# A Connectionist Model of the Coordinated Interplay of Scene, Utterance, and World Knowledge

**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**

The interaction of utterance comprehension and information from a visual scene is characterized by the closely time-locked coordination of incremental comprehension and attention in the scene. Comprehension is also anticipatory, as revealed by attention to objects in a scene before they are mentioned. The interaction is further marked by the rapid and seamless integration of, and adaptation to, diverse information sources in both the utterance and visual scene. These sources can interact dynamically, both complementarily and, at times, conflictingly. A recurrent sigma-pi neural network is presented that implements an attentional mechanism to model these behaviors, directly instantiating the *coordinated interplay account* that suggests the utterance guides attention in the scene, which in turn rapidly provides information that influences comprehension. A key aspect of the account is that the immediacy of depicted events in the scene takes precedence over stereotypical knowledge when these two information sources conflict. Crucially, the model captures this behavior without being explicitly trained to resolve the conflict, even when the relative frequency of the information sources differs greatly.

**Keywords:** Connectionist modelling; situated utterance comprehension; language-scene interaction; attention

## Introduction

All human communication occurs in context. Indeed, even the so-called isolated phrase, coveted by linguists for its self-contained syntactic and semantic properties, is understood only within the context of human experience. In this way, the study of how language relates to its context provides insight into the very nature of language itself: how it *means* anything at all. Understanding the interaction of language and context, such as a visual environment, serves to identify and delineate the cognitive mechanisms involved in language comprehension, and how resources such as linguistic and world knowledge, as well as information from the visual context, are utilized. This challenge is especially daunting because language is inherently dynamic, and the utilization of these various information sources must be coordinated in real time.

Fortunately, a growing body of psycholinguistic research in the *visual worlds* experimental paradigm, wherein subjects' eye movements over a visual scene are monitored as they listen to an utterance, has begun to yield tangible data on the nature of the on-line interaction of utterance comprehension and context. Typically, that context is a visual scene that can establish referents and relations, together with the participants' own linguistic and world knowledge. The analysis of eye movements in a scene during utterance comprehension under the controlled manipulation of a variety of information sources has revealed five fundamental characteristics of on-line situated utterance comprehension. First, on-line comprehension occurs *incrementally* and is closely time-locked with attention to the scene (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). Second, attention to objects in a scene before they are mentioned in an utterance shows that *anticipation* plays a vital role in comprehension (Altmann & Kamide, 1999). Third, all available information sources–linguistic and world knowledge, as well as scene information–are rapidly and seamlessly *integrated* during on-line comprehension (Knoeferle, Crocker, Scheepers, & Pickering, 2005; Kamide, Scheepers, & Altmann, 2003; Sedivy, Tanenhaus, Chambers, & Carlson, 1999; Tanenhaus et al., 1995). Fourth, sentence comprehension is highly *adaptive* to the dynamic availability of information from these multiple sources. Fifth, these sources of information are *coordinated*: the interaction between language and visual scene processing is a two-way street. Comprehension of the unfolding utterance both rapidly guides attention to objects in the scene and, in turn, the attended region of the scene tightly constrains and influences comprehension, a process Knoeferle and Crocker (in press) dub the *coordinated interplay account* (CIA). Furthermore, a full account of this interaction must address the issue of what happens when information sources conflict: which sources take precedence and why? Recent research on the interaction between world knowledge and information from a visual scene indicate that immediate depicted events are preferred over knowledge about stereotypical relationships when these conflict. Knoeferle and Crocker suggest that such a preference may have its basis in the role the immediate visual environment plays in child-directed speech during language acquisition (e.g., Snow, 1977).

These characteristics of situated utterance comprehension pose an interesting challenge for modellers. The successful model should operate incrementally, anticipate upcoming referents, rapidly and seamlessly integrate information from multiple sources, adapt to available information, exhibit the observed attentional shift during utterance comprehension, and demonstrate the observed preference for the depicted information over world knowledge when these information sources conflict.

