First things first? Top-down influences on event apprehension

Johannes Gerwien (jo.gerwien@uni-heidelberg.de)
Institut für Deutsch als Fremdsprachenphilologie, Heidelberg University
Plöck 55, 69117 Heidelberg, Germany

Monique Flecken (monique.flecken@mpi.nl)
Max Planck Institute for Psycholinguistics
Wundtlaan 1, 6525 XD Nijmegen, the Netherlands

Abstract

Not much is known about event apprehension, the earliest stage of information processing in elicited language production studies, using pictorial stimuli. A reason for our lack of knowledge on this process is that apprehension happens very rapidly (<350 ms after stimulus onset, Griffin & Bock 2000), making it difficult to measure the process directly. To broaden our understanding of apprehension, we analyzed landing positions and onset latencies of *first fixations* on visual stimuli (pictures of real-world events) given short stimulus presentation times, presupposing that the first fixation directly results from information processing during apprehension.

Keywords: apprehension, visual attention, event construal, language production, cross-linguistic analyses

Introduction

What is event apprehension? In most sentence production research, it is assumed that during this phase the 'gist' of a scene is extracted from a visual stimulus (Henderson & Ferreira 2004). It is, however, not clear what features constitute the 'gist' of a scene, nor is there a clear differentiation between event apprehension and the next phase of the production process, namely, the generation of a pre-verbal representation ('message') in which the information that a speaker intends to verbalize, and in what way, is specified. Understanding the interrelation between these two processes though, is an important step towards better insights into the early phases of the language production process, and potential factors affecting it.

To date, studies on event apprehension show that it is a rapid process, during which scene category information and associated semantic knowledge becomes available (Henderson & Ferreira 2004), scene coherence can be evaluated (Dobel et al. 2007), and an event representation can be constructed in which visual entities are linked to event roles (agent/patient) (Griffin & Bock 2000). However, it has also been argued that only those features of objects or scenes needed for the task at hand can be perceived (Williams & Simmons 2000, Dobel et al. 2007), showing that top-down factors, i.e., task-demands, play a role already during early visual processing, and that apprehension is thus a *flexible* process - a process which can be adapted to specific circumstances. Furthermore, studies employing change blindness paradigms suggest that another top-down factor, i.e., the cultural and language background of the viewer, may also influence basic perceptual processes. Masuda & Nisbett (2006), for example, showed that East Asians are more sensitive to information in the periphery (changes in context) compared to Americans who are more sensitive to changes in focal objects. Here, we aim to explore the dimensions of the flexibility of the scene apprehension process, using eye tracking methodology.

Background

How has apprehension been studied previously? Using eye tracking, apprehension has been linked to the first, or the first few overt fixations in a trial (Griffin & Bock 2000, Bock et al. 2003). However, those early fixations have been reported to be unspecific, in the sense that no clear pattern of fixation locations was detected. First fixations tend to be placed in the middle of the stimulus, a location from where relevant information at different locations can be visually processed more or less equally well (Holmqvist et al. 2011). Moreover, Griffin & Bock (2000) interpret an increase in fixations toward one specific region of the stimulus, for example the region displaying the agent of an action, as indicating the completion of apprehension and the beginning of further linguistic processing, i.e., syntactic and phonological encoding.

Another approach to studying apprehension is to investigate what information can be extracted from a visual stimulus in what amount of time (Dobel et al. 2007). Even without overt fixations speakers are surprisingly good at identifying agents, objects, and actions, under very short presentation duration conditions (100-250ms). However, in this approach it is difficult to precisely understand what is going on during apprehension: When participants are asked to describe what they have seen, a representation of what the informant intends to say (the message) is obligatorily constructed. It is thus more than difficult to distinguish between apprehension and message generation, based on using the linguistic product (the descriptions produced) as the only measure.

In order to investigate questions about apprehension independent of message generation or other aspects of the production process, an empirical measure is needed which mainly reflects apprehension and which does not directly relate to linguistic processing. Here, we record and analyze *first fixation locations and latencies* given a manipulation of task requirements, i.e., by constraining the amount of time participants have for visual information uptake (presentation duration manipulation: 300ms, 500ms, 700ms) and by using

two different participant groups (speakers of German and Spanish). We argue that, under brief exposure times, first fixations are informative about apprehension, and not reflecting (other, later) stages in the production process.

