Stored Knowledge versus Depicted Events: what guides auditory sentence comprehension?

Pia Knoeferle (knoeferle@coli.uni-sb.de)
Department of Computational Linguistics,
Saarland University, 66041 Saarbrücken, Germany

Matthew W. Crocker (crocker@coli.uni-sb.de)
Department of Computational Linguistics,
Saarland University, 66041 Saarbrücken, Germany

Abstract

In a seminal article, Tanenhaus, Spivey-Knowlton, Eberhard & Sedivy (1995) showed that eye-movements to real-world objects reflect a rapid interplay of utterance and visual environments in sentence comprehension. Further, Kamide, Scheepers & Altmann (2003) found that when *linguistic and world knowledge* constrain the domain of reference in a visual scene, people even anticipate as yet unmentioned arguments/referents. Studies by Knoeferle, Crocker, Scheepers & Pickering (2003) have since revealed that when linguistic and world knowledge did not disambiguate an initial syntactic and role ambiguity, *depicted agent-action-patient events* permitted anticipation of thematic role-fillers online. This paper opposes linguistic and world knowledge on the one hand, and visual scenes on the other hand in order to determine their relative importance in auditory comprehension. We observed a preferred reliance of auditory sentence comprehension processes on information that had to be extracted from depicted event scenes. Determining the nature and time-course of the interaction between linguistic/world knowledge and visual scenes is a first step towards developing a model of real-time auditory sentence comprehension in visual environments. Our finding has implications for theories of the language faculty (e.g., Jackendoff, 2002).

Introduction

How do we understand utterances online in visual environments? As a first hypothesis, we might assume that the presence of visual environments does not affect language comprehension processes. Crucially, however, language refers to things in the world. It has further been demonstrated that in on-line comprehension reference to entities is established rapidly, and that referentially relevant non-linguistic information influences how the linguistic input is structured (Tanenhaus, Spivey-Knowlton, Eberhard & Sedivy, 1995). In auditory comprehension in visual environments, the unfolding utterance provides linguistic, semantic, and world knowledge. The scene in turn *affords* information about entities and events in the immediate environment (e.g., an outstretched hand affords the prospect of shaking hands) (Gibson, 1966; see Steedman, 2002, for a formal description of object and event affordances). Mental representations that issue in parallel from distinct cognitive components such as the auditory and visual system have to be integrated in "real-time". Comprehension in visual scenes hence moves from single-mode understanding to bi- or even multi-modal understanding.

The architecture of the language faculty

Since visual scene and linguistic information interact in online comprehension (Tanenhaus et al., 1995; Knoeferle et al., 2003), adequate description of auditory comprehension in visual environments requires that we embrace a theoretical account which situates language comprehension with respect to other cognitive systems such as the visual, auditory, or motor system. Jackendoff (2002) proposes one such framework. The individual levels in Jackendoff's architecture are modular in the sense of being domain-specific (i.e., their representational vocabulary is specialized), but unlike Fodorian modularity (Fodor, 1983), linguistic structures interact with one another, and with other cognitive sub-systems. This variety of modularity permits communication between phonological, syntactic, and conceptual structure. Moreover, it also allows communication between conceptual structure and perception or action via interface processors (Jackendoff, 2002, pp. 220f.).

While Jackendoff's theory provides for the interaction of linguistic/world knowledge and visual information, it is underspecified with respect to the precise nature and time-course of this interaction. In order to develop a model of real-time comprehension on the basis of his architecture, further experimental work is required. As a first step in this direction, we need to clarify how linguistic and world knowledge is integrated with object and event affordances that have to be extracted from the visual environment.

Previous work emphasizes both the importance of visual scenes in determining online comprehension when the linguistic input was structurally ambiguous (Tanenhaus et al., 1995; Knoeferle et al., 2003), and the importance of stored linguistic/world knowledge in anticipating which object in a scene will be referred to next (Kamide et al., 2003)[1].

[1] We use the term 'stored knowledge' to refer to linguistic/world knowledge which is stored in memory. The terms 'world knowledge' and 'stereotypical knowledge' are used synonymously to mean stereotypical relationships between scene entities (e.g., a cat is a stereotypical agent for chasing mice).

