A Connectionist Model of Language-Scene Interaction

Marshall R. Mayberry, III Matthew W. Crocker Pia Knoeferle
Department of Computational Linguistics
Saarland University
Saarbrücken 66041, 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

People learn language within the context of the surrounding world, and use it to refer to objects in that world, as well as relationships among those objects (e.g., Gleitman, 1990). Recent research in the *visual worlds* paradigm, wherein participants' gazes in a scene while listening to an utterance are monitored, has yielded a number of insights into the time course of sentence comprehension. The careful manipulation of information sources in this experimental setting has begun to reveal important characteristics of comprehension such as incrementality and anticipation. For example, people's attention to objects in a scene closely tracks their mention in a spoken sentence (Tanenhaus et al., 1995), and world and linguistic knowledge seem to be factors that facilitate object identification (Altmann and Kamide, 1999; Kamide et al., 2003). More recently, Knoeferle et al. (2005) have shown that when scenes include depicted events, such visual information helps to establish important relations between the entities, such as role relations.

Models of sentence comprehension to date, however, continue to focus on modelling reading behavior. No model, to our knowledge, attempts to account for the use of immediate (non-linguistic) context. In this paper we present results from two simulations using a Simple Recurrent Network (SRN; Elman, 1990) modified to integrate input from a scene with the characteristic incremental processing of such networks in order to model people's ability to adaptively use the contextual information in visual scenes to more rapidly interpret and disambiguate a sentence. In the modelling of five visual worlds experiments reported here, accurate sentence interpretation hinges on proper case-role assignment to sentence referents. In particular, modelling is focussed on the following aspects of sentence processing:

- anticipation of upcoming arguments and their roles in a sentence
- adaptive use of the visual scene as context for a spoken utterance
- influence of depicted events on developing interpretation
- multiple/conflicting information sources and their relative importance

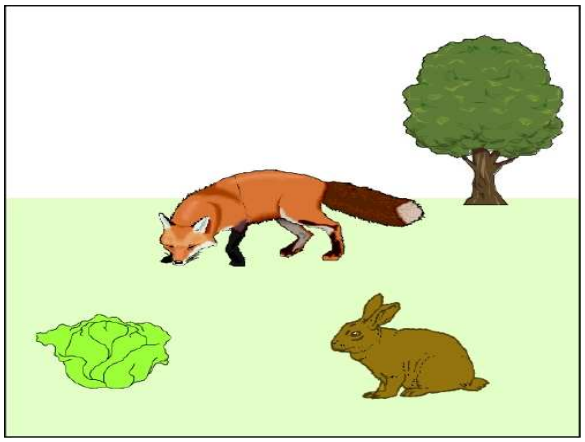

Figure 1: **Selectional Restrictions.** Gaze fixations depend on whether the hare is the subject or object of the sentence, as well as the thematic role structure of the verb. These gaze fixations reveal that people use linguistic and world knowledge to anticipate upcoming arguments.

2 Simulation 1

In the first simulation, we simultaneously model four experiments that featured revealing contrasts between world knowledge and context. These four experiments show that the human sentence processor is very 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.

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 alone cannot be relied upon 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.

2.1 Anticipation depending on stereotypicality

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

Experiment 1: Morphosyntactic and lexical verb information. In order to examine the influence of linguistic knowledge of case-marking, Kamide et al. (2003) presented experiment participants 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 (NP) together with verb meaning and world knowledge allowed anticipation of the correct post-verbal referent. Participants made anticipatory eye-movements to the cabbage after hearing "The hare$_{nom}$ eats ..." and to the fox upon encountering "The hare$_{acc}$ eats ...". Thus, people are able to predict upcoming referents when the utterance is unambiguous and linguistic/world knowledge restricts the domain of potential referents in a scene.

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 interchanged. 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$_{nom}$.

2.2 Anticipation depending on depicted events

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.

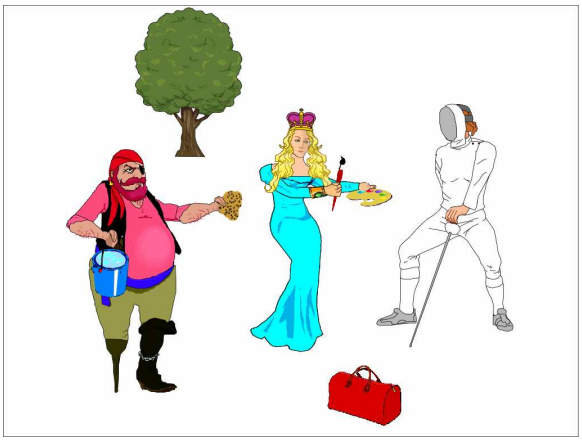

Figure 2: **Depicted Events.** The depiction of actions allows role information to be extracted from the scene. People can use this information to anticipate upcoming arguments even in the face of ambiguous linguistic input.