Two recently proposed models feature several of these characteristics. The Fuse model by Roy and Mukherjee (2005) uses an attentional mechanism to constrain the number of referents to improve speech recognition. The system does predict different ways a person might describe objects in a scene and biases how the words are recognized. The scene employed contains only objects, and is always assumed

Figure 1: **Coordinated Interplay Account** When presented with a sentence such as *Den Piloten bespitzelt gleich der ...* ("The pilot$_{acc}$ spies-on shortly the ..."), participants could either look at the *Detektiv* ("detective") as the most likely upcoming agent based on its stereotypical association with the verb *bespitzelt* ("spies on"), or at the *Zauberer* ("wizard"), depicted as doing the spying. Empirical results show that people prefer the depicted event over stereotypical knowledge.

to be relevant to the speech signal being processed. On the other hand, the model proposed by Mayberry, Crocker, and Knoeferle (2005) operates both with and without a scene, and the scene can contain both objects and actions that explicitly depict relationships between the objects. It processes sentences incrementally and is able to use the information about objects and events to predict upcoming arguments. The network modelled results from five distinct experiments in two separate simulations. In one of these simulations, the model also demonstrated the observed preference for the scene over stereotypical knowledge, but only after being explicitly trained to perform that resolution. However, the model did not feature an attentional mechanism.

In this study, a novel system called CIANet is presented that improves upon the model in Mayberry et al. (2005) in four important ways:

- it exhibits the proper *cognitive properties* of incrementality, anticipation, integration, adaptation, and coordination,
- it models the *empirically* observed preference for depicted information over stereotypical knowledge,
- it employs an *innovative* attentional mechanism that gives rise to this cognitively plausible behavior,
- it implements a *simpler* account of language-scene interaction, resulting in faster training and better performance.

These characteristics allow the model to more directly implement the CIA, described next.

## Coordinated Interplay Account

Knoeferle and Crocker (in press) 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 spying on a pilot, to whom a detective is also serving food (see Figure 1). The item sentences were in German, a language that allows both subject-verb-object (SOV) and object-verb-subject (OVS) word order, with grammatical function often indicated by case marking on the articles. For this experiment, item sentences had an OVS order. When people heard (Cond 1), 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, as revealed by more inspections to the agent of the depicted event (detective) than to the other agent. In contrast, when people heard the verb in sentence (Cond 2), stereotypical knowledge about jinxing 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.

(Cond 1) *Den Piloten verköstigt gleich der Detektiv.*
The pilot$_{acc}$ serves shortly the detective$_{nom}$.

(Cond 2) *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 (Cond 3 & 4) where the verb identified both a depicted (wizard) or a stereotypical agent (detective). When faced with this conflict, people preferentially relied upon the immediate event depiction over stereotypical knowledge, looking more often at the wizard, the agent in the depicted event, than at the other, stereotypical agent of the spying action (the detective).

(Cond 3) *Den Piloten bespitzelt gleich der Zauberer.*
The pilot$_{acc}$ spies-on shortly the wizard$_{nom}$.

(Cond 4) *Den Piloten bespitzelt gleich der Detektiv.*
The pilot$_{acc}$ spies-on shortly the detective$_{nom}$.

Combining insights from this study and prior psycholinguistic research, Knoeferle and Crocker (in press) propose the coordinated interplay account (CIA) of situated utterance comprehension. The CIA stipulates that initially the unfolding utterance guides attention in the visual scene to establish reference to objects and events. Once identified, the attended information rapidly constrains comprehension of the utterance, allowing anticipation of upcoming arguments not yet mentioned. Moreover, the immediacy of depicted events takes priority over learned world knowledge such as stereotypical associations.

## Modelling Dynamic Event Selection

Neural networks are a type of computational model that operates through parallel computation over massively interconnected simple processing units. These units take an input pattern and integrate it with activation from other units to produce an output pattern. Because their operation involves summation and compression over often thousands of weights, these connectionist systems are able to seamlessly integrate disparate information sources, making them a natural choice for modelling aspects of multimodal human information processing, such as the interaction of language and scene in the eye-tracking experiment just described.