Stored knowledge versus depicted event scenes

Kamide et al. (2003) have shown that unambiguous case-marking, lexical expectations and world-knowledge influence anticipation of post-verbal arguments depicted in the scene. In German, a case-marked article can determine the grammatical function and thematic role of the noun phrase it modifies. Both SVO (subject-verb-object) and OVS (object-verb-subject) orders are grammatical. Participants inspected images showing a hare, a cabbage, a fox and a distractor object while hearing sentences such as *Der Hase frisst gleich den Kohl* ('The hare (subj) eats soon the cabbage (obj)') and *Den Hasen frisst gleich der Fuchs* ('The hare (obj) eats soon the fox (subj)'). The subject and object case-marking on the article of the first noun phrase together with world knowledge extracted at the verb allowed anticipation of the correct post-verbal referent. This was evidenced by anticipatory eye-movements to the cabbage after participants had heard 'The hare (subj) eats …' and to the fox after having encountered 'The hare (obj) eats …'. Hence, when the utterance is unambiguous, and linguistic/world knowledge restricts the domain of potential referents in a scene, the comprehension system may anticipate mention of scene objects.

Knoeferle et al. (2003), in contrast, considered ambiguous utterances, where neither case-marking nor stereotypical knowledge could assist in disambiguation. Specifically, they examined the time course with which listeners were able to resolve an initial structural and thematic role ambiguity in German sentences. As the linguistic input did not determine the correct syntactic analysis and thematic role-assignment of the sentence, listeners had to rely on depicted events in the scene for interpretation of the utterance. The events showed, e.g., a princess washing a pirate, while a fencer painted her. The princess was thus determined as either patient or agent of an event depending on the depicted action (washing/painting respectively). Listeners heard *Die Prinzessin wäscht/malt den Pirat/der Fechter*. ('The princess (amb.) washes/paints the pirate (obj./patient)/the fencer (subj./agent)'). Once the verb had identified the relevant depicted action, anticipatory eye-movements to the appropriate other event participant (the pirate or the fencer) were observed. The anticipation of a patient and agent role-filler for initially ambiguous German subject-verb-object and object-verb-subject sentences respectively suggests rapid use of depicted events in resolving structural and thematic role ambiguity online. This finding was also shown for the English main clause (MC)/reduced relative (RR) ambiguity, and hence generalized to another language and construction.

Teasing apart the relative effects of visually perceived events and stored linguistic/world knowledge in online sentence comprehension is of relevance for theories of the architecture of the language system such as the one proposed by Jackendoff (2002). Such endeavor may ultimately allow us to propose a more concrete model of processing mechanisms within such a framework and to hence develop it into a situated model of real-time sentence comprehension. What the above studies have shown, is that stored knowledge and visual-scene information are both rapidly applied, and that each may guide comprehension processes online. It is further clear, that utterance, world knowledge and the immediate visual scene interact during online comprehension. What remains unclear is the nature and time-course of that interaction. Among other questions, we might ask: What is the relative importance, or priority of different information sources, such as linguistic, scene, and world knowledge? Does scene information guide comprehension, or is the use of scene information determined by the utterance?

