Analyzing gaze allocation during language planning: a cross-linguistic study on dynamic events[1]

MONIQUE FLECKEN, JOHANNES GERWIEN, MARY CARROLL and CHRISTIANE VON STUTTERHEIM

Language and Cognition / Volume 7 / Issue 01 / March 2015, pp 138 - 166
DOI: 10.1017/langcog.2014.20, Published online: 06 June 2014

Language and Cognition **7** (2015), 138–166. doi:10.1017/langcog.2014.20

Analyzing gaze allocation during language planning: a cross-linguistic study on dynamic events[1]

MONIQUE FLECKEN

Donders Centre for Cognition, Radboud University Nijmegen, the Netherlands, and Institut für Deutsch als Fremdsprachenphilologie, Heidelberg University, Germany

JOHANNES GERWIEN

MARY CARROLL

AND

CHRISTIANE VON STUTTERHEIM

Institut für Deutsch als Fremdsprachenphilologie, Heidelberg University, Germany

(*Received* 21 *May* 2013 – *Accepted* 10 *March* 2014 – *First published online* 6 *June* 2014)

ABSTRACT

Studies on gaze allocation during sentence production have recently begun to implement cross-linguistic analyses in the investigation of visual and linguistic processing. The underlying assumption is that the aspects of a scene that attract attention prior to articulation are, in part, linked to the specific linguistic system and means used for expression. The present study concerns naturalistic, dynamic scenes (video clips) showing causative events (agent acting on an object) and exploits grammatical differences in the domain of verbal aspect, and the way in which the status of an event (a specific vs. habitual instance of an event) is encoded in English and German. Fixations in agent and action areas of interest were timelocked to utterance onset, and we focused on the pre-articulatory time span to shed light on sentence planning processes, involving message generation and scene conceptualization.

[1] The execution of this study was supported by a DFG (German Research Foundation) grant (STU-131/8-2). M. Flecken is currently supported by an NWO *Veni* grant (275-89-011). We would like to thank both funding institutions for financial support. Address for correspondence: Monique Flecken, Donders Centre for Cognition, Radboud University, Montessorilaan 3, 6525 HR Nijmegen, The Netherlands. e-mail: m.flecken@donders.ru.nl

Findings are threefold: (i) English speakers mark the status of an event as specific in relation to the action, with progressive aspect marking on the verb in each utterance. German speakers do so by elaborating specific characteristics of the agent; (ii) participants display significantly different gaze allocation patterns to agent and action regions although the sentences produced in both languages follow the same subject–verb word order; and (iii) the analysis of gaze patterns during sentence production given dynamic scenes provide complementary results from a more naturalistic paradigm, to those obtained in studies with still images.

KEYWORDS: language production, cross-linguistic analysis, eye-tracking, dynamic stimuli, event construal, grammatical aspect.

1. Introduction

Cross-linguistic variation in language production patterns has been studied extensively using a range of different elicitation stimuli (e.g., pictures, films of single events or chains of events; see Berman & Slobin, 1994; Soroli & Hickmann, 2010; Strömqvist & Verhoeven, 2004). Studies have focused on what information, presented in a visual stimulus, speakers of different languages select for verbalization, and how it is organized and structured in sentences or longer texts. Decisions made during language planning (processes of information selection and structuring; cf. Bock & Levelt, 1994; Griffin & Bock, 2000; Levelt, 1989) have been shown to relate to the types of lexical and grammatical structure which the language of the speakers in question offer ('thinking for speaking'; cf. Slobin, 1996; v. Stutterheim & Nüse, 2003). However, it is not clear at which point in the production process language-specific requirements are taken into account. The present study focuses on potential cross-linguistic differences in attention allocation during the time span of sentence planning, from the presentation of a visual stimulus leading up to utterance onset. Analyses of gaze allocation in event description show that there is an initial phase in which speakers extract the gist of an event (e.g., Griffin & Bock, 2000; Griffin & Spieler, 2006), i.e., the identification of the event type in general terms. Furthermore, the phase between stimulus onset and utterance onset includes, besides linguistic encoding processes related to the first mentioned sentence element, processes of 'message generation' and conceptualization (cf. Levelt, 1989), the generation of a conceptual representation of the event, specifying sentence meaning and 'what to say', which precedes the encoding of the first word of an utterance (cf. Bock & Levelt, 1994). It is still an open question to what extent the specific language used or the specific linguistic structures planned affect processing already during the phase of conceptualization.

Here, we address this question, looking at effects of grammar on patterns of visual attention during a pre-articulatory time window, thus including processes of message generation related to the domain of events.[2] We focus on the grammatical category of verbal aspect and its implications for conceptualization. Broadly speaking, verbal aspect encodes a particular temporal viewing point or perspective in relation to a situation (Comrie, 1976; Klein, 1994), with a basic distinction between imperfective/progressive – presenting a situation as ongoing and unbounded – and perfective, presenting a situation as bounded. Languages differ in how they encode aspect. Aspect can be grammaticalized, expressed via morphological marking on the verb, as in English (the focus of the present study), many Slavic languages, or Arabic, or it can be optionally lexically encoded by adverbs or particles as in German (the focus of the present study). Grammatical markers of aspect need to be expressed obligatorily in specific contexts, which has implications for the cognitive salience of the corresponding temporal concepts (e.g., ongoingness, boundedness) and of the visual features of stimuli associated with these concepts (see, for the notion of 'saliency', Slobin, 1996; for a specific view of the relation between grammar and cognition, Lucy, 1992). Given that aspect contributes to the construction of sentence meaning (the 'message') by conveying an explicit perspective on a situation, we assume its presence or absence in a language system to already play a role during conceptualization processes in production.

Besides contributing a temporal perspective, aspect also has implications for the modal interpretation of a sentence. Whenever a speaker describes an event, its modal status will be part of the linguistic description: when a person is asked to describe "What is happening?" in relation to the contents of a scene (showing a woman baking cupcakes, for example), the linguistic output produced needs to be a finite sentence, referring to a specific event. Propositional content such as [bake, a woman, cupcakes] is transformed into an assertion, i.e., an interpretable linguistic unit, by anchoring it with respect to a referential frame of times, spaces, and worlds. A basic distinction that is made at this level is the one between reference to a specific event, i.e., a singular occurrence of a situation, and a habitual or generic reference to an event of the same type. An aspect-language such as English encodes this distinction by means of grammatical aspect marking on the verb (the progressive *to be* V-*ing*, for example, *the woman is baking cupcakes*, specific

[2] We use the term 'event' as denoting a dynamic situation in which a change of state or change of place takes place. An event therefore always has an internal temporal structure (see, for a comprehensive discussion of the notion 'event', Tenny & Pustejovsky, 2000).

event, vs. *the woman bakes cupcakes,* generic or habitual statement). The function of grammatical aspect is therefore not only to highlight and specify the temporal contours of an event, but also to convey the modal status of an event (Dahl, 1995; v. Heusinger, 2002; Klein, 1994; Rijkhoff & Seibt, 2005; v. Stutterheim, Carroll, & Klein, 2009). A non-aspect language like German does not provide such grammaticalized verbal means. *Die Frau bäckt Kuchen* 'the woman bakes cupcakes' is ambiguous with respect to a specific or unspecific interpretation. German speakers show a tendency to convey the status of an event described as specific by giving more detailed information on either the entities involved (agents, objects, instruments) or the referential frame of a situation (its location, for example) (see v. Beek, Flecken, & Starren, 2013; Carroll & v. Stutterheim, 2011) (e.g., *Die (ältere) Frau bäckt Kuchen (in der Küche)* 'the older woman bakes cakes in the kitchen'). This means that in sentence production, besides its relevance for the conceptualization of action-features of an event (i.e., its temporal contours) and thus the linguistic encoding of the verb in a sentence ('simple' verb or aspectually marked verb), the fact that one speaks an aspect or non-aspect language may also affect how one extracts information relevant for other components of a sentence (e.g., the conceptualization and description of entities), and how events in general are visually processed and conceptualized.

