of anticipatory eye-movements to the stereotypical agent (wizard) than to the other agent.

(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. Participants heard utterances (11 & 12) where the verb identified both a stereotypical (detective) or a depicted agent (wizard). When faced with this conflict, people preferred to rely on the immediate event depictions over stereotypical knowledge, and looked more often at the wizard, the agent in the depicted event, than at the other, stereotypical agent of a 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$_{nom}$.

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 that it overrode the influence from the scene.

The solution adopted in this study is motivated by a theory of language acquisition that takes 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. The training corpus was generated by exhaustively combining participants and actions while holding out all test sentences. The network was trained both with and without scenes.

The experiment was modelled with the same basic architecture as described in Simulation 1 with some minor modifications: only the event-to-hidden layer links were used for scene integration and weight updates were roughly five times stronger. Neither change is expected to adversely affect the previous results when experiments from simulation 1 are integrated with simulation 2 in future work.

Early results from five separate runs with slightly different training parameters (e.g., 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.

General Discussion and Future Work

Experiments in the visual worlds paradigm have clearly reinforced the view of language comprehension as an active, incremental, highly integrative process in which anticipation of upcoming arguments plays a crucial role. Visual context not only facilitates identification of likely referents in a sentence, but helps establish relationships between referents and the roles they may fill. Research thus far has shown that the human sentence processor seems to have facile access to whatever relevant information is available, whether it be syntactic, lexical, semantic, or visual, and that it can combine these sources to achieve as complete an interpretation as is possible at any given point in comprehending a sentence.

The modelling results reported in this paper are an important step toward the goal of understanding how the human sentence processor is able to accomplish these feats. The SRN provides a natural framework for this research because its operation is premised on incremental and integrative processing. Trained simply to produce a representation of the complete interpretation of a sentence as each new word is processed (on the view that people learn to process language by reviewing what they hear), the model automatically develops anticipations for upcoming arguments that allow it to demonstrate the early disambiguation behavior observed in the visual worlds experiments modelled here.

The simple accuracy results belie the complexity of the task in both simulations. In Simulation 1, the network has to demonstrate early disambiguation 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 first NP and verb processed in the linguistic stream. This task is rendered more difficult because the proper 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 scene information 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 immediate than 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.

Knoeferle and Crocker (2004a,b) suggest that the preferred reliance of the comprehension system on the visual context over stored knowledge can broadly be accomodated by appealing to boot-strapping accounts of language acquisition such as that of Gleitman (1990). Gleitman suggests that partial interpretation of a sentence can direct the child's attention to relevant entities and events in the environment. A child can in turn extract event structure from the world around it. Indeed, the rapid impact of the immediate situation identifies the comprehension system to be highly adapted towards acquiring new information from its environment. The fact that the child can draw on two informational sources (sentence and scene) enables it to infer information that it has not yet acquired from what it already knows.

Knoeferle and Crocker (2004a) argue that the substantial part of our lives we have spent acquiring language 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).

Connectionist models such as the SRN have been used to model aspects of cognitive development, including the timing of emergent behaviors (Elman et al., 1996), 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). If there are developmental reasons for the preferred reliance of listeners on the immediate scene during language comprehension, then the finding that modelling that development provides the most efficient (if not only) way to naturally reproduce the observed experimental behavior promises to offer deeper insight into how such knowledge is instilled in the brain.

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 (Mayberry and Crocker, 2004).

Conclusion

We have presented a neural network architecture that successfully models the results of five recent experiments designed to study the interaction of visual context with sentence processing. The model shows that it can adaptively use information from the visual scene such as depicted events, when present, to anticipate roles and fillers as observed in each of the experiments, as well as demonstrate traditional incremental processing behavior when context is absent. Furthermore, more recent results show that training the network in a visual environment, with stereotypical knowledge gradually learned and reinforced, allows the model to successfully negotiate even conflicting information sources.

Acknowledgements

This research was supported by SFB 378 project "ALPHA" to the first two authors and a PhD scholarship to the last, all funded by the German Research Foundation (DFG).

References

Altmann, G. T. M. and Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73:247–264.

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

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., and Plunkett, K. (1996). *Rethinking Innateness: A Connectionist Perspective on Development*. MIT Press, Cambridge, MA.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, 1:3–55.

Kamide, Y., Scheepers, C., and Altmann, G. T. M. (2003). Integration of syntactic and semantic information in predictive processing: Cross-linguistic evidence from German and English. *Journal of Psycholinguistic Research*, 32(1):37–55.

Knoeferle, P. and Crocker, M. W. (2004a). The coordinated processing of scene and utterance: evidence from eye-tracking in depicted events. In *Proceedings of International Conference on Cognitive Science*, Allahabad, India.

Knoeferle, P. and Crocker, M. W. (2004b). 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. 714–719.

Knoeferle, P., Crocker, M. W., Scheepers, C., and Pickering, M. J. (2005). The influence of the immediate visual context on incremental thematic role-assignment: evidence from eye-movements in depicted events. *Cognition*, 95:95–127.

Mayberry, M. R. and Crocker, M. W. (2004). Generating semantic graphs through self-organization. In *Proceedings of the AAAI Symposium on Compositional Connectionism in Cognitive Science*, pages 40–49, Washington, D.C.

Miikkulainen, R. (1997). Natural language processing with subsymbolic neural networks. In Browne, A., editor, *Neural Network Perspectives on Cognition and Adaptive Robotics*, pages 120–139. Institute of Physics Publishing, Bristol, UK; Philadelphia, PA.

Sedivy, J., Tanenhaus, M., Chambers, C., and Carlson, G. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, 71:109–148.

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