CIANet is based on a simple recurrent network (SRN; Elman, 1990) that has been modified to optionally take the representation of a scene and produce a case-role interpretation

of the input utterance (see Figure 2). Processing is incremental, with each new input word interpreted in the context of the scene, if present, and the sentence processed so far, as 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 a sentence is completely processed. Moreover, during the course of processing a sentence these expectations can be overridden with subsequent input, often resulting in the abrupt revision of an interpretation in a manner strongly 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 adaptation to new information that have endeared connectionist models to cognitive researchers.

The encoding of the scene used by CIANet features three characters involved in two events (cf., **wizard spies-on pilot** and **detective serves pilot** in Figure 1). The middle character (e.g., **pilot**) is involved in both events as a patient. Only one of the events, however, will be relevant to the input utterance.

The representations for the characters and actions in each event are fed into the network's hidden layer by shared agent, action, and patient connections. The result is that the two events' constituents are effectively and separately superimposed. Thus, there is no explicit binding of each event's constituents; with shared weights, any of the constituents could go with any other. CIANet solves this problem through the use of an attentional mechanism that dynamically binds events, and is described in the next section.

The who-did-what-to-whom was encoded for the events, when depicted; grammatical information came from the linguistic input. The SRN consisted of input and output assemblies of 144 units each. The input assemblies comprised the six constituent representations in the scene and the current word from the input sentence. The output assemblies made up the verb, the first and second nouns, and a discriminator 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 CIANet performed better when the task was recast as having to learn to isolate the nouns in the order in which they are introduced in the utterance, 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. It was trained with backpropagation-through-time (Rumelhart, Hinton, & Williams, 1986) with a learning rate of 0.002.

## Event Selection using Sigma-Pi Units

Mayberry et al. (2005) used explicit *event layers* to build compressed representations of the two events in the scene. The compression process served to bind the entities and actions in the event together so that the network could access the compressed information and make reliable predictions about an upcoming argument once it had enough information (such as the patient and verb to predict the relevant agent). The task was complicated considerably by the fact that the two events

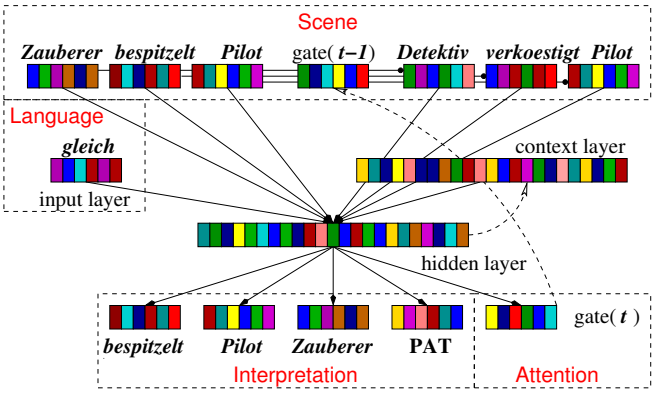


Figure 2: **Attention through Multiplicative Connections** The network modulates each event through a vector called a gate that functions as an assembly of sigma-pi units multiplied element-wise with each constituent of an event. The black circles indicate that the complement (one minus each element of the gate) is multiplied element-wise against each constituent of the other event. At each step of processing, the gate is updated so that the most informative event has the most influence on reducing the overall error.

were in effect superimposed over each other by virtue of the shared links between the event layers and the hidden layer of the network. Yet, the shared weights were deemed important as they were meant to represent a single pathway used to recognize each event in a scene.

Pilot studies suggested that event binding could instead be effectively achieved by scaling the two events themselves by differing degrees, and the network could use the resulting distinction to produce the correct output. The agents, the actions, and the patients of the two events would then be fed directly to the hidden layer through shared weights, leaving open the possibility of adding more events in future experiments. The empirical question, then, was how to most effectively scale the two events.