2. Background

2.1. GAZE ALLOCATION IN SENTENCE PRODUCTION

In general, studies on language production using eye-tracking have largely focused on the naming of individual objects (i.e., the production of words or simple noun phrases), with a relatively small number of studies on the production of longer sequences relating to object arrays, or static depictions of scenes (see, e.g., Bock, Irwin, & Davidson, 2004; Brown-Schmidt & Tanenhaus, 2006; Griffin & Bock, 2000; Huettig, Rommers, & Meyer, 2011; Meyer, 2004; Meyer & Dobel, 2003). Findings point to a tight timelock between looking and speaking, described as the eye–voice span, a measure which has been taken as a starting point in numerous follow-up studies (e.g., Kuchinsky, Bock, & Irwin, 2011). This relation was also found in paradigms that allow variation in the type of constructions used to describe a scene (Bock, Irwin, Davidson, & Levelt, 2003; Griffin & Bock, 2000; Meyer, Sleiderink, & Levelt, 1998) (note that in all cases static stimuli, i.e., pictures, were used to elicit descriptions of events). The first studies that took into account the effect of word order in different sentence structures on gaze patterns were carried out by Griffin and Bock (2000). The authors examined eye-movement to parts of still pictures depicting events, before and during

the time speakers were describing them with either an active or a passive sentence. The order in which the two participants in the event were fixated matched the order of mention in the active or passive sentences, following an initial event apprehension phase in which the general gist of the scene was extracted (lasting for approximately 300 ms from stimulus onset; see also, e.g., Bock et al., 2003). Griffin and Bock (2000) thus concluded that eye-movements can predict the order of mention of elements in a sentence. Similarly, van der Meulen, Meyer, and Levelt (2001) found a tight serial order between object gazes and object naming. Gleitman, January, Nappa, and Trueswell (2007) also looked at eye-movement during the description of event scenes (pictures). Interestingly, they showed that the early phases of event construal correlated with specific word order variations in sentences, and they discussed, besides perceptual factors, certain conceptual and linguistic factors that can 'control' sentence production. Meyer and Dobel (2003) showed that in event description tasks speakers tended to fixate action-relevant regions in still pictures early on (e.g., the hands of the characters involved in a 'giving' event). This was interpreted as necessary for the extraction of information relevant for verb selection, regardless of word order in the sentences produced (see also Dobel, Glanemann, Kreysa, Zwitserlood, & Eisenbeiss, 2010). Significantly, word order in the language of the speakers tested (Dutch) is not verb-initial, but subject-initial.

Sauppe, Norcliffe, Konopka, Van Valin, and Levinson (2013) investigated event descriptions by speakers of Tagalog, a verb-initial language, which encodes agreement on the transitive verb with either the agent or the undergoer of an action, depending on the assignment of the role of 'privileged semantic argument' (PSA) to either of these participants. Interesting for the present paper is the following fact: the two noun phrases (NPs) referring to actor and undergoer do not have syntactically fixed order at sentence level. However, speakers have to make an early syntactic decision on which participant to select as PSA. This thus allows the disentanglement of planning processes related to participants' internal dependencies (PSA and respective verb agreement), and planning processes related to the sequential order of mention of participants in sentences. Attention allocation was measured in the time span before utterance onset. Findings showed that the participant who was selected as PSA was fixated more frequently during this phase, independent of order of mention. The authors concluded that there are two phases in language planning, an early phase in which grammatical constraints are integrated in a message, which is generated independent of the actual linear order of elements, and a later phase in which the elements are prepared for encoding, which follows the order of mention as encoded in a sentence.

A study by Coco and Keller (2010) showed that the situational context of an event affects attention allocation to a significant extent. Participants had to describe a scene for which they were given a verbal prompt. This cue word was ambiguous with respect to two referents depicted in the scene. Fixations patterns showed that ambiguity resolution was one of the factors affecting sentence planning and the accompanying fixation patterns: If a to-be-mentioned referent had a competitor in the visual scene, competition between visual referents seemed to override the standard eye–voice span effect. The authors concluded that "the simple view according to which referents are fixated in the order in which they are mentioned with a fixed eye–voice span between fixation and mention, does not seem to generalize to more realistic settings" (Coco & Keller, 2010, p. 1075).

There is thus evidence that pre-articulatory gaze allocation patterns follow complex conceptual principles, to some extent involving grammatical requirements, and they do not only reflect linguistic encoding processes (e.g., form retrieval) that follow surface sentence form and structure (e.g., word order).

2.2. CROSS-LINGUISTIC CONTRASTS AS A TOOL IN THE ANALYSIS OF GAZE ALLOCATION IN SENTENCE PRODUCTION

Cross-linguistic studies provide a particularly interesting basis for insights into patterns in attention allocation and their link to phases in language planning. Such studies show how attention patterns may vary in line with different linguistic structures used by speakers of different languages, and during what phases of the timecourse of the production process differences arise (see general overview in Brown-Schmidt & Konopka, 2008; Jaeger & Norcliffe, 2009; Papafragou, Hulbert, & Trueswell, 2008; Sauppe et al. 2013; Soroli & Hickmann, 2010; v. Stutterheim, Andermann, Carroll, Flecken, & Schmiedtová, 2012). Papafragou et al. (2008) investigated how English and Greek speakers scanned and described motion events (short animations showing different manner of motions, e.g., skate, run to a tunnel). Greek is a path-language, encoding information on the path of a motion event mainly in the verb (and manner can be optionally expressed in sentences), whereas English encodes manner information in the verb and path information in particles or adjuncts in sentences (cf. Talmy, 1985). Different degrees of manner salience were hypothesized to be reflected in visual attention to, and the linguistic encoding of, manner. Already shortly after stimulus onset, English participants allocated more attention to the manner of motion. Differences in the type of information encoded in motion verbs thus affected EARLY on-line scene processing and attention allocation to two sources of motion-relevant information. Both were attended to by speakers of both languages, but language affected when,

and the intensity with which they were inspected, causing an early manner focus in English.

Another set of empirical studies looked at the role of grammatical aspect for gaze allocation in event description tasks, comparing aspect and non-aspect languages. The extent to which aspect affects description and gaze patterns in relation to motion events was investigated for the aspect-languages Standard Arabic, Russian, and English, in contrast with the non-aspect languages German and Dutch (v. Stutterheim et al., 2012). Speakers of aspect-languages showed fewer linguistic encodings of endpoints of motion event scenes (e.g., a video clip showing a vehicle travelling along a road leading to a town in the distance), and this correlated with a lower frequency and a shorter duration of fixations on endpoints (e.g., the town in the distance). In the non-aspect languages, higher fixation rates correlated with higher frequencies of endpoint encoding. The authors concluded that 'seeing for speaking' patterns differed in correlation with the grammatical means available (v. Stutterheim et al., 2012). Another paper studied the description of causative events in Dutch (Flecken, 2011). Dutch speakers, who described the events using progressive aspect, fixated action-regions of the scenes longer and more frequently during the entire period of stimulus display than Dutch speakers, who did not use aspect.

In short, language-specific lexical and grammatical features (which may have scope over larger linguistic units) play an important role already during the generation of the 'message', and the conceptualization of an event, and this can be reflected in patterns of gaze allocation during (early phases of) scene processing.

2.3. GRAMMATICAL ASPECT AND EVENT COMPREHENSION

Empirical studies have investigated whether progressive aspect, when encoded in a sentence, influences how listeners or readers perceive and interpret the event described. The basic hypothesis is that, given an aspectual perspective, action-specific features and temporal contours of actions should be highlighted. Anderson, Matlock, and Spivey (2013), for example, using a computer-mouse tracking paradigm, showed that aspect in English systematically shaped perceptual simulations of events, in the sense that descriptions of events with the past progressive displayed an inherent match with a temporal context encoding a recent past; this match between conditions elicited smoother and faster mouse movements. Matlock (2010) showed that the use of progressive aspect in an event description caused comprehenders to conceptualize more action in a given time span, compared to sentences unmarked for aspect. In a slightly different vein, Flecken and Gerwien (2013) found that progressive aspect affects people's perception of the duration of everyday events and actions. Also, from an embodied perspective, Bergen and Wheeler (2010) showed

that the progressive leads people to mentally simulate the core, intermediate, phases of an action. These studies thus show specific psycholinguistically real processing patterns associated with grammatical aspect, in particular in relation to action-features of events.