To further investigate this empirical question, consider an example from German. As noted above, German has a rich case marking system where grammatical function is usually indicated by unambiguous case morphemes. Word order constraints are less rigid in German than in English, and both subject-verb-object (SVO) and object-verb-subject (OVS) order are grammatical with SVO being the preferred reading (e.g., Hemforth, 1993). On hearing an OVS sentence fragment such as *The pilot* (object/patient[2]) *jinxes*… while inspecting the example scene in Figure 1, a number of processes occur. When we hear *The pilot*, object case-marking permits assignment of a patient role to the noun phrase while establishing reference to the pilot in the scene. As there are no constraints on inspecting the scene, perceivers might notice a wizard holding a telescope, and a character resembling a detective who is serving some food. Having encountered an agent and the verb, we might expect post-verbal mention of an agent. At this point, our knowledge that a wizard is a likely jinxing-agent can combine with the fact that the wizard is the only entity whose affordances match the expectations raised by the verb. The combination of entity affordance and stereotypical knowledge allows us to anticipate the wizard as a likely-to-be-mentioned agent. The decisive contribution is, however, made by the utterance, as the verb provides knowledge of stereotypical thematic role-fillers of a jinxing-action (a wizard). The scene affords no information about a jinxing-*relation* between two event participants, as there is no depicted jinxing-action. In contrast, when we hear *The pilot* (object/patient) *serves food to*…, verb-based knowledge of stereotypical agents of a serving-food action (e.g., a cook), cannot provide any guidance, as the scene affords no such entity. However, the scene does afford a depicted food-serving event performed by the detective. While stereotypical knowledge does not allow determination of thematic role-relations in this case, the affordances of the depicted scene events do. Based on findings by Kamide et al. (2003) and Knoeferle et al. (2003), we would expect unmistakable resolution of the temporal uncertainty regarding the yet-to-be-mentioned agent in both of the above examples once people have heard the verb.

Imagine we heard instead *The pilot* (object/patient) *spies-on*… In this case, the scene affords both a stereotypical

[2] In our materials, a grammatical subject and object correspond to an agent and patient respectively in the scenes.

agent (the detective), and an immediately depicted agent of a spying event (the wizard) as potential agents (see Fig. 1). When we hear *spies-on*, lexical access makes available the meaning of the lexical item and stereotypical knowledge related to it (see Ferretti, McRae & Hatherell, 2001). After encountering the verb, word meaning, stereotypical knowledge of *spy-on*, and scene affordances are available to anticipate either a depicted spying-event and its agent (the wizard), or a stereotypical agent (the detective).

Do listeners rely more on extracting thematic role relations from stereotypical knowledge provided by the utterance (*spy-on* + WORLD KNOWLEDGE –> *detective*), or do they rely on thematic relations afforded by the scene (*spy-on* + WIZARD-SPYING-EVENT -> *wizard*) in incremental interpretation? For the ambiguous *spy-on* example thematic role-relations that are provided by the visual scene, conflict with stereotypical knowledge of who-does-what-to-whom. The comprehension system has to choose between two available, yet conflicting types of information in determining online thematic role-assignment.

While Jackendoff's framework does not make explicit predictions, there are reasons to expect a priority of stereotypical/world knowledge in online thematic role-assignment in such architecture. Jackendoff (2002, pp. 282) argues against a strict separation of linguistic meaning and world knowledge (see also Levinson, 2000). Experimental evidence confirms such an assumption. Ferretti et al. (2001) found that verbs immediately activated stereotypical knowledge of agents (*arresting-cop*) or patients (*arresting-criminal*), but not locations (*swam-ocean*). They conclude that this type of world knowledge is part of thematic-role knowledge, and immediately activated upon encountering the verb (see also McRae, Ferretti & Amyote, 1997). Stored knowledge about stereotypical agents is hence readily available for online thematic role-assignment processes. Non-stereotypical thematic role-relations afforded by scene events, however, must be newly acquired by a perceiver, and via a different perceptual system (the visual system). On the basis of experimental evidence for the tight coupling between word meaning and world knowledge (e.g., Ferretti et al., 1997), we expect comprehension processes to rely in preference on stored knowledge over scene information. Indeed, such a prediction would also appear to follow from traditional assumptions concerning the modularity of the language faculty with respect to other perceptual faculties such as the visual system (e.g., Fodor, 1983).