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 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: Weak 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$_{acc}$ 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 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.

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 to the current hidden layer. Because these types of associationist models automatically develop correlations among the sentence constituents 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 this study, the four experiments described

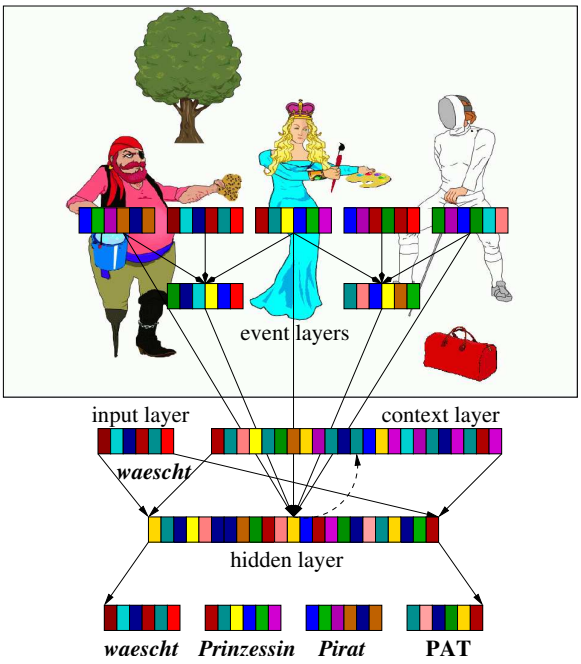

Figure 3: **Scene Integration.** A simple conceptual representation of the information in a scene, along with compressed event information from depicted actions when present, is fed into a standard SRN to model adaptive processing. The links connecting the depicted characters to the hidden layer are shared, as are the links connecting the event layers to the hidden layer.

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. 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 representation of this scene information and its integration into the model's processing was the main 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. The network developed representations for the events 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, but grammatical information only came from the linguistic input. As the focus of this study was on whether sentence processing could adapt to information from the scene when present or from stored knowledge, lower-level perceptual processes such as attention were not modelled.

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 (Miikkulainen, 1997). 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 exhaustively generate sentences and scenes 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* 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 original 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, while 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 over the vertices of a 100-dimensional hypercube, which resulted in marked improvement over sampling from within the hypercube (Noelle et al., 1997).

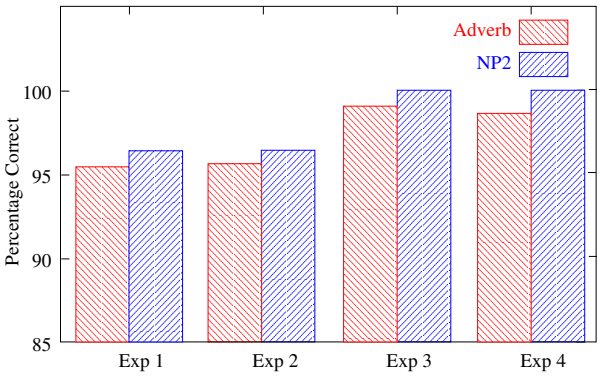

Figure 4: **Results.** In each of the four experiments modelled, anticipation of the upcoming argument at the adverb is nearly as accurate as at sentence end. However, the network has some difficulty with distinguishing stereotypical arguments.

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 four 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. Ten splits were run on 1.6Ghz PCs and took a little over two weeks to complete.

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 from the encoded scene or stereotypicality and combine it with the incrementally presented sentence to anticipate forthcoming arguments.

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.

3 Simulation 2

The previous set of experiments examined how people are able to use either stereotypical knowledge or depicted information to anticipate forthcoming arguments in a sentence. But how does the human sentence processor handle these information sources when both are present? Which takes precedence when they conflict? The experiment modelled in this section was designed to provide some insight into these questions.