3. The present study

The present study used naturalistic dynamic scenes in gaining insight into the conceptualization and encoding of events, in real-life settings.[3] Native speakers of English (aspect language) and German (non-aspect language) were shown a series of short video clips and instructed to describe what was happening in each video on-line, in one sentence. Speakers' utterances were audio-recorded, and eye-movement was recorded with an eye-tracker during stimulus display (see details below). We analyzed the contents and structure of the event descriptions and the accompanying pre-articulatory fixation patterns (fixation duration and frequency) in relation to two Areas of Interest (AoIs) in the stimuli. The type of event studied involved causative events in which an agent performs an action on a specific object (e.g., a person knitting a scarf, a person folding a paper airplane). This event type allowed the investigation of potential differences in the distribution of attention allocated to the agent and the action in the events. The order of mention of these components is identical in German and English finite declarative main clauses; word order is SVO, thus referring to the agent first, the action second, and the object last (a person (Subject) is doing X (Verb) to/with Y (Object)).[4] In order to conceptualize and describe causative events, speakers of each language thus have to attend to both agent and action features, given the assumption that overt fixations are, at least to some extent, necessary for message generation, and, to a large extent, for formulation processes (i.e., form retrieval) of individual sentence elements (cf. Griffin, 2004).

---

[3] Animated scenes in which components may be adapted in size (unusually big or small), or where the timecourse contradicts real-world experience in subtle ways, may disrupt highly automated systems and trigger processes that over-ride established patterns (cf. methodological considerations in Coco & Keller, 2010; see also Jaeger et al., 2012, for studies on natural speech). In this sense, static scenes depicting only one screenshot or frame of a dynamic scene may be less likely to activate processes involved in event construal. This assumption is based on the fact that the key factor distinguishing states from events – dynamicity – is not immediately evident with static stimuli, but has to be inferred.

[4] Other options, which are grammatically possible, but highly marked, are contextually not supported in the given task. Object fronting in German, *Eine Kette fädelt ein Mädchen auf* 'a necklace (object) is beading a girl (subject)', or the use of a passive, *Eine Kette wird aufgefädelt* 'a necklace is getting beaded', did not occur in any of the utterances produced.

The relevant cross-linguistic contrast lies in the domain of grammatical aspect, and the way in which the modal status of an event is encoded in finite sentences: English encodes progressive aspect grammatically on the verb, whereas German does not have grammatical means to encode aspectual distinctions. In order to fulfil the task (describe 'what is happening' in each video, on-line), the utterances produced needed to be finite and specific. In English, both criteria can be met via morphological markings on the verb, encoding, amongst other things, tense (making descriptions finite) and aspect. Aspect marks an explicit perspective on the state of affairs, and temporally anchors the event as a specific case (*The boy plays football* vs. *the boys are playing football*, the latter expressing a specific instance of a football-playing event, ongoing at present, rather than a generic statement about the boy's activities). In German, only the former criterion (finiteness) is met by verb morphology. Speakers may convey specificity of the event on any part of the description, outside the verb, to disambiguate between generic and specific interpretations of the utterance. One of the options for marking specificity is the linguistic expression of the agent of the causative action.[5] The analysis of the pre-articulatory distribution of gaze to agent versus action regions in the scenes allows a specific investigation of the encoding of the agent, given that the agent will be the first mentioned element of the event. We are thus able to look at the timecourse of gaze allocation to the agent versus gaze allocation to the action region during this time window.

Our hypotheses with respect to description and fixation patterns were the following: English speakers use progressive aspect; they thus mark the status of the events described as specific in relation to the action, by taking this specific temporal perspective on the situation. German is a non-aspect language and the specificity of the event may be encoded in elements of the description other than the verb; speakers may, for example, elaborate on specific details of the agents in the scenes (the other important aspect of the event depicted). In the time window analyzed, we expected German speakers to attend more to agents than English speakers, as this event element could be relevant for conveying the event status as specific, and (at least) the planning and encoding of the subject noun phrase needs to be finished prior to utterance onset. English speakers were hypothesized to show predominant allocation of attention to the action, and a smaller interest in the agent, given that the agent is less relevant for marking event status in English.

Potential gaze allocation differences could, in principle, be due either to differences in surface form of produced sentence elements, such as the

[5] Previous studies show that specificity can also be marked by highlighting instruments of actions, or locations at which events take place, for different event types (v. Beek et al., 2013; Carroll & v. Stutterheim, 2011).

first mentioned sentence element (agent) which should be detectable in roughly the second before speech onset time (SOT), or to more global processes of conceptualization of the event as a whole and the generation of the 'message' of the planned utterance, which should arise early in the timecourse before SOT. In order to tease apart these two types of language effects, besides an analysis of overall looking times in both AoIs, fixation patterns over the whole time span leading up to utterance onset were analyzed, investigating explicitly WHEN along the timecourse differences emerge.

4. Experiment

4.1. PARTICIPANTS

Two groups of German and English native speakers took part in the experiment ($N = 19$ in each group), with comparable socio-cultural backgrounds (students and postgraduates), aged between twenty and thirty-five. Numbers were balanced for gender, and the participants had normal or corrected-to-normal vision. Data collection was carried out in the eye-tracking laboratory at the institute for German as a foreign language philology, Heidelberg University. Participants were given a questionnaire on their social and linguistic background. They were excluded from the analysis if they listed a bilingual background or residence for more than three months in a country where a language other than their mother tongue was spoken. German participants were local students, and English speakers were participants at a summer school at Heidelberg University, all with very little knowledge of German. They were recruited for the experiment during the first five days of their stay in Germany so that they would be as monolingual as possible with hardly any knowledge of German. All participants were paid for participation.

4.2. STIMULI

The analyses were conducted on the basis of six dynamic video clips, embedded in a total of sixty video clips, each of six seconds in length, which were presented in pseudo-randomized order. Each of the six video clips depicted one event in which an agent was seen performing an ongoing action on an object without rapid movements or interruptions:

1. a woman beading a necklace;
2. a man folding a paper airplane;
3. a man drawing a tree with a pencil;
4. a woman knitting a scarf;
5. a woman decorating a cake with cream;
6. a woman building a tower with blocks.

All items represented causative events with a comparable spatial distribution of the areas of interest. In all video clips, two non-overlapping spatial regions could be identified: the upper side of the body of the agent in one region, and in the other the action involving the affected/effected object and the hands of the agent while engaged in the activity (see Figure 1 below).

The remaining fifty-four clips, which include motion events, activities (e.g. 'jogging') and states (an object displayed against a specific background), are not discussed in the present paper. All stimuli were pre-tested for ease of action recognition and homogeneity of labelling. The inter-stimulus-interval (a black screen with a white fixation cross) lasted eight seconds. All video clips showed realistic, everyday situations that were filmed and cut for this specific experiment.

4.3. PROCEDURE

Eye-movements were recorded with a remote *Eye Follower* eye-tracker (LC Technologies, Inc). Cameras were attached to the monitor for binocular eye-tracking and the eye-gaze system accommodated all natural head movements during normal computer operation. The gaze point sampling rate was 120 Hz, with a 0.45 degree gaze-point tracking accuracy throughout the operational head range. Stimuli were displayed on a 20" TFT monitor and participants were seated approximately 60 to 70 cm from the screen. Calibration was carried out once for each participant before the experiment (tracking fixations on yellow dots on a black screen, appearing at specific positions on the screen). The NYAN® software was designed to meet the requirements of analyzing eye-movements in relation to a dynamic visual input.[6] NYAN recorded eye-movements and audio data using an external microphone synchronously and timelocked (allowing the analysis of speech onset times in relation to stimulus onset).

Each recording was preceded by a training session with six video clips. Participants were given the following instructions in writing:

> *You will see a set of* 60 *video clips showing everyday events which are not in any way connected to each other. Before each clip starts, a black screen with a white fixation cross will appear. Please focus on this cross. For each video clip, it is your task to tell "what is happening", and you may begin as soon as you recognize what is happening in the clip. It is not necessary to describe the video clips in detail (e.g., "the sky is blue"). Please focus on the event only.*

[6] NYAN Eye tracking Data Analysis Suite 2.0 is developed by and available for purchase at Interactive Minds GmbH in Dresden, Germany.