As mentioned above, a manipulation of stimulus presentation duration has been applied previously; here, we are specifically interested in how the quality of event descriptions may change given different exposure times. By encouraging participants to produce full sentences, we can shed light on how *specific* the information extracted concerning different event elements under these circumstances can be. Furthermore, by analyzing linguistic content as well as first fixation locations/latencies we are able to compare the type of information overtly attended first to the type of information encoded in the linguistic product.

The rationale behind choosing two different languages is that, though message generation processes should be the same in speakers of any language (and they should be dependent on the same set of task- or stimulus-related factors), there is evidence of cross-linguistic variation in viewing behavior during scene description (Brown-Schmidt & Konopka 2008; Sauppe et al. 2013; Flecken et al. 2015; v. Stutterheim et al. 2016): Cross-linguistic differences concerned what type of information is considered most relevant, when, and effects were observed during message planning and/or formulation. Here, we ask whether language already exerts influence on apprehension in a description task, and thus whether the process may show flexibility in adapting to language-specific demands. Our test case is Spanish vs. German: The two languages have been shown to differ in the prominence accorded to two core event elements, i.e., the agent and the action; Spanish speakers tend to represent accidental events *action*- rather than agent-oriented (Fausey & Boroditsky 2011), whereas German speakers, conceptualize events with a strong focus on *agents* (Flecken et al. 2015). Differences are driven by specific linguistic means typically used to describe events in the two languages. We will use this variation in global agent- vs. action-prominence in event encoding to shed light on the flexibility of the apprehension process.

Aims of the present study

We address the contents and features of initial visual processing, i.e., scene apprehension, in sentence production. Our window on this process is the location of the very first fixation on stimuli showing photographed events (agent performing action on an object). Given that during apprehension the location for this first fixation must be calculated, analyzing in which region of the depicted scene the first fixation is registered should allow us to infer, at least to some extent, what information was derived from the stimulus up to this moment and what information the cognitive system demands next in order to fulfill the task, i.e., constructing a representation that can be verbalized.

A manipulation of stimulus exposure time (300ms-500ms-700ms) allows us to address the *flexibility* of the process under different degrees of 'pressure' on the system, given the complex task to produce full-fledged event descriptions. Furthermore, we investigate the flexibility of the process by contrasting speakers with native languages which differ in agent- vs action-prominence in event encoding.

If the presentation duration manipulation has no impact on the locations and/or latencies of first fixations, the apprehension process must be considered rigid, in the sense that it cannot adapt to time constraints on visual information uptake. If language shows no effect it must be concluded that apprehension is unaffected by linguistic or cultural variation.

Experiment 1 - Pretest

In Experiment 1 native speakers of German and Spanish described and rated a collection of photographs depicting everyday events, each involving an agent performing an action on an object. This experiment was done to ensure homogeneity in event recognition and labeling (e.g., choice of action verbs), and to control for a potential visual bias towards specific elements in the scenes.

Method

Participants Native speakers of German and Spanish (N = 10 in each group; university students 20-30 years of age) took part in the experiment. Data were collected at Heidelberg University, which has a large population of Spanish speaking exchange students (Erasmus programme).

Materials 73 different actions, each performed by a male and female agent, involving one specific object, were photographed. All scenes were staged identically against a white background, controlling for distance between agent and object, head, body and object position and the amount of space covered by agent and action in the photograph. There was also a mirrored version of each item.

Procedure Photographs were printed in black and white and presented to participants in a paper catalogue, in a randomized order. Each participant saw each action once, half performed by a male, half by a female actor. Participants were first asked to write down the type of event depicted (e.g. "to squeeze a lemon"). Then, they were asked to rate each photograph for a number of dimensions: - naturalness of the event, on a scale from 1 to 5 (1: unnatural; 5: natural); - ease of recognition of the action (1: action not recognizable; 5: action perfectly recognizable); - a potential visual bias of one over the other element (is either the agent or the action more easily recognizable, identifiable or prominent? or are both elements equal? (open question)). Participants were given as much time as needed, but they were instructed not to look back at previous items and to rate the pictures on the basis of their first impression.