The most straightforward approach would have employed a single *gating unit* that selected one event or the other by scaling them to sum to 1, so that the more one event is activated, the less is the other. Thus, if one event was scaled by 0.9, then the other would be simply scaled by 0.1. Yet, it proved difficult to train the network to perform this appealingly simple operation with just one gating unit. The reason is that minimizing the error between outputs and targets meant $(4 \cdot 144 =)$ 576 units were contributing to the overall error, overwhelming the signal from the single gating unit and resulting in erratic behavior when miniscule changes in the gating unit were amplified through recurrency. Various attempts to improve its performance, such as altering the learning rate, changing the gain, and phased training had little effect.

However, extending the gating unit into a *gating vector* (or gate) of the same size as the lexical representations (144 units) proved effective. The gating vector basically transforms the architecture into a recurrent sigma-pi network (Rumelhart et al., 1986). The units of the gate are multiplied element-wise with the corresponding units in each of the three lexical representations comprising the agent, action, and patient of an event (see Figure 2). To maintain the con-

straint that, the more active one event is, the less active the other, each unit of the gate is subtracted from one to derive a vector complement that then modulates the other event's constituents. In effect, the gate serves as a common *mask* for the constituents of one event that is optimized through training to increase contrast by suppressing the elements of the other event so as to minimize the error from the target case-role interpretation of the sentence. The result is that the average activation of the gating vector directly correlates with greater activation of the attended event in a scene, effectively implementing an attentional mechanism. Crucially, the network is never taught which event to attend to. Because the gate functions essentially as a second hidden layer, attention to the most relevant event develops automatically on the basis of error information from the multiplicative connections to the modulated constituent representations of each event which is backpropagated recurrently during training.

## Training and Test Data

Recall that the four conditions in the experimental design were used to measure the interaction of stereotypical thematic role knowledge and information from depicted events in a scene. In two of the conditions, only one or the other of these two information sources was available, whereas in the other two conditions, both sources were available and conflicting. A major objective of the current study was to show that the model could learn to correctly resolve the conflicting conditions when trained only on the non-conflicting conditions. Additionally, the model should perform correctly in the absence of a scene, anticipating the stereotypical agent at the verb. Once it reads the final noun, it should produce that noun as the correct agent for the utterance, possibly overriding the anticipated filler.

The training corpus used in this study was based on sentence templates of the two conditions with nonconflicting information sources. These sentences either involved stereotypicality, in which case neither depicted event showed an action that corresponded with the verb in the sentence; or they involved only the scene, in which case no stereotypical agent for the verb in the sentence was depicted. Twenty-four verbs were used, together with their stereotypical agents, which is half of that used in Mayberry et al. (2005), but SVO versions of all sentences were added to expose the network to greater sentence variation. The training corpus was generated from all possible combinations of referents in both OVS and SVO word orders, while strictly holding out the original experimental materials as the test set. Because all the scenes in these materials featured an action in one event and a plausible agent for the action in the other event (cf. *bespitzelt* and *Detektiv* in Figure 1), the network could potentially learn to use this purely scene-based correlation to accomplish its task. Accordingly, all such cases were filtered from the training corpus to remove any source of subtle bias that might confound the results. These measures ensured that the test set was as novel to the network as possible. Indeed, where an event was relevant during training, it was irrelevant during testing, so the network had to learn to ignore or suppress it to produce the correct output. There were 13,632 sentences in the training corpus, each of which had an event that was most relevant to the sentence, and could be paired with one