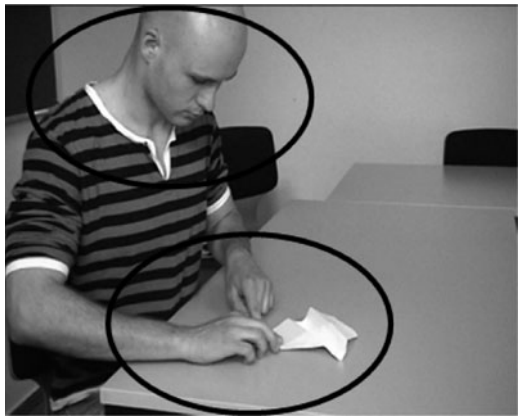


Fig. 1. Screenshot of stimulus (video clip) with the two areas of interest marked by ellipses.

Instructions were translated into German by a native speaker, and the experimenter was a native or highly proficient speaker of the language tested.[7] This means that all exchanges before and during the experiment took place in the participant's native language. Each recording session lasted approximately 15 minutes with no option of manipulating the presentation pace of the video clips. Video clips automatically started playing on the screen for 6 seconds, followed by a black screen lasting 8 seconds. A pre-test ensured that participants had sufficient time to inspect the scene and produce a full sentence to describe the event on-line. There was no cue as to when they should start speaking; this was left to the participant. Following the eye-tracking experiment, participants spent approximately 5 minutes filling out a questionnaire in their native language concerning their educational and linguistic background.

4.4. DATA CODING AND ANALYSES

4.4.1. *Pre-processing of the eye-tracking data*

The raw eye-tracking data were visually inspected for technical problems. Then, prior to further analyses, two ellipsoidal areas of interest were defined

[7] The critical question in the German instruction was *Was passiert?* 'What is happening?', using the simple present (unmarked for aspect). In German, any kind of expression of progressivity would be inappropriate in this context. In order to test the effect of the instruction formulation on the verb form used by English subjects, a pilot study was conducted in which instructions were varied for aspect ('What happens?' vs. 'What is happening?'). The results showed that use of the progressive in English was not affected by the phrasing of the instruction (see Carroll et al., 2004).

for each stimulus, depicting the agent and the action. These AoIs remained fixed over all recordings for all participants. The agent area of interest included the upper part of the body of a single agent (a man or a woman), and the action area of interest included the hands of the agent and one specific object (e.g. scarf, paper airplane) which was being acted upon (see Figure 1). The AoIs were defined on a frame-per-frame basis, ensuring that they always accounted for slight movement of the relevant scene elements during the entire presentation time. In fact, the agent AoI was stable in all stimuli, whereas the action AoI was dynamic (movements of the hands of the actor) – the size of the action AoI did not change over the course of the experiment. Depending on the size of the individual object acted upon, the action AoIs differed only marginally in their spatial dimensions between stimuli. We are aware of the fact that this slight variance represents a potential cause for discrepancy in gaze allocation patterns between items. We regard this as a necessary concession, however, when dealing with dynamic, live-recorded stimuli.

The NYAN system uses an area-based algorithm where a set of fixations with a maximum deviation of 25 screen pixels (corresponding to eye-movement of less than roughly 0.5° at approximately 68 cm distance from eye to screen), and a minimum sample count of six, is recognized as a fixation. Accordingly, movements that cover more than 0.5° on the scene were treated as saccades (at the average distance and monitor dimensions given).

Furthermore, NYAN detected speech onset times measured from stimulus onset for each participant. These were checked manually for accuracy. We manually timelocked the fixation timecourse to speech onset by dividing the stimulus presentation time (6000 ms) into intervals of 60 ms, taking speech onset time as temporal zero. To check for significant differences in speech onset times between languages, the data were analyzed using linear mixed effects models (package lme4; Bates, Maechler; & Bolker, 2012, in R version 3.0.2). We set up a model with language as fixed factor, participant and item as random factors, and a random slope for items (1 + Language | Item) (German was set as reference level). The results show no significant differences between groups (average SOT: 1800 ms from stimulus onset) (Factor *language*: estimate −0.101, standard error 0.153, *t*-value −0.662, *p*-value .508, n.s.).[8] For each 60 ms time interval, the occurrence and location of individual fixations was then registered with regard to five categories: 'AoI agent', 'AoI action', 'outside both AoIs', 'no fixation', and 'no measurement'

[8] The model reported here is one in which the random slope for items was excluded, given a correlation of −1 for the term (1+Language | Item) in the model with the random slope, indicating that it did not add much explanatory power to the model (cf. Bates, 2010).

(e.g., looking away from the screen, eyes closed). NYAN calculated the duration of fixations for each AoI before speech onset, for each individual participant.

4.4.2. *Audio data pre-processing and coding of the language data*

The audio data were transcribed by a native speaker of the respective language. The transcripts were then checked for accuracy by a second researcher. Since all event descriptions obtained in the current study made reference to agents and actions, the linguistic structures related to both units were analyzed in detail. All utterances (defined as clauses with finite verbs) were coded for each participant and each stimulus. Coding covered counts of the use of progressive aspectual markers on the main finite verb in each utterance (progressive versus non-progressive verb form; see Section 5.1.1), interpreted as markers of specificity of the action supplied on the verb. The degree of specificity in reference to the agent was coded as well, covering the use of qualifiers which elaborate specific visual characteristics of the agent (see Section 5.1.2): the complexity of the noun phrases that refer to the agent was captured with three coding categories, (a) an unspecific reference to the agent '(a) person' (category 'noun – unspecific'); (b) a more specific reference encoding the gender of the agent '(a) man/woman' (category 'noun'); and (c) more elaborate noun phrases that include adjectives or postverbal attributive structures ('a bald man', 'a blond woman', 'a man with glasses', 'a man in T-shirt') (category 'noun – specific').

4.4.3. *Statistical analyses of the eye-tracking data*

Overall gaze duration in the agent and action AoIs before utterance onset and the frequency of fixations in both AoIs over the pre-articulatory time window were analyzed. Both measures have been shown to relate to attention allocation to the respective components of a stimulus and language processing (Griffin, 2004). Gaze duration (looking time) in the agent and action AoIs was averaged and normalized in relation to the specific utterance onset on each trial, and compared between languages. Both sets of analyses – overall fixation duration and fixation frequency over time – were carried out by setting up linear mixed effects regression models (package lme4; Bates et al., 2012, in R version 3.0.2). First, a maximally specified random effects structure was included in the models, to account for random variation between participants and between items. Complex random terms which turned out not to have any explanatory power, as indicated by a correlation of around 1 or −1, or parameters of around 0, were removed from the eventual models reported below (following Bates, 2010). The models analyzing fixation duration included a random intercept for participants. For items, a random intercept and a random slope

for language was included, given that the present study concerns a within-items design, i.e., all participants in each group viewed the same video clips. The models analyzing fixation frequencies over time included a random term for participants with a random slope for time (bin).

5. Results

5.1. EVENT DESCRIPTIONS

5.1.1. *Type of verb form used*

Findings showed that German speakers did not use progressive aspectual constructions (0/114 utterances, typical description, e.g., *Ein Mann mit einer Glatze faltet einen Papierflieger* 'a bald man folds a paper airplane'), whereas English speakers used progressive aspect (*to be* V-*ing*) in each utterance (114/114 utterances, 100%, e.g., *A man is folding a paper airplane*).

5.1.2. *Specificity: references to the agent*

Table 1 shows the absolute and relative frequencies of the coding categories for the subject noun phrase (referring to the agent) in each language.

We treated the data as binary, analyzing the number of trials with specific agent phrases ('noun – specific') versus those without specific details about the agent ('noun – unspecific' and 'noun') using a mixed model with LANGUAGE as a fixed factor, in which German was set as the reference category. The model showed a significant effect of language (intercept: estimate −3.464, standard error 1.086, *z*-value −3.189; language: estimate −3.439, standard error 1.623, *z*-value −2.118, *p*-value .03),[9] showing that German participants produced significantly more complex subject noun phrases than English participants.