Results

In total, 13 items had to be discarded due to a high level of heterogeneity in event description, most of them given by

cross-linguistic differences. For the selection of stimuli for Experiment 2, we made sure that the events were described in Spanish and German with a similar verbal structure and level of syntactic complexity, and that they could be described with a compact action verb. For example, the action 'ein Ei koepfen' ('to top off the lid of an egg') cannot be described by a single verb in Spanish (neither in English, actually). On the basis of the descriptions and the ratings, we selected two sets of 60 homogeneous scenes with a male and a female actor performing the same actions for Experiment 2. Selected pictures were rated 4-5 for the dimensions reported above; no visual biases were reported.

Experiment 2 - Online event description

Method

Participants In total, 40 participants were tested. Eight participants from each language were assigned to the 300ms presentation duration condition, eight from each language to the 500ms presentation duration condition and four from each language to the 700 ms condition.

Materials Figure 1 shows an example stimulus. Picture size was 600x400px.

Figure 1: Example stimulus with three Areas of Interests (AoIs); the AoIs were not visible to subjects

Procedure Eye movements were recorded with a remote Eye link 1000 system, running on Experiment Builder (SR research), with a 500 Hz sampling rate. The monocular recording mode was used. Participants were seated in front of a 19 inch computer screen, at a distance of appr. 65 cm.

Participants were given written instructions in their native language. A native speaker experimenter (Spanish/German) was present to answer questions. The instructions explicitly aimed at eliciting complete sentences, referring to dynamic events, rather than descriptions of individual elements / static properties of the scenes. A trial started with a centred fixation cross (1000ms), after which a picture appeared randomly in one of the four quadrants of the screen. This was done (1) to force participants to execute a fixation, and (2) to prevent strategies related to the predictability of a stimulus' location. The number of pictures showing agent left-action right, and agent right-action left was counterbalanced across subjects and order of presentation was randomized within subjects. Stimulus presentation duration was manipulated between subjects. After picture offset, a blank screen was shown for 9000 ms, providing sufficient time for producing utterances. Speech was audiorecorded with an external microphone.

Data treatment and Results

We analyzed first fixations[1] on the pictures for three locations (AoIs) separately: Agent (head and upper part of agent's body, white oval), Action (hands and object, white oval) and 'In-between' (part of agent's body and closely surrounding it, dark-grey area, Figure 1). We included the in-between area because there are no theoretical grounds to assume that first fixations that are placed neither in the agent nor in the action AoI are arbitrary or inaccurate. In fact, we consider the location of the first fixation as the outcome of specific calculations made by the cognitive system while apprehending.

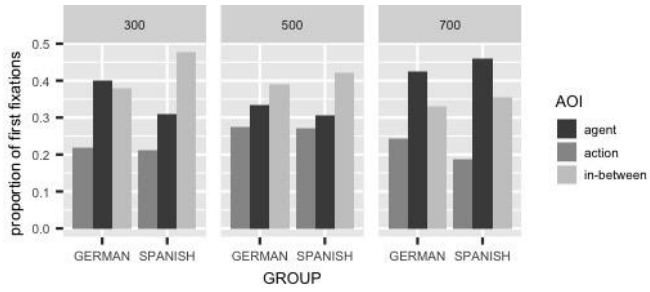

Figure 2: Overall proportions of first fixations in the AoIs, by presentation duration and language

Figure 2 shows the overall proportions of first fixations in each AoI within each group, per condition (300ms, 500ms, 700ms). First fixation locations (binary coded for each AOI separately; 1=hit, 0=no hit) were analysed with linear mixed effects regression models, including the fixed factors Language (sum coded) and Condition (treatment coded) and random intercepts for subjects and items (pictures; both versions were collapsed).