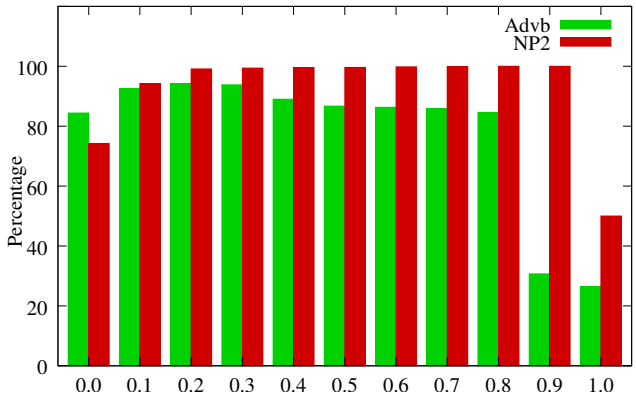


Figure 3: **Effect of Stereotypicality parameter** The bar plot shows the average accuracy of CIANet in correctly identifying the agent as measured both at the adverb (anticipation) and sentence-final (comprehension) for stereotypicality ratios from 0.0 to 1.0. A clear preference for the depicted agent is evident across all ratios up to 0.8.

of 6912 events generated randomly. Both OVS and SVO test sets had 96 sentences and scenes based on the twenty four verbs across the four conditions. All lexical items were given 144-dimensional binary random representations to remove any features the network could use to develop correlations that might confound the study. As an unbiased estimate of human exposure to language in situated settings, the network was trained on sentences with scenes half of the time, and half of the time without.

In order to measure the relative influence of stereotypical information versus depicted events, a single parameter, the *stereotypicality ratio*, was manipulated during training that controlled the relative frequency of sentences that appeared with stereotypical agents to those with non-stereotypical agents. Because the lexicon in the current study featured 24 verbs, each with its own stereotypical agent, the stereotypicality ratio had to be greater than 1/24 (0.04167) for the network to learn stereotypicality at all; otherwise, any ostensibly stereotypical agent would appear as frequently as any other. The greater this ratio, the stronger the association of a verb with its stereotypical agent. If the ratio is too large, then the network would learn the stereotypical association to the exclusion of all others.

## Results

Figure 3 reports the performance of CIANet for stereotypicality ratios from 0.0 to 1.0. Accuracy is given as the percentage of targets at the network's output layer that the model correctly matches (based on human performance), both as measured at the adverb (anticipation) and at the end of the sentence (comprehension). Performance is measured at the adverb rather than the verb because integration of the verb with information from the scene causes the network to shift attention to the relevant event, which manifests itself on the next word. The process is loosely analogous to the use of the adverb region in the analysis of the eye movements to allow time for people to process the verb in the utterance and attend to the scene. The model clearly demonstrates the qualitative behavior observed in the experiment in that it is able to access

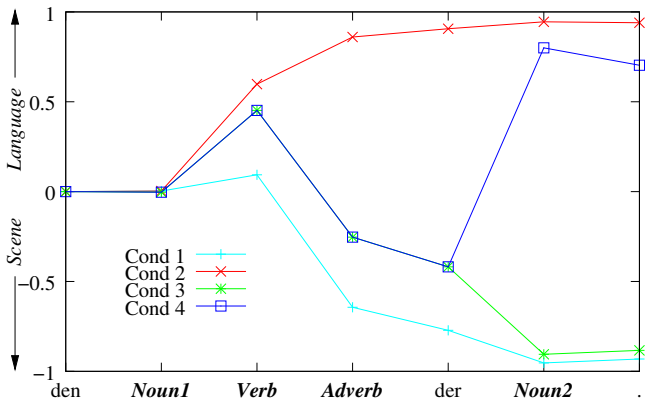


Figure 4: **Coordinated Interplay of Information Sources** The varying preference for the stereotypical (positive) versus depicted (negative) agent averaged for each of the four conditions over the test set clearly shows the model's ability to adapt to information as it processes a sentence incrementally.

the scene information and combine it with the incrementally presented sentence to anticipate forthcoming arguments.