5.2. EYE-TRACKING DATA

5.2.1. *Fixation duration analyses*

The mean duration of all fixations (looking time) in each AoI was compared between languages. In order to control for minor fluctuations in speech onset latency (note that SOT was not fixed), gaze durations were normalized relative to speech onset of each participant on each trial. This was carried out by dividing the total gaze duration before speech onset by the speech onset latency on each trial.

[9] The model reported here included random intercepts for participants and items. The random slope for language and item was taken out, because the initial model returned a correlation of −1 for the term (1+Language | Item), probably due to the low number of datapoints.

TABLE 1. *Type of noun phrase used to encode the agent (the subject in all sentences) in German and English*

			Agent: level of specificity			
			noun - unspecific: a person	noun: a (wo)man	noun - specific: a young blond woman	Total
Language	*German*	Frequency	9	81	24	114
		%	7.9	71.1	21.1	100
	English	Frequency	7	103	4	114
		%	5.8	90.8	3.3	100

To pinpoint potential gaze allocation differences related only to the retrieval of more linguistic form concerning the first mentioned sentence element in the time window analyzed (e.g., adjectives, prepositional phrases used to describe the sentences' subjects, referring to the agent of the event), the subsequent analyses of gaze duration were performed on two datasets, one including all datapoints, and one disregarding those trials in which complex subject NPs were planned and produced by participants. This led to the exclusion of twenty-eight data points (of which 24 in the German dataset). The latter type of analysis is thus more informative with respect to our aim of isolating potential processing effects of aspectual perspective-taking (in English), or the absence thereof (in German), and general implications of these grammatical differences for event conceptualization and pre-articulatory visual processing in sentence production.

Table 2 shows the mean normalized gaze duration for each AoI before speech onset per language group, on all trials, and when excluding trials in which complex subject NPs were produced. [10]

The models evaluating fixation duration included LANGUAGE as a fixed factor. To exclude the possibility that the patterns of results obtained were due to the presence of outliers, extreme values with a standardized residual at a distance greater than 2.5 standard deviations from zero were removed from the data and the model was refitted (cf. Baayen, 2008). In the model including complex NP trials (all datapoints) six data points were excluded (1.32% each for the agent and the action AoI), and in the model on data excluding complex NPs six datapoints were excluded as well (2.01% for the agent AoI, 1.01% for the action AoI). The fixed effects are reported in Table 3 (language 'German' was the reference category).[11]

[10] In the German data, two participants were fully excluded, given the production of complex NPs on all six trials (one subject), or on five out of six trials.

[11] In the models reported in this section, the random slope for items was excluded, given a correlation of −1 for the term (1+Language|Item), indicating that it did not add much explanatory power to the models (cf. Bates, 2010).

TABLE 2. *Normalized gaze duration before utterance onset (total gaze duration before speech onset / SOT) (ms) (N = 19 in each group)*

		All data points		Excl. complex NPs	
AoI	Language	mean	SD	mean	SD
Agent	German	273	157	236	137
	English	209	145	205	142
Action	German	346	176	380	168
	English	356	200	360	198

TABLE 3. *Results of mixed models on normalized gaze duration data*

All data points				
AoI *Agent*	estimate	std. error	*t*-value	*p*-value[12]
(intercept)	0.23604	0.02998	7.872	<.001
Language	-0.03408	0.01288	-2.646	.009*
AoI *Action*	estimate	std. error	*t*-value	*p*-value
(intercept)	0.34931	0.03992	8.750	<.001
Language	0.00339	0.01853	0.183	.855 n.s.
Excl. complex NPs				
AoI *Agent*	estimate	std. error	*t*-value	*p*-value
(intercept)	0.21874	0.02586	8.459	<.001
Language	-0.02574	0.01183	-2.176	.031*
AoI *Action*	estimate	std. error	*t*-value	*p*-value
(intercept)	0.363393	0.038248	9.501	<.001
Language	-0.005448	0.018746	-0.291	.794 n.s.

There were no significant language effects with respect to gaze durations in the action AoI.

The fact that English speakers produced longer and more complex verb phrases than German speakers, given the morphological marking of aspect (V-*ing* vs. V), was not reflected in longer looks in the action region in the time window analyzed. With respect to gaze duration in the agent AoI, results showed a language effect in both datasets. This effect was more pronounced in the dataset including complex NP trials, indicating that, to some extent, the production of more linguistic form in the subject phrase was reflected in longer looking times to the agent. Nevertheless, factors other than lexical and phonological retrieval and encoding processes of the first mentioned sentence element seem to contribute to pre-articulatory gaze patterns. To investigate

[12] Following Barr (2012), we calculated *p* values on the basis of the *t*-values, using the following code in R: tvalues <- fixef(model) / sqrt(diag(vcov(model))) pvalues <- 2*(1-pnorm(abs(tvalues))).

those factors further, the following sets of analyses focused only on the dataset matched for overall subject NP complexity.

5.2.2. *Analyses of fixation frequencies over time*

For visual inspection we plotted fixation proportions in the agent and the action region for German and English participants, in the phase leading up to utterance onset (Figure 2; data plotted are matched for subject NP complexity).[13]

Visual inspection of Figure 2 allows us to infer a number of things: in the English dataset, the proportion of looks in the agent region never exceeded those in the action region throughout the time window plotted. In the German data, this is the case for a short time window around 1200 ms before stimulus onset. There seems to be a steeper increase in action fixations in the English data compared to German, from around 900 ms until about 360 ms before SOT. In particular between −600 ms and −300 ms there seem to be more action fixations in English than German. With respect to agent fixation frequencies, languages at first sight differ between approximately −900 ms and −300 ms.

To statistically evaluate these observations, two mixed effects models (one for the agent AoI, one for the action AoI) were set up in which the evolvement of fixation patterns over time was represented by a sequence of six successive time bins of 300 ms each, covering the whole pre-articulatory time window. Data aggregation was carried out to account for a potential non-independence of data in subsequent time intervals, typical for timecourse-related eye-movement data. We only report analyses on data aggregated by subjects. This is motivated by the fact that items were balanced across the language condition (i.e., all participants saw all items), and the experiment involved only a small number of items making the by-items statistics of limited power. Elogits for each AoI were used as the dependent variable (cf. Barr, 2012). They were calculated as follows:

$$\text{elogit} = \log\left(\frac{(\text{sum of registered fixations in AOI} + 0.5)}{(\text{total of fixations} - \text{sum of registered fixations in AOI} + 0.5)}\right)$$

Total fixations in the above equation were a multiplication of the number of items per bin for each subject, and the number of time bins over which the aggregation was done (six time bins). The levels of the factor 'time prior to SOT' BIN were coded as follows: (0): 0 to −300 ms, (0.3): −300 to −600 ms, (0.6): −600 to −900 ms, (0.9): −900 to −1200 ms, (1.2): −1200 to −1500 ms, (1.5): −1500 to −1800 ms. The reference level was the first bin (0), thus taking the time window shortly before and at utterance onset as the basis for comparison with fixation patterns in all bins prior to this window. Note that at SOT both

[13] For the interested reader, 'Appendix A' also shows plots depicting fixation proportions over time, based on all datapoints (including complex subject NP trials).

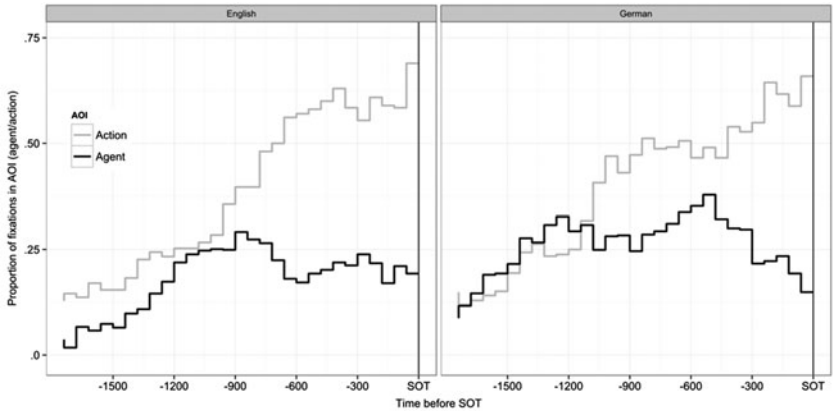


Fig. 2. Proportion of fixations in agent (black) and action (grey) areas of interest per 60 ms time interval (time plotted: −1800 ms until utterance onset, SOT), in English (left panel) and German (right panel) participants.

groups of participants exhibited almost the same distribution of attention to the agent and the action (similar proportions of agent and action looks). The levels of the predictor LANGUAGE were coded as follows: (−0.5): German, (0.5): English. Weights were included in the models because the variance of the elogit depends on the mean (cf. Barr, 2012). They were calculated as follows:

$$\text{Weights} = \frac{1}{\text{sum of registered fixations in AOI} + 0.5} + \frac{1}{\text{Total of fixations - sum of registered fixations in AOI} + 0.5}$$

Agent fixation proportions over time. The analysis of agent-elogits included LANGUAGE and BIN as predictor variables, testing for potential interaction effects. Bin was treated as a factor. A significant interaction means that the difference between languages at bin 0 (−300 ms to SOT) significantly differed from that in a specific other bin. Significant effects were found in the time bin from −300 ms to −600 ms, indicating fewer agent fixations in the English participants. The same pattern was obtained for the two bins from −1200 ms to −1800 ms before SOT (see 'Appendix B').