		Estimate	Std. Error	z value	Pr(>|z|)	
1	(Intercept)[2]	-0.633	0.201	-3.156	0.002	
2	Ger vs Spa	-0.212	0.164	-1.288	0.198	
3	300 vs 500	-0.171	0.185	-0.924	0.356	
4	300 vs 700	0.449	0.223	2.014	0.044	*
6	(Intercept)[3]	-0.803	0.201	-3.997	0.000	
7	500 vs 700	0.619	0.223	2.780	0.005	**
8	Ger vs SPA 300	-0.490	0.2539	-1.925	0.054	.
9	Ger vs SPA 500	-0.141	0.2531	-0.558	0.577	
10	Ger vs SPA 700	0.180	0.3491	0.516	0.606	

Table 1: 'Agent' first fixations; reference level is indicated by the first term in each line

With respect to the agent AoI, our analyses revealed an effect of condition (presentation duration): There were more first looks in the agent AoI in the 700ms condition compared to both other conditions, but there was no difference between 300ms and 500ms (Table 1, lines 3, 4, and 7). There was no effect of Language. Comparing the 3

[1] A first fixation was defined as the event that followed the first saccade after stimulus onset, as registered by the eye tracker. On each trial, the initial position of the participants' gaze was in the center of the screen due to a centred fixation cross being presented before each stimulus.

[2] Model includes both main predictors.

[3] Same model specifications as in lines 1-4 but with reference level changed to 500ms.

presentation duration conditions between languages, the model detected a small effect: In the 300ms condition, Spanish participants displayed fewer first looks to the agent than German participants (Table 1, line 8).

With respect to the action AoI, our analyses showed no effects (models are not reported).

Regarding first fixations in the 'in-between' area, there were significantly fewer of those in the 700ms compared to the 300ms condition (Table 2, line 5). A main effect of Language was detected: Spanish speakers placed significantly more first fixations in the in-between area than German speakers (Table 2, line 2). Comparing the 3 conditions between languages, the model showed a significant effect in the 300ms condition, with Spanish participants fixating the 'in-between' AoI more than German participants (Table 2, line 8).

		Estimate	Std. Error	z value	Pr(>|z|)	
1	(Intercept)	-0.317	0.130	-2.427	0.015	
2	Ger vs Spa	-0.132	0.067	-1.987	0.047	*
4	300 vs 500	-0.112	0.149	-0.753	0.452	
5	300 vs 700	-0.396	0.183	-2.166	0.030	*
6	(Intercept)	-0.428	0.130	-3.288	0.001	
7	500 vs 700	-0.284	0.183	-1.555	0.120	
8	Ger vs SPA 300	-0.226	0.104	-2.179	0.029	*
9	Ger vs SPA 500	-0.079	0.103	-0.762	0.446	
10	Ger vs SPA 700	-0.052	0.146	-0.355	0.722	

Table 2: 'In-between' first fixations; reference level is indicated by the first term in each line

First Fixation latencies

To test whether participants modulate the time for executing their first fixation depending on stimulus presentation duration, log-transformed first fixation latencies were analyzed using mixed effects linear regression models with Language and Condition as fixed factors and random intercepts for subjects and items. Results show that first fixation latencies are significantly smaller in the 300ms condition compared to the 500ms condition (Est.=0.050, SE=0.025, t=1.96, p<.05) and marginally smaller compared to the 700ms condition (Est.=0.060, SE=0.031, t=*1*.95, p=.05). There was no significant effect of Language, nor any interaction effects.

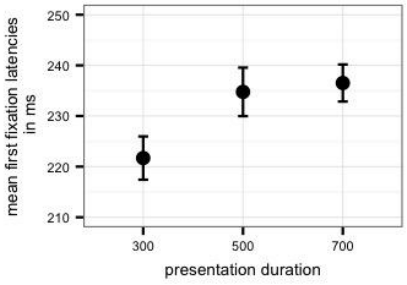

Figure 3: Mean first fixation latencies

Language data: specificity of event descriptions

The transcribed utterances were coded for specificity with respect to agent, action, and object features. Agent reference was coded as 'specific' when the subject contained a noun (phrase) with information on the agent's gender (*[there is / I see] a/the woman*), and as 'unspecific' when the sentence contained a gender-neutral noun (*a person/someone*), a pronoun or null subject. Action-specific sentences contained a specific action verb (*to draw*); the action was coded as unspecific in case of general action verbs (*to do/hold something*) or state-verbs (*to sit at a table*). Objects were coded as specific when referring to concrete items (*a sandwich/a bottle*); other object references were coded as unspecific (*something [big]*) (all specificity factors were binary coded). Mixed effect models were used to estimate the impact of the factors Language and Condition.