In summary, we expect the following: when the verb determines either a depicted or a stereotypical target agent only (Table 1, a1 and a2), both verb-derived knowledge of stereotypical role-fillers and affordances of the scene events should allow anticipation of the appropriate target agent. This would replicate - within a single study - findings by Kamide et al. (2003) and Knoeferle et al. (2003). When the verb is *serves-food-to* (a1), we expect a higher percentage of anticipatory looks to the only depicted food-serving agent (the detective) than to the respective other agent in the scene (the wizard). Conversely, when the verb is *jinxes* (a2), more looks should occur to the stereotypical agent (the wizard) than to the other agent in the scene (the detective). In the interesting case of competition, when the verb (*spy-on*) allows more than one potential scene entity as target agent, no interaction is expected. Rather, we should observe a main effect, with the point of interest being the direction of the main effect. The fact that stored stereotypical knowledge is readily available from memory, argues for guidance of online thematic role-assignment processes by stereotypical knowledge rather than by event affordances acquired from the scene in a Jackendoffian framework. This should be revealed in a higher percentage of inspections to the stereotypical spying-agent (the detective) than to the depicted spying-agent (the wizard) for sentences b1 and b2. Crucially, these looks should occur before people hear the disambiguating second noun, and hence reveal online expectations of thematic role interpretation (see Fig. 1 and Table 1).

Experiment 1

Method

Participants Twenty-four German native speakers with normal or corrected-to-normal vision were paid 5 euro for taking part in the experiment. Some of them had already participated in an eye-tracking-during-listening experiment.

Materials We created 48 images using commercially available clipart and graphic programs. For each of these images, a female native German speaker recorded 4 sentences, which described either a depicted event (e.g., wizard-spying) or a stereotypical event (e.g., detective-spying, see Fig. 1; Table 1).

Design A set of 24 items was created. Each item consisted of 8 spoken sentences and 2 images (Table 1 and Fig. 1 show examples for the 4 sentences for one image of an item set). The two versions of an image only differed in the actions performed by the respective characters. This ensured that each of the target agents (wizard, detective) was a stereotypical and a depicted agent in turn, and that each verb referred once to a depicted event, and once to a stereotypical event. Actions were typically depicted as a character holding an instrument. The way in which actions or characters were depicted did not differ between the two image versions. The middle character on each image (e.g., the pilot) was always a patient ('being acted upon'). The entities to the left and the right of the patient character were performing an action upon the patient entity, and hence always had an agent-role. An example image (see Fig. 1) showed two such agent-action-patient events, e.g., wizard-spying-on-pilot and detective-serving-food-to-pilot.

Table 1: Sentences for the example image in Figure 1

Image	Condition	Sentence	PATIENT	VERB	|------ ADV----|	AGENT	
Fig. 1	No-Competitor & depicted agent	a1	Den Piloten	*verköstigt*	gleich der	Detektiv.	
			The pilot (PAT.)	serves-food-to	soon the	detective	(depicted AGENT)
			'The detective will soon serve food to the pilot.'				
Fig. 1	No-Competitor & stereotypical agent	a2	Den Piloten	*verzaubert*	gleich der	Zauberer.	
			The pilot (PAT.)	jinxes	soon the	wizard	(stereotypical AGENT)
			'The wizard will soon jinx the pilot.'				
Fig. 1	Competitor & depicted agent	b1	Den Piloten	*bespitzelt*	gleich der	Zauberer.	
			The pilot (PAT.)	spies-on	soon the	wizard	(depicted AGENT)
			'The wizard will soon spy on the pilot.'				
Fig. 1	Competitor & stereotypical agent	b2	Den Piloten	*bespitzelt*	gleich der	Detektiv.	
			The pilot (PAT.)	spies-on	soon the	detective	(stereotypical AGENT)
			'The detective will soon spy on the pilot.'				