Action fixation proportions over time. This analysis concerned elogits for the action AoI, testing for an interaction of the predictor variables LANGUAGE and BIN. Significant interactions were found for the bin from −300 ms to −600 ms, showing more looks in the action region in the English group than in the German group. Furthermore, in the two bins from −900 to −1500 ms there were fewer action fixations in the English group (see 'Appendix C').

Evolvement of attention allocated to agent and action in German and English. The above analyses revealed differences in direct language comparisons for specific time bins, suggesting that the evolvement of attention allocated to agent and action regions over time differed within the two languages. Two additional analyses were performed, specifying main effects of BIN for each AoI and each language separately, to pinpoint more directly the evolvement of agent and action fixations over time in each group (see 'Appendix D').

In German, with respect to the agent AoI, compared to the time bin shortly before and at SOT (the reference level), there were significantly more fixations in all preceding bins, except for the one furthest away from SOT (−1500 to −1800 ms); in this bin, the model detected significantly fewer agent fixations. In the English group the evolvement of agent fixations over time was different: there was only one bin prior to SOT with significantly more agent looks than at SOT, i.e., between −900 and −600 ms (a small peak is visible in Figure 2). In the two bins furthest away from SOT we find fewer agent looks.

With respect to the action region in the German data, significantly fewer fixations were registered during all pre-articulatory time bins. Model estimates indicated similar and high proportions of action fixation between −300 ms and SOT (estimates in the two bins prior to SOT hardly differed from each other), fewer fixations during earlier time bins (between −300 and −900 ms), and much lower proportions closer to stimulus onset. Fixation patterns in action regions looked different in English participants: proportions of action fixation were similarly high between −600 ms and SOT. Before that time, fixation frequencies were significantly lower; an evaluation of model estimates from bin to bin showed large differences, meaning a steady decrease in action looks going back towards stimulus onset from −600 ms.

6. Discussion

The present study shows how the distribution of attention allocated to agents and actions in dynamic scenes during pre-articulatory sentence planning is, to some extent, affected by language differences between English and German. We postulate that differences with respect to grammatical aspect lead to different patterns of event conceptualization and sentence planning, as reflected in pre-articulatory visual processing patterns, and event descriptions. Progressive aspect (in English) encodes a particular temporal perspective on an event. At the same time, it marks that the utterance produced refers to a specific instance of a situation, thus distinguishing it from generic statements about, or habitual instances of, a situation.

Part of the requirements for satisfactory task fulfilment, given the on-line description of 'what is happening' in dynamic scenes as in the present task, is

for a speaker to explicitly refer to a specific event, rather than to make a generic statement about a situation. English speakers met this requirement in marking progressive aspect on the action verb in each utterance produced. In a non-aspect language like German, there are no options for marking the status of an event as specific, as opposed to generic, by means of verbal aspect. Instead, the data displayed a tendency in German speakers to do this by referring to more specific details of the main participant in the event, the agent involved in the action, reflected in the more frequent occurrence of complex subject noun phrases in the event descriptions.

With respect to attention patterns before utterance onset, overall longer looking times in agent regions were found in German, compared to English. Interestingly, this could not be explained by the use of more complex linguistic surface forms (subject NPs referring to the agent) in German speakers alone, given the occurrence of this language effect also (though less pronounced) when analyzing only utterances matched for subject NP complexity. Analyses of fixation frequencies over time in fact showed an early increased allocation of attention to agent regions in German speakers, roughly between 1800 and 1200 ms before utterance onset. This time window is very likely to precede lexical and phonological encoding (formulation) processes related to the first mentioned element in the sentences produced, which is the agent in all cases. This allows us to trace the early differences in attention to event agents between German- and English-speaking participants to differences in the early planning processes of scene conceptualization and message generation, related to the event and sentence as a whole. During these stages of processing, the protagonist in an event seems to be of greater importance to German speakers (a non-aspect language), than to speakers of an aspect-language (like English): the agent of a causative event provides one of the options for conveying the event status as specific. On the surface, this was reflected in the production of a number of complex NPs referring to the agent in German; at a processing level, this was reflected in an early allocation of gaze to agents in German participants.

The analyses of pre-articulatory looking times did not display a language effect for the action regions in the scenes. The production of longer and more complex aspectually marked verb phrases in English (i.e., *to be* V-*ing* verb phrases, compared to unmarked verb phrases in German) was not reflected in this measure during the time window analyzed.

Main findings with respect to pre-articulatory processing reflected in fixation proportions in agent and action regions over time were twofold: first, attention allocation patterns to agents and actions of events evolved differently over time in the two languages. Second, in both languages action regions generally attracted more attention than agent regions throughout almost the entire timecourse up to speech onset. With respect to the first main finding,

language differences were detected in several time bins before speech onset. During early phases we found more agent fixations in the German group, that is, in the time span between 1800 and 1200 ms before utterance onset, which is very likely to precede formulation processes of the first mentioned referent (see above), the start of which has been shown to approximate a second before naming (eye–voice span; cf. Bock et al., 2003; Griffin & Bock, 2000). Strikingly, in fairly early time bins (between 1500 and 1200 ms, and 1200 and 900 ms before utterance onset), there were higher proportions of action fixations in German participants than English, even though, in general, action fixation frequencies were lower in time windows further away from utterance onset. Again, it seems unlikely that these fixations reflected linguistic encoding processes, in part given by the fact that the action is only the second mentioned element in the sentence (referred to by means of the finite verb). This pattern of fixations may well reflect similar processing to that postulated for the early allocation of gaze to agents; German speakers have been shown to encode specificity also in relation to the action, not by verbal means, but rather by the use of qualifiers, e.g., prepositional phrases related to instruments of actions, for example (cf. van Beek et al. 2013). The action region may thus also be relevant for marking event status in German in early event processing – at this point, we can only speculate, however, given that there were no such productions in the current dataset. During part of this same time window a peak in agent fixations starts in English participants, lasting until around 600 ms before utterance onset. A time span of 900 to 600 ms between looking and naming (eye–voice span) would fit the current picture: the peak for agent fixations falls roughly around 900 ms, thus within the eye–voice span for the naming of the agent (sentence subject).

In the time interval from 600 to 300 ms before speech, English speakers showed a greater preference for the action region than the German speakers, which was complemented by the finding that speakers of German showed more agent looks than English speakers in this time window. For German speakers, highest action fixation proportions were detected between −300 ms and utterance onset, i.e., starting later and lasting for a shorter period of time. This earlier and longer-lasting 'action preference' in English speakers could be related to processing implications of aspect in general, i.e., a stronger attentional focus on action contours, or to an earlier start of formulation processes of the verb phrase in English than in German (given the increased complexity of aspect marking on the verb), or both.

To summarize our first main finding, overall, in German speakers we found more attention allocated to the agents depicted in the event scenes, which could not be exclusively related to the retrieval of linguistic form from the mental lexicon to describe the agents. English participants showed high action fixation proportions for a longer time window before utterance onset than German speakers. Given the fact that utterance onset latencies were roughly similar in

English and German speakers, our results show that pre-articulatory attention allocation patterns are, at least in part, language-specific; they indicate a different temporal coordination of information uptake, information selection and organization, and linguistic encoding, in the two languages.