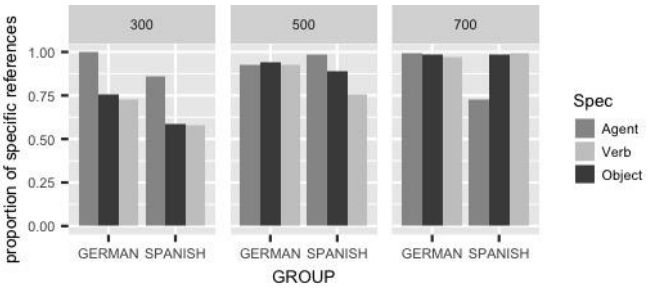

Figure 4: Overall proportions of specific encodings of agent, verb, object, by presentation duration and language

Agent specificity

There were no language or condition effects (nor interactions) for agent specificity (model is not reported). Note that in some cases Spanish participants omitted overt reference to the agent, which is licensed in specific contexts (pro-drop). It occurred mainly in the 700ms condition and these instances are coded as unspecific agent references.

		Estimate	Std. Error	z value	Pr(>|z|)	
1	(Intercept)	0.855	0.378	2.263	0.024	
2	Ger vs Spa	0.418	0.241	1.732	0.083	.
3	300 vs 500	1.562	0.522	2.992	0.003	**
4	300 vs 700	3.963	0.745	5.320	0.000	***
5	(Intercept)	2.417	0.393	6.147	0.000	
6	500 vs 700	2.403	0.746	3.219	0.001	**
7	Ger vs SPA 300	0.385	0.350	1.099	0.272	
8	Ger vs SPA 500	0.790	0.364	2.167	0.030	*
9	Ger vs SPA 700	-0.627	0.662	-0.948	0.343	

Table 3: Verb specificity; reference level is indicated by the first term in each line

Verb and object specificity

Analyses showed an effect of Condition: The longer the stimulus was displayed on the screen, the more specific action verbs were produced (Table 3, lines 4 and 5). There was also a significant Condition by Language interaction: In the 500ms condition, Spanish speakers produced fewer specific action verbs than German speakers (Table 3, line 8).

With respect to object encoding there were significant effects of Condition (Table 4, lines 3, 4, and 6), with object specificity increasing with longer exposure time. There was also an effect of language: Overall, Spanish speakers produced fewer specific object references. Most pronounced is the effect when comparing Spanish and German speakers within the 300ms condition (Table 4, line 7).

		Estimate	Std. Error	z value	Pr(>|z|)
1	(Intercept)	0.830	0.257	3.232	0.000
2	Ger vs Spa	0.406	0.164	2.475	0.013*
3	300 vs 500	2.063	0.354	5.834	0.000***
4	300 vs 700	3.862	0.554	6.908	0.000***
5	(Intercept)	2.893	0.288	10.056	0.000
6	500 vs 700	1.763	0.557	3.167	0.002**
7	Ger vs SPA 300	0.515	0.232	2.215	0.027*
8	Ger vs SPA 500	0.391	0.258	1.513	0.130
9	Ger vs SPA 700	-0.027	0.488	-0.056	0.955

Table 4: Object specificity; reference level is indicated by the first term in each line

General Discussion

The main findings of Experiment 2 are the following: (1) Presentation duration modulates the specificity of the event descriptions: Reference to actions and objects is more precise given longer viewing times; (2) presentation duration affects the timing of parafoveal and foveal information uptake, i.e., given a short viewing time participants move the eyes faster towards a location from where acute vision is possible; (3) presentation duration has an effect on the location of first looks: There were generally fewer first looks in the in-between region and more looks to the agent region in the longest presentation duration condition; (4) the language of the participants exerted an influence on first fixation locations.