Figure 1: Example of an image for Experiment 1

In addition to the affordances of the depicted events (a telescope affording a spying action), each of the characters also had entity affordances (a detective affording a spying action). The event and entity affordances were always incongruous for any one entity in this experiment (e.g., a detective was never depicted as performing a spying action). Rather, one agent on each image was a stereotypical competitor for the depicted event performed by the other agent (e.g., the detective was a stereotypical competitor for the depicted wizard-spying event), while carrying out a different action (serving-food) himself. By manipulating the verb people heard, we created four conditions, crossing the factors *competitor* (competitor, no-competitor) with *information type* (depicted target, stereotypical target). For the no-competitor conditions (see Table 1, sentences a1 and a2), the verb permitted either a depicted or a stereotypical agent only: "verköstigen" ('serve-food-to') determined the detective (Table 1, a1) as depicted agent; "verzaubern" ('jinxes') identified the wizard as stereotypical agent (Table 1, a2). For the competitor condition (see Table 1, sentences b1 and b2) the verb "bespitzeln" ('spy-on') allowed two scene entities as likely agents (Fig. 1): the wizard, being depicted as performing a spying-action, and the detective, a stereotypical agent for a spying-action. Sentences were unambiguous OVS sentences (see Table 1). They always started with an object case-marked noun phrase referring to a patient role-filler (Fig. 1, the pilot). The middle character was not engaged in an action, and its gaze and position did not bias towards either the left- or the rightward entity. Conditions were matched for length and frequency as much as possible (CELEX). For the image in Figure 1, the sentences in Table 1 were recorded.

Procedure An SMI Eye-Link head-mounted eye-tracker monitored participants' eye-movements. Images were presented on a 21″ multi-scan color monitor at a resolution of 1024 x 768 pixels together with the spoken sentences. Each participant saw only one condition of each item, and the order of appearance of items was randomized individually for every participant. It was further ensured that no participant heard any utterance or part of it more than once. There were eight experiment lists. Each consisted of 24 experiment and 32 filler items. Consecutive experiment trials were separated by at least one filler trial. Before the experiment, participants were instructed to listen to the sentences and to inspect the images. There was no other task. The entire experiment lasted approximately 30 min.

Analysis The critical time region we chose for the analysis extended from the late verb (200 ms prior to adverb onset) to the determiner region of the second noun phrase (labeled 'ADV', see Table 1). For this region, participants had heard most of the verb, but had not heard the disambiguating second noun. The *X-Y* coordinates of participants' fixations were assigned to regions for the entities and scene background. Consecutive fixations within one object region (i.e., before a saccade to another region occurred) were added together, being counted as one *inspection*. For the inferential analysis hierarchical log-linear models were used, which combine characteristics of a standard cross-tabulation chi-square test with those of ANOVA. Log-linear models are adequate for count variables because they

neither rely upon parametric assumptions concerning the dependent variable (e.g., homogeneity of variance), nor require linear independence of factor levels (Howell, 2001). Entities were coded depending on their event role. For Figure 1, for instance, the wizard was coded as 'stereotypical agent', and the detective as 'depicted agent' for the no-competitor conditions (a1 and a2 respectively), and vice versa for the competitor conditions (b1 and b2, see Table 1). The inspections to a character within the ADV time region were a dependent variable in the statistical analysis. Inspection counts for the ADV analysis region were adjusted to factor combinations of *target character* (stereotypical agent, depicted agent), *competitor condition* (competitor, no competitor), *information type* (depicted target, stereotypical target) and either participants (N = 24) or items (N = 24).

Results

Figures 2 and 3 show the proportion of inspections to the target characters (depicted agent, stereotypical agent) during the ADV time interval. Figure 2 shows inspection percentages to the entities in the two information type conditions (depicted target, stereotypical target) for the No-Competitor condition, Figure 3 for the Competitor condition.

For the No-Competitor condition (a1, a2, see Table 1), when the verb singled out either a depicted (a1) or a stereotypical agent (a2), a significant interaction of information type (depicted target, stereotypical target) and target character (depicted agent, stereotypical agent) revealed clear disambiguation using either depicted or stereotypical information (all ps < 0.01). This was due to a significantly higher percentage of inspections to the depicted agent in the depicted (a1) than in the stereotypical target condition (a2), and a significantly higher percentage of inspections to the stereotypical agent in the stereotypical target condition (a2) than in the depicted (a1) (see Fig. 2).

In contrast, for the Competitor condition (b1, b2, see Table 1), when stored stereotypical knowledge and scene affordances competed and provided conflicting information, we expected no interaction since the stimuli between b1 and b2 were identical (see Table 1). Rather, we found a main effect. We observed more anticipatory looks to the agent of the depicted spying-event (the wizard), than to the stereotypical agent (the detective) for sentences b1 and b2 (see Fig. 3). These looks occurred after people had heard *The pilot* (object/patient) *spies-on*…. and before they heard the respective second noun, which then disambiguated towards the depicted or stereotypical target-agent type. Log-linear analyses showed that the main effect of depicted agent was significant (p < 0.001 by part. and items) in the absence of a significant interaction (ps > 0.6).