Crucially, the model learns the observed resolution in favor of the depicted action when the two information sources conflict, despite never having been trained to do so. Moreover, this performance is not simply a matter of setting the stereotypicality ratio just right, but is robust over a wide range of parameter settings. The best overall performance is obtained for settings of $0.2$ and $0.3$, at which the model predicts more than 94% of the upcoming arguments correctly, and matches over 99% of all arguments at the end of the sentence. The anticipation accuracy does decrease as the stereotypicality ratio is increased, but even at $0.8$, it correctly identifies the depicted agent over 83% of the time. The amount of training for the network to converge on this level of performance also increases, taking approximately five times as long for a setting of $0.8$ as for $0.2$. For settings greater than $0.8$, performance failed to improve, despite extensive training. For $1.0$, this failure is hardly surprising since the network only learns stereotypical associations, but for $0.9$, performance may yet improve with more training.

Is the gating vector even necessary? Could the network learn to produce the correct response just from the superimposed patterns of the events to identify and correlate their constituents? Several models were trained and tested with no gating vector whatsoever to select events. On average, these models achieved approximately 50% (i.e., chance) correct anticipation of the upcoming agent.

Figure 4 gives a clearer view of how attention shifts over the course of sentence processing for a stereotypicality ratio of 0.5. The plot shows the difference between the Euclidean distances of the network's second noun output to the two agents in the scene, normalized so that positive values indicate a preference for the stereotypical (language) agent, and negative values indicate a preference for the depicted (scene) agent. These values were collected for all OVS test sentences and averaged for each of the four experimental conditions. The network shows no preference for either event agent as it processes the first noun phrase, *den Noun1*, because the patient appears in both events. A preference for the stereotypical agent over all conditions is evident at the point that the model has just processed the input verb, but not yet shifted attention to the most relevant event. This behavior is a prediction of CIANet that should be amenable to experimental verification. The initial preference makes sense for Cond 2-3 because a stereotypical agent does appear in the scene, but for Cond 1 it reflects a very slight negative correlation between the input verb and depicted agent that has developed as an artifact of the limited number of verbs used in the study. Nonetheless, the effect is completely overridden on the next step at the adverb once the network has shifted attention to the most relevant event, and the relative influence of language and scene are clearly manifested. For Cond 1, in which only case marking on the first NP and thematic role information from the processed verb combine with information from the depicted event, there is a strong preference for the depicted agent. For Cond 2, there is likewise a strong preference for the stereotypical agent since the processed input verb has no corresponding depicted action. The two conflicting conditions are identical up to the final noun phrase, and the interaction between the language and scene is evident in the network's shifting anticipation. Yet, the network does show a clear preference for the depicted over the stereotypical agent at the adverb. Finally, the zigzag form of the Cond 4 curve in Figure 4 attests to the ability of CIANet to rapidly adapt to information as it becomes available: at the verb, stereotypicality is the most informative source, which is integrated with information from the scene on the next step to shift attention to the relevant event supporting anticipation of the depicted agent, but finally overridden on the final noun, which turns out to be the stereotypical agent.

## General Discussion and Future Work

CIANet is a recurrent sigma-pi neural network that was motivated by, and directly implements, the coordinated interplay account (CIA) of situated utterance comprehension. The use of an attentional mechanism enables the model to exhibit a number of important *cognitive properties*. The model operates incrementally, integrating an utterance word by word with information from a scene, if present. It is also adaptive, able to perform correctly when there is no scene, and, in general, avails itself of whatever information is present. The model accurately anticipates upcoming arguments based on either stereotypical knowledge or information from the scene. The manner in which the events are selected can be seen as instantiating the CIA: the utterance causes the network to activate the gating vector to select the most relevant event in the scene, which then directly influences the network's full interpretation, as revealed by what it anticipates.

In addition to this cognitive behavior, the *primary experimental modelling* result of this study is that the network correctly learns to resolve conflicting information sources in favor of the immediate scene over stereotypical knowledge, despite only being trained on nonconflicting sentences. This means that the model is no longer "just fitting" the data, but generalizing in a novel manner. Significantly, the result holds over a wide range of ratios for the relative frequency of sentences in the training corpus that have a stereotypical versus a non-stereotypical agent. The network takes longer to learn to use the scene information correctly as the stereotypicality

ratio increases, but the scene ultimately has a stronger influence on the interpretation once it has been integrated with the input sentence. The reason for this behavior is that information from the scene is available as each word of the input sentence is processed, whereas the stereotypical information only comes into play once the verb or its stereotypical agent is processed. Because the network must learn to identify and attend to the relevant event in the scene, its relative influence becomes amplified with training.