The finding that verbal responses were more specific regarding actions and objects given longer presentation durations, but at the same time there was no increase with respect to first fixations registered in the corresponding Area of Interest (Action region), suggests that the location of the first fixation does not directly relate to the process of message construction, i.e., linguistic processing. Furthermore, the location of first fixations does not correlate with the sentence format produced (only approximately 40% of all first looks were registered in the agent region, but sentences were exclusively SV(O)). We thus take both findings to show that our measure indeed reflects processing *prior* to linguistic planning, i.e., scene apprehension.

To explain the effect of presentation duration on the location of the first fixation, three things must be taken into account. First, information extraction is better the nearer the eye rests on the area which contains critical information, because viewing acuity asymptotically decreases from the foveal region (Holmqvist et al. 2011). Second, the time available for information extraction *after* the first fixation is placed on the stimulus (acute vision), differs between all three conditions. In the 300ms condition participants have approximately 80ms for further visual processing after the eye has arrived at the landing site and before the stimulus disappears, whereas this time is already more than three times longer in the 500ms condition, and almost 6 times longer in the 700ms condition. Third, the time before the first fixation is launched also differs between the presentation durations, as is evident from the differences in first fixation latencies. In the 300ms condition participants remained significantly shorter at the fixation cross before directing focal attention onto the stimulus, compared to the 500ms or 700ms condition. The location of the first fixation thus reflects what type of visual information the processing system demands with the highest acuity possible, given the time available. Since first fixation patterns differ most clearly between the 300ms and the 700ms condition (generally more first looks to agents, and fewer to the in-between region in the longer presentation duration), the strongest predictor for the location of the first fixation is the probability of a following fixation: In the longest presentation duration condition participants nearly in all cases had enough time to place a total of two fixations on the stimulus before it disappeared. Calculating the location and timing the execution of a first fixation thus draws on the ability of the cognitive system to predict the approximate future time course of the processes relevant for task completion, in our study the task of constructing a message on the basis of the information available. Given these results, we can conclude that apprehension is a flexible process.

Interestingly, we did obtain a difference with respect to first fixation *latencies* but not (generally) first fixation *location* when comparing the 300ms and the 500ms condition. Given that verbal responses were more precise in the 500ms condition we can conclude that the landing site of the first fixation in the 300ms condition was already "good enough" to fulfill the task, even with this short exposure time (i.e., participants were already able to provide information on both the agent and the action in the 300ms condition). In future research it thus seems reasonable to explicitly analyze the interrelation between first fixation locations and first fixation durations.

The analysis of event descriptions revealed that the representation which is ultimately encoded by verbal means, i.e., the preverbal message, is not as "rich" in the shortest condition as it is in both longer conditions – participants produced fewer specific action verbs and object references in the 300ms condition. This implies that less specific information was passed on from the representation constructed during apprehension to the message generator. Put the other way around, a message was constructed on the basis of whatever information was available. Since messages are thought to be composed of word meanings (Levelt et al. 1999) the quality of information available from event apprehension directly influences what word meanings can be activated. This is especially evident from the main effect of presentation duration on verb specificity.

Given the role of the factor language in our experiment we may infer that during scene encoding speakers of different languages focus on different things at different times. Under time pressure, mainly Spanish speakers resort to fixating the area in between the agent and the action first, and this seems to be sufficient for retrieving information concerning both event elements, as there is no difference in encoding specificity with German speakers. German speakers, in turn, fixate agents first more frequently in the 300ms condition. In sum, language seems to matter but not in a straightforward way: Given limited exposure time, we

have identified that German and Spanish speakers have different *starting points* for their linguistic encoding processes. Nonetheless the *outcome* seems to be the same, in the present task setting. Previous research has put forward the hypothesis that in general, the Spanish language may be less agent-oriented than German or English. Our data can be interpreted in this way, but the effect mainly appeared under increased task demands: Spanish speakers showed fewer agent-first fixations than Germans in the 300ms condition. Note that this lack of a prominent agent-first strategy in Spanish is not directly linked to the fact that the language allows subject-drop; there were only few such instances in the data, limited to the longest presentation duration condition. We can conclude that the current experimental design does not provide an appropriate pro-drop licensing context in Spanish. At this point, further research is required, introducing, e.g., pro-drop biasing context (e.g., a discourse context), or more variation regarding agents and event-types (i.e., ditransitive events; controlling for animacy, another relevant factor) in the experimental stimuli. The use of different language contrasts would also shed more light on whether language background of the speaker structurally affects scene apprehension.