With respect to our second main finding, in general, the present results concerning fixation patterns in relation to the verbalization of dynamic events diverge from findings obtained in previous studies using still images (e.g., Bock et al., 2003; Griffin & Bock, 2000; overview in Bock et al., 2004).[14] The present study provides an important source of complementary results from a more naturalistic paradigm to studies using non-dynamic stimuli. Studies using the latter type of stimuli have identified a tight sequential link between looking at elements of events, and naming these. These studies have proven to be a good tool for investigating formulation processes mainly. In the present study, an eye–voice span of around 900 to 600 ms before naming, during which a particular object or entity to be named is fixated most frequently and more frequently than other elements in the scene, could not straightforwardly be identified for the agents of the events. Fixation frequencies in agent regions never exceeded fixations in action regions during any time interval before utterance onset, which could be related to formulation processes, even though this point in time marked the starting point of the articulation of the sentence's subject, in all cases referring to the agent of the action. Findings point to a more complex set of factors that determine pre-articulatory gaze allocation patterns. Coco and Keller (2010), though using a different paradigm, argue along the same lines, by stating that a 'simple view', according to which referents are fixated in order of mention with a fixed eye–voice span, may not seem to generalize to more real-life settings in which speakers describe naturalistic scenes. Current findings thus underline task-driven and context-related processing. Video clips showing causative events clearly foreground the ongoing action as it unfolds in real time, and speakers may need to direct attention to possible unforeseeable developments in the action being performed. The timecourse analyses, showing a rapidly increasing fixation frequency in AoI action, and an overall higher frequency of fixations in the action AoI, support this view of visual processing given dynamic events. By contrast, under the hypothesis of a close timelock between viewing and naming, a dominance of looks in the action region should be expected to occur only after speech onset, following the formulation of the agent, given the relevance of the action region for information extraction related to the planning of the verb as well as the object. This is thus not entirely reflected in our data. The present

[14] Besides the use of still images depicting events, the studies cited also used a preview paradigm, adding to the homogeneity of attention and description patterns across participants, making the task less naturalistic, however.

study indicates how the timelock between speaking and gazing may be less strictly sequentially ordered, dependent on the context and task conditions, which are naturalistic in the present case. Looking and naming are, however, linked, since the highest frequency of agent fixations was found in roughly the second before speech onset for speakers of both languages (English: between 1200 and 600 ms before SOT; German: between 600 and 300 ms before SOT).

All in all, we argue that the different processing patterns obtained in the current production task, as reflected in a different pre-articulatory distribution of attention to two relevant event elements, could be due to structural differences between English and German. The findings thus correspond to results in Sauppe et al. (2013), who concluded that grammatical requirements at sentence level, involving several constituents in a sentence, are reflected in visual attention before SOT, to some extent overruling effects of sequential order of surface form. The present study adds to the growing body of evidence showing differences in processing, in this case in a complex production task, between speakers of different language backgrounds (see overview in Jaeger & Norcliffe, 2009). We contribute to this debate in showing visual processing differences along a specific timecourse of language planning, given structural differences between languages in the way in which full-fledged finite and specific event descriptions are produced: we find intricate processing differences between speakers of languages that differ in the grammatical domain, and thus evidence for the implications of grammaticalized categories, or the absence thereof, on information organization in a complex production task, using naturalistic stimuli.

7. Conclusions

The present study on attention allocation and sentence generation in an unscripted production task, using naturalistic scenes of events which unfold over time, illustrates how the processes of message planning and the encoding of linguistic form are inter-related and reflected in gaze in a complex manner. Findings show how the timelock between speaking and gaze may be less strictly sequentially ordered, even though clearly linked, compared to when events are presented in static form or when single objects have to be named sequentially. The results point to the mediation of linguistic means and grammatical structures (and their functions in event construal) in event processing during language planning phases before utterance onset. The present study demonstrates how abstract linguistic knowledge (linked to the presence or absence of grammatical categories in specific languages), and associated means of expression related to the category of 'event status', for example, can function as a guiding factor for visual attention – a factor which so far has not been subject to extensive analysis (see also Huettig et al., 2011). A challenge lies in the disentangling of different factors that are inter-related and integrated

in the course of the message planning process. The cross-linguistic production paradigm used in the present study may serve to surmount this problem. Linguistic contrasts facilitate the identification of the role of different types of factors induced by language: on the one hand, grammatical factors which attribute saliency to specific aspects of a scene, and, on the other hand, factors determined by processing constraints such as order of elements in a sentence.

REFERENCES

Anderson, S., Matlock, T., & Spivey, M. (2013). Grammatical aspect and temporal distance in motion descriptions. *Frontiers in Psychology*, **4**, 1–9.

Baayen, H. (2008). *Analyzing linguistic data*. Cambrige: Cambridge University Press.

Barr, D. (2012). Walkthrough of an 'empirical logit' analysis in R, online: <http://talklab.psy.gla.ac.uk/tvw/elogit-wt.html>.

Bates, D. (2010). Correlated random effects in lmer and false convergence, online: <https://stat.ethz.ch/pipermail/r-sig-mixed-models/2010q2/003921.html> (last accessed 8 January 2014).

Bates, D., Maechler, M., & Bolker, B. (2012). Package 'lme4'. Linear mixed effects models using S4 classes, online: <http://cran.r-project.org/web/packages/lme4/index.html>.

Beek, v. G., Flecken, M., & Starren, M. (2013). Aspectual perspective-taking in L1 and L2 Dutch. *International Review of Applied Linguistics*, **51** (2), 199–227.

Bergen, B., & Wheeler, K. (2010). Grammatical aspect and mental simulation. *Brain and Language*, **112**, 150–158.

Berman, R., & Slobin, D. (1994). *Relating events in narrative: a crosslinguistic developmental study*. Hillsdale, NJ: Erlbaum.

Bock, K., Irwin, D., & Davidson, D. (2004). Putting first things first. In J. Henderson & F. Ferreira (Eds.), *The interface of language, vision, and action: eye movements and the visual world* (pp. 224–250). New York: Psychology Press.

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

Bock, K., & Levelt, W. (1994). Language production: grammatical encoding. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 945–984). London: Academic Press.

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

Brown-Schmidt, S., & Tanenhaus, M. (2006). Watching the eyes when talking about size: an investigation of message formulation and utterance planning. *Journal of Memory and Language*, **54**, 592–609.

Carroll, M., & Stutterheim, C. v. (2011). Event representation, time event relations, and clause structure: a cross-linguistic study of English and German. In E. Pederson & J. Bohnemeyer (Eds.), *Event representation* (pp. 68–83). Cambridge: Cambridge University Press.

Carroll, M., Stutterheim, C. v., & Nüse, R. (2004). The language and thought debate: a psycholinguistic approach. In C. Habel & T. Pechmann (Eds.), *Approaches to language production* (pp. 183–218). Berlin: Mouton de Gruyter.

Coco, M., & Keller, F. (2010). Sentence production in naturalistic scenes with referential ambiguity. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 1070–1075). Austin, TX: Cognitive Science Society.

Comrie, B. (1976). *Aspect: an introduction to the study of verbal aspect and related problems*. Cambridge: Cambridge University Press.

Dahl, Ö. (1995). The marking of the episodic/generic distinction in tense-aspect systems. In G. Carlson & F. Pelletier (Eds.), *The generic book* (pp. 412–425). Chicago: University of Chicago Press.

Dobel, C., Glanemann, R., Kreysa, H., Zwitserlood, P., & Eisenbeiss, S. (2010). Visual encoding of coherent and non-coherent scenes. In J. Bohnemeyer & E. Pederson (Eds.),

Event representation in language and cognition (pp. 189–215). Cambridge: Cambridge University Press.

Flecken, M. (2011). Event conceptualization by early bilinguals: insights from linguistic and eye tracking data. *Bilingualism: Language and Cognition*, **14** (1), 61–77.

Flecken, M., & Gerwien, J. (2013). Grammatical aspect modulates event duration estimations: findings from Dutch. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 2309–2314). Austin, TX: Cognitive Science Society.