Scene vs Stored Knowledge. Based on the findings from the four experiments in Simulation 1, Knoeferle and Crocker (2004b) examined two issues. First, it verified that stored knowledge about non-depicted events 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 (Figure 5). When people heard condition 1 (example sentence 9), the case-marking on the first NP identified the pilot as a patient. Stereotypical knowledge identified the wizard as the only relevant agent, as

Figure 5: **Scene vs Stored Knowledge.** Experimental results show that people rely on depicted information over stereotypical knowledge when both are present during sentence processing.

indicated by a higher proportion of anticipatory eye-movements to the stereotypical agent (wizard) than to the detective. In contrast, when people heard the verb in condition 2 (sentence 10), it uniquely identified the detective as the only food-serving agent, revealed by more inspections to the agent of the depicted event (detective) than to the wizard.

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

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

Second, the study determined the *relative importance* of depicted events and verb-based thematic role knowledge when the information sources were in competition. In both conditions 3 & 4 (sentences 11 & 12), participants heard an utterance in which the verb identified both a stereotypical (detective) and 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 the spying-action (the detective).

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

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

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 (11) 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 we adopted 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 ratios of stereotypicality to non-stereotypicality that permitted the network to develop an early reliance on information from the scene while it gradually learned the stereotypical associations. When the ratio was lower than 6:1, the network developed too strong a reliance on stereotypicality, overriding information from the scene. When the ratio was greater than 15:1, the scene always took precedence when it was present, but stereotypical knowledge was used when the scene was not present. Within this range, however, the network quickly learns to extract information from the scene because the scene representation remains static while a sentence is processed incrementally. It is the stereotypical associations, predictably, that take longer for the network to learn in rough proportion to their ratio over non-stereotypical agents.

Figure 6 shows the effect this training regime had over 6000 epochs on the ability of the network to accurately anticipate the missing argument 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 conditions 3 and 4 are the same up to the second NP, their curves

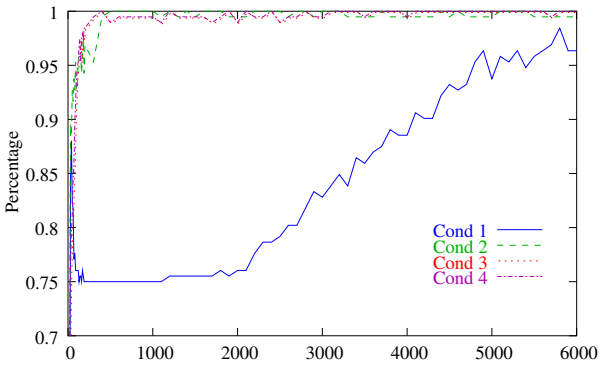

Figure 6: **Acquisition of Stereotypicality.** Stereotypical knowledge (condition 1) is acquired much more gradually than information from the scene (conditions 2-4).

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 1200 or so and then gradually improves as the network learns the appropriate stereotypical associations. The condition 1 curve asymptotically approaches 100% over the course of 10,000 epochs.

Results from several 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 so far exceed 99% accuracy in correctly anticipating the proper roles and 100% accuracy at sentence end.

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 the 0.05 used in the first simulation. The 130 words in the lexicon were given random binary representations from the vertices of a 100-dimensional hypercube as described before.

Analysis of the network after successful training suggests why the training regime of holding the ratio of stereotypical to non-stereotypical sentences constant works. Early in training, before stereotypicality has been encoded in the network's weights, patterns are developed in the hidden layer as each word is processed that enable the network to accurately decode the words in the output layer. Once the verb is read in, its hidden layer pattern is available to pro-

duce the correct output representations for both the verb itself and its stereotypical agent. Not surprisingly, the network thus learns to associate the hidden layer pattern for the verb with its stereotypical agent pattern in the second NP output slot. 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

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 easy access to whatever 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 difficult 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.

Knoeferle and Crocker (2004a,b) suggest that the preferred reliance of the comprehension system on the visual context over stored knowledge might best be explained by appealing to a boot-strapping account of language acquisition such as that of Gleitman (1990). 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. 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).

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 com-

prehension, 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).

5 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 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 negotiate even conflicting information sources.

6 Acknowledgements

This research was funded by SFB 378 project "ALPHA" to the first two authors and a PhD scholarship to the last, all awarded 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.

Noelle, D. C., Cottrell, G. W., and Wilms, F. (1997). Extreme attraction: The benefits of corner attractors. Technical Report CS97-536, Department of Computer Science and Engineering, UCSD, San Diego, CA.

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.