Importantly the observed main effect for the Competitor condition (b1 and b2) is only meaningful in comparison to the significant interaction found in the No-Competitor condition (a1 and a2). The difference between the main effect for the Competitor condition (b1 and b2), and the significant interaction for the No-Competitor condition (a1 and a2, see Table 1) was significant. Analyses revealed a three-way interaction between target character (depicted agent, stereotypical agent), competitor (competitor, no-competitor) and information type (depicted target, stereotypical target) (p < 0.05 by part. and items).

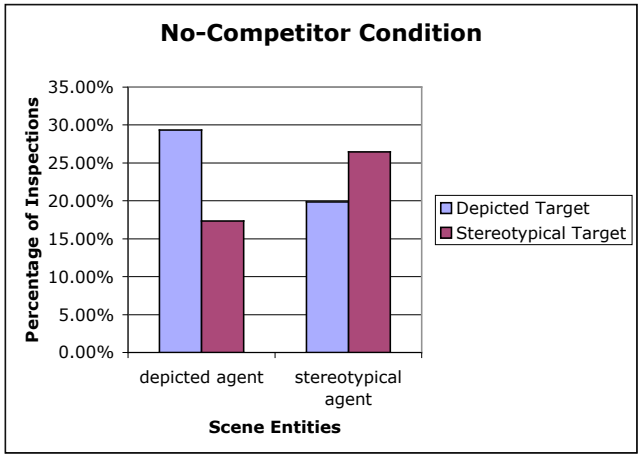

Figure 2: Percentage of inspections to characters during the analysis region ('ADV') in the No-Competitor Condition

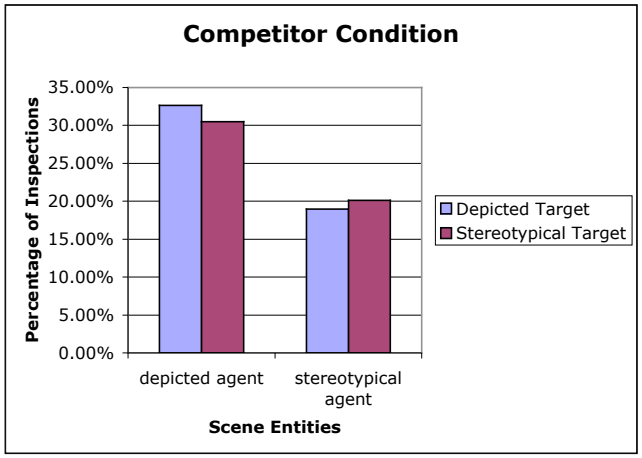

Figure 3: Percentage of inspections to characters during the analysis region ('ADV') in the Competitor condition

Discussion

Within a single study the observed pattern of eye-gazes has shown that both stereotypical knowledge and information that has to be newly acquired from depicted event scenes, allow rapid thematic role interpretation of an unfolding utterance. In the face of competition, however, people have a clear preference for relying on thematic relations that have to be extracted from depicted events. We argue that our findings are an important step towards developing a fully specified model of comprehension that is able to make explicit predictions of auditory sentence comprehension in visual environments. For architectures of the type proposed by Jackendoff (2002), they point to the necessity of incorporating a processing (rather than an architectural) account for the preferred reliance of the comprehension

system on thematic role-relations afforded by depicted event scenes.

Clearly, these findings of the relative importance of information that listeners had to newly acquire from depicted scenes, and via a different perceptual system (the visual system), counter our initial expectations of a priority of stored stereotypical knowledge. When utterance and scene compete and provide conflicting information, it is not stereotypical knowledge, which influences our interpretation of the scene. Rather, when verb meaning identified relevant depicted events, the scene guided interpretation of the utterance. Our findings hence indicate an active contribution of thematic role-relations afforded by scene events in online thematic role-assignment. It should nonetheless be highlighted that scene events only influenced thematic role-assignment once they had been identified by the verb.