On a similarly critical note, given the between-subjects manipulation of stimulus presentation duration, we cannot fully exclude that participants developed some strategy for solving the task in the specific condition put to them. Importantly, however, for our participants it was impossible to predict the locations of the agent and the action on the screen, given left-right counterbalancing of agents and actions within the pictures, and the random placement of pictures on the screen. Regardless, even if explicit strategy-formation played a role of some sort, the data still show effects of task demands and language background, evidencing top-down influences on scene apprehension.

Conclusions

We measured the location of first fixations and their latencies to gain insight into the flexibility of the event apprehension process. We find top-down effects: For the first time it was shown that the time available for visual processing (presentation duration manipulation: 300ms-500ms-700ms) directly affects *when* and *where* people move their eyes first. However, first fixation locations and latencies did not directly relate to the contents (i.e., the level of specificity) of our participant's verbal responses. We thus argue that first fixations are driven by apprehension (initial gist extraction) and not message generation processes, which underlines that both processes are distinct in nature. Furthermore, we obtained small, but measurable effects of the language spoken by our participants, but mainly when time pressure on information uptake was strongest. These differences in first fixations, however, were not directly reflected in the information provided in the utterances produced. Further research is required to address language background as a factor affecting scene apprehension.

Acknowledgements

This study was made possible by a grant from the Ibero-Amerika-Zentrum (IAZ), Heidelberg University, to both authors. We thank Lina Zambrano for help with data collection and Kay Bock for discussing preliminary results with us.

References

Bock, K., Irwin, D., Davidson, D., & Levelt, W. (2003). Minding the clock. *Journal of Memory and Language, 48*(4), 653-685.

Brown-Schmidt, S., & Konopka, A. (2008). Little houses and casas pequeñas: message formulation and syntactic form in unscripted speech with speakers of English and Spanish. *Cognition, 109(2),* 274-280.

Dobel, C., Gumnior, H., Bölte, J. & Zwitserlood, P. (2007), Describing scenes hardly seen. *Acta Psychologica, 125(2),* 129-143.

Fausey, C., &. Boroditsky. L. (2011). Who dunnit? Cross-linguistic differences in eye-witness memory. *Psychonomic Bulletin & Review, 18(1),* 150-157.

Flecken, M., Gerwien, J., Carroll, M., & v. Stutterheim, C. (2015). Analyzing gaze allocation during language planning: a cross-linguistic study on dynamic events. *Language and Cognition, 7(1),* 138-166.

Griffin, Z., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science, 11(4),* 274-279.

Henderson, J., & Ferreira, F. (2004). Scene perception for psycholinguists. In J. Henderson & F. Ferreira (Eds), *The Interface of Language, Vision, and Action*, pp. 1-58. New York: Psychology Press.

Holmqvist, K., Nystrцm, M., Andersson, R. Dewhurst, R. Jarodzka, H., & Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. OUP.

Levelt, W., Roelofs, A., & Meyer, A. (1999). A theory of lexical access in speech production. *Behavioral and brain sciences, 22(1),* 1-38.

Masuda, T., & R. Nisbett (2006). Culture and change blindness. *Cognitive Science, 30(2),* 381-399.

Sauppe, S., Norcliffe, E., Konopka, A., Van Valin, R., & Levinson, S. (2013). Dependencies first: eye tracking evidence from sentence production in Tagalog. In M. Knauff et al. (Eds.), *Proceedings of the 35th Annual Meeting of CogSci* (pp. 1265−1270). Austin, TX: Cognitive Science Society .

v. Stutterheim, C., Bouhous, A. & Gerwien, J. (2016). The impact of aspectual categories on the construal of motion events. The case of Tunisian and Modern Standard Arabic. Poster presented at EXAL+, NYU Abu Dhabi.

Williams, P., & Simons, D. J. (2000). Detecting changes in novel, complex three-dimensional objects. *Visual Cognition, 7,* 297–322.