To model the empirical results, the *main innovation* of the model is the use of a gating vector to directly modulate the two events fed into the SRN through shared weights. The purpose of the gating vector is to implement an attentional mechanism that can be more directly compared with human behavior as observed in psycholinguistics experiments. The assembly of sigma-pi units allows the network to select the relevant event itself by effectively molding the shared weights so that they capture the distributional characteristics of the events within the task of producing the desired output for the unfolding utterance. More research is needed to understand the exact mechanism through which the multiplicative units select the appropriate event, but current analysis suggests that they function like a mask to increase contrast between the two events by reducing the bits in the agent and verb representations of the irrelevant event that interfere with the recognizable propagation of the relevant event. This behavior accords well with evidence that attention–at least at the cellular level–also works by increasing the discriminatory response among stimuli (Taylor, Hartley, & Taylor, 2005).

Finally, the gating vector leads to a more *parsimonious* model in which the attentional mechanism is also directly involved in binding the event participants together. This approach is a fundamental improvement over Mayberry et al. (2005) because the elimination of the event layers simplifies the architecture and results in training times that are faster by up to an order of magnitude on the same corpus. Furthermore, CIANet is able to learn to reliably make the correct conflict resolutions when trained only on the two nonconflicting conditions, whereas the earlier model was not.

Future research will focus on adding material to the current training set to cover broader experimental results, including the experiments initially modelled in the first simulation reported in Mayberry et al. (2005). A particularly promising direction, moreover, would be to exploit the attentional mechanism to handle more complex, possibly dynamic, scenes. Lastly, the gating vector representations and their modulation of the scene constituents suggest that they may provide a more accessible way to develop a linking hypothesis between the attentional mechanism and the gaze probabilities observed in Knoeferle and Crocker (in press).

## Conclusion

CIANet is a recurrent sigma-pi neural network architecture that successfully models situated utterance comprehension, both when the individual information sources uniquely identify an interpretation and when they conflict. The model also performs correctly with and without the scene. The primary innovation of the network is the introduction of an attentional mechanism to select the scene event most congruous with the developing interpretation.

## Acknowledgements

This research was supported by SFB 378 project "ALPHA", funded by the German Research Foundation (DFG).

## References

Altmann, G. T. M., & 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.

Kamide, Y., Scheepers, C., & 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., & Crocker, M. W. (in press). The coordinated interplay of scene, utterance, and world knowledge: evidence from eye-tracking. *Cognitive Science*.

Knoeferle, P., Crocker, M. W., Scheepers, C., & 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., Crocker, M. W., & Knoeferle, P. (2005). A connectionist model of sentence comprehension in visual worlds. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society.* Mahawah, NJ: Erlbaum.

Miikkulainen, R. (1997). Natural language processing with subsymbolic neural networks. In A. Browne (Ed.), *Neural network perspectives on cognition and adaptive robotics* (pp. 120–139). Bristol, UK; Philadelphia, PA: Institute of Physics Publishing.

Roy, D., & Mukherjee, N. (2005). Towards situated speech understanding: Visual context priming of language models. *Computer Speech and Language*, *19*(2), 227–248.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition, Volume 1: Foundations* (pp. 318–362). Cambridge, MA: MIT Press.

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

Snow, C. E. (1977). Mothers' speech research: from input to interaction. In C. Snow & C. Ferguson (Eds.), *Talking to children: language input and acquisition.* Cambridge, MA: Cambridge University Press.

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

Taylor, J., Hartley, M., & Taylor, N. (2005). Attention as sigma-pi controlled ACh-based feedback. In *Proceedings of the International Joint Conference of Neural Networks.* IEEE Computer Society.