Such utterance-mediated influence of depicted events suggests a highly efficient interaction between the visual and comprehension systems, where lexical items single out scene entities and events, hence making scene information available, which then in turn may influence online interpretation. However, this observation contradicts the assumption that there is *unconstrained* guidance of auditory comprehension by visual scenes. Under this view, we would expect that reference from verbs to the depicted actions in the scene is a necessary pre-requisite for the influence of scene events on online comprehension processes.

Studies by Knoeferle, Crocker, Scheepers & Pickering (2004) found, however, that depicted scene events allowed online thematic role-assignment and structural disambiguation even before people had encountered a sentence-final verb. The insight that emerges from their finding, and the ones presented in this paper, is that visual scene information has great importance in online comprehension. Indeed, reference from verbs to depicted actions is not an indispensable pre-requisite for the guiding influence of depicted events in auditory comprehension. In sum, this speaks to a *strong* – albeit not necessarily unconstrained - guidance account of the influence of visual scenes on online comprehension processes. Further research is required to establish whether there are constraints on the influence of depicted event scenes on online thematic role-assignment, and under which conditions they apply.

An Adaptive Perspective

In many cases, people will find themselves in relevant situations where both spoken language and immediate scene context are available. When watching movies, for instance, people seem to rapidly integrate their knowledge of the world with the movie events unfolding over time in front of their eyes. Clearly, cognitive processes such as understanding an unfolding utterance and an immediate event may occur simultaneously. Yet, when both types of information are available in our environment, the pattern of eye-gazes we observed provides strong evidence for a preference of the immediate environment over expectations of stereotypical events. Further, the rapid impact of the immediate situation identifies the comprehension system to be highly adapted towards acquiring new information from its environment rather than relying on linguistic and world knowledge, a conclusion which bears important implications for both developmental and evolutionary accounts of the language comprehension system.

Acknowledgements

We thank Martin J. Pickering for his comments on an earlier version of the design for the experiment, and Ulrich Pfeiffer for his help in running it. This research was funded by a PhD scholarship to the first author and by SFB 378 "ALPHA" to the second author, both awarded by the German research foundation (DFG).

References

Ferretti, T. R., McRae, K., & Hatherell, A. (2001). Integrating verbs, situation schemas, and thematic role concepts. Journal of Memory and Language, 44, 516-547.

Fodor, J. (1983). *Modularity of mind*. Cambridge, MA: MIT Press.

Gibson, J. (1966). *The Senses considered as perceptual systems*. Boston, Mass.: Houghton-Mifflin Co.

Hemforth, B. (1993). *Kognitives Parsing: Repräsentation und Verarbeitung sprachlichen Wissens.* Berlin: Infix.

Howell, D. C. (2001). *Statistical methods for psychology*. Belmont, CA: Duxbury Press.

Jackendoff, R. (1983). *Semantics and Cognition*. Cambridge, MA: MIT Press.

Jackendoff, R. (2002). *Foundations of language*. Oxford: Oxford University Press.

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*, 37-54.

Knoeferle, P., Crocker, M.W., Scheepers, C., & Pickering, M.J. (2003). Actions and roles: using depicted events for disambiguation and reinterpretation in German and English. In: *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (pp. 681-686), Boston, MA.

Knoeferle, P., Crocker, M.W., Scheepers, C., & Pickering M.J. (2004). The influence of the immediate visual context on incremental thematic role-assignment: evidence from eye-movements in depicted events. Manuscript submitted fro publication.

Levinson, S.C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. Cambridge, Mass.: MIT Press.

McRae, K., Ferretti, T. R. & Amyote, L. (1997). Thematic roles as verb-specific concepts. *Language and Cognitive Processes*, 12, 137-176.

Steedman, M. (2002). Formalizing affordance. In: *Proceedings of the 24th Annual Conference of the Cognitive Science Society* (pp. 834-839), Fairfax VA.

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.