Gleitman, L., January, D., Nappa, R., & Trueswell, J. (2007). On the give and take between event apprehension and utterance formulation. *Journal of Memory and Language*, **57** (4), 544–569.

Griffin, Z. (2004). Why look? Reasons for eye movements related to language production. In J. Henderson & F. Ferreira (Eds.), *The integration of language, vision, and action: eye movements and the visual world* (pp. 213–247). New York: Taylor & Francis.

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

Griffin, Z., & Spieler, D. (2006). Observing the what and when of language production for different age groups by monitoring speakers' eye movements. *Brain and Language*, **99**, 272–288.

v. Heusinger, K. (2002). Specificity and definiteness in sentence and discourse structure. *Journal of Semantics*, **19** (3), 245–274.

Huettig, F., Rommers, J., & Meyer, A. (2011). Using the visual world paradigm to study language processing: a review and critical evaluation. *Acta Psychologica*, **137**, 151–171.

Jaeger, F., Furth, K., & Hilliard, C. (2012). Incremental phonological encoding during unscripted sentence production. *Frontiers in Psychology*, **3**, 1–22.

Jaeger, T., & Norcliffe, E. (2009). The cross-linguistic study of sentence production: state of the art and a call for action. *Language and Linguistics Compass*, **3** (4), 866–887.

Klein, W. (1994). *Time in language*. London: Routledge.

Kuchinsky, S., Bock, K., & Irwin, D. (2011). Reversing the hands of time: changing the mapping from seeing to saying. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **37** (3), 748–756.

Levelt, W. (1989). *Speaking: from intention to articulation*. Cambridge, MA: MIT Press.

Lucy, J. (1992). *Grammatical categories and cognition: a case study of the linguistic relativity hypothesis*. Cambridge: Cambridge University Press.

Matlock, T. (2010). Abstract motion is no longer abstract. *Language and Cognition*, **2**, 243–260.

Meulen, F. van der, Meyer, A., & Levelt, W. (2001). Eye movements during the production of nouns and pronouns. *Memory & Cognition*, **29**, 512–521.

Meyer, A. (2004). The use of eye tracking in studies of sentence generation. In J. Henderson & F. Ferreira (Eds.), *The interface of language, vision, and action: eye movements and the visual world* (pp. 191–211). New York: Psychology Press.

Meyer, A., & Dobel, C. (2003). Application of eye tracking in speech production research. In J. Hyöna, J. Radach, & H. Deubel (Eds.), *The mind's eye: cognitive and applied aspects of eye movement research* (pp. 253–272). Oxford: Elsevier Science.

Meyer, A., Sleiderink, A., & Levelt, W. (1998). Viewing and naming objects: eye movements during noun phrase production. *Cognition*, **66**, 25–33.

Papafragou, A., Hulbert, J., & Trueswell, J. (2008). Does language guide event perception? Evidence from eye movements. *Cognition*, **108** (1), 155–184.

Rijkhoff, J., & Seibt, J. (2005). Mood, definiteness and specificity: a linguistic and a philosophical account of their familiarities and differences. *Tidsskrift for Sprogforskning*, **3** (2), 85–132.

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

Slobin, D. (1996). From 'Thought and language' to 'Thinking for speaking'. In J. Gumperz & S. Levinson (Eds.), *Rethinking linguistic relativity* (pp. 70–96). Cambridge: Cambridge University Press.

Soroli, E., & Hickmann, M. (2010). Language and spatial representations in French and in English: evidence from eye-movements. In G. Marotta, A. Lenci, L. Meini, & F. Rovai (Eds.), *Space in language* (pp. 581–597). Pisa: Editrice Testi Scientifici.

Strömqvist, S., & Verhoeven, L. (2004), *Relating events in narrative. Vol. 2: typological and contextual perspectives*. Mahwah, NJ: Erlbaum.

Stutterheim, C. v., Andermann, M., Carroll, M., Flecken, M., & Schmiedtová, B. (2012). How grammaticized concepts shape event conceptualization: insights from linguistic analysis, eye tracking data and memory performance. *Linguistics*, **50** (4), 833–867.

Stutterheim, C. v, Carroll, M., & Klein, W. (2009). New perspectives in analyzing aspectual distinctions across languages. In W. Klein & P. Li (Eds.), *The expression of time* (pp. 195–216). Berlin: Mouton de Gruyter.

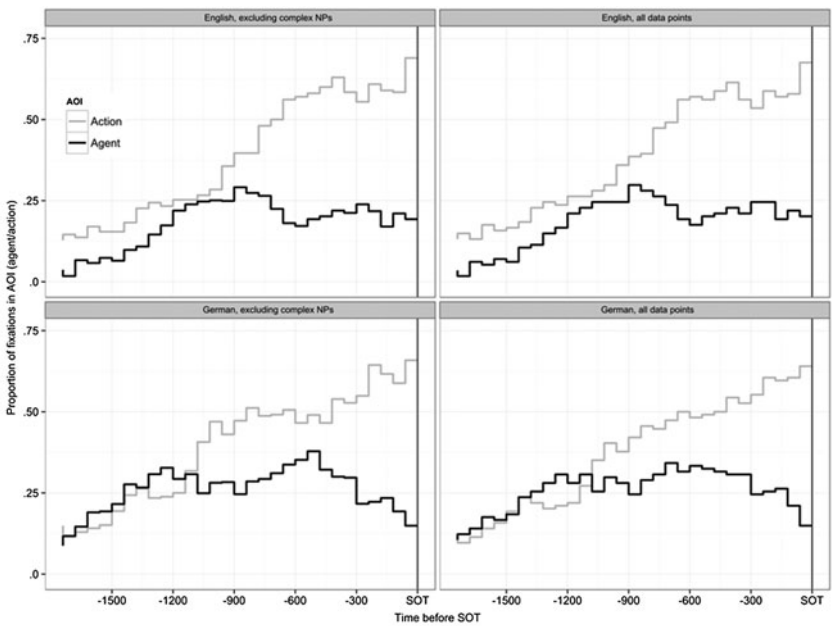
Stutterheim, C. v., & Nüse, R. (2003). Processes of conceptualization in language production: language-specific perspectives and event construal. *Linguistics*, **41** (5), 831–881.

Talmy, L. (1985). Lexicalization patterns: semantic structure in lexical forms. In T. Shopen (Ed.), *Language typology and syntactic description. Vol. 3: grammatical categories and the lexicon* (pp. 57–149). Cambridge: Cambridge University Press.

Tenny, C., & Pustejovsky, J. (2000). A history of events in linguistic theory. In C. Tenny & J. Pustejovsky (Eds.), *Events as grammatical objects* (pp. 3–37). Chicago: University of Chicago Press.

APPENDIX A

Plots depicting agent (black) and action (grey) fixation proportions before utterance onset (SOT): original data sets including all datapoints (right panel); datasets excluding trials with complex subject noun phrases (left panel) in English (top) and German (bottom) participants.



APPENDIX B

Output of mixed effect model on elogit values for AoI AGENT (German was coded as −0.5, English as 0.5).

Formula: lmer(elog~LANGUAGE*BIN+(1+BIN|Subject),data=data,subset=data$AOI=="-1",weights=1/wts)

	AIC/BIC/ logLik/ deviance/ REMLdev		estimate	std. error	*t*-value
German vs. English	877.3/992/ −404.6 / 791/ 809.3/	(Intercept)	−1.16985	0.06838	−17.107
		LANGUAGE	−0.06496	0.13677	−0.475
		BIN0.3	0.24992	0.07831	3.191
		BIN0.6	0.29501	0.07319	4.031
		BIN0.9	0.23687	0.08809	2.689
		BIN1.2	−0.15009	0.10057	−1.492
		BIN1.5	−0.94136	0.12116	−7.77
		LANGUAGE:BIN0.3	−0.50841	0.15662	−3.246
		LANGUAGE:BIN0.6	−0.10351	0.14639	−0.707
		LANGUAGE:BIN0.9	−0.21412	0.17619	−1.215
		LANGUAGE:BIN1.2	−0.85040	0.20115	−4.228
		LANGUAGE:BIN1.5	−1.02172	0.24232	−4.216

APPENDIX C

Output of mixed effect model on elogit values for AoI ACTION (German was coded as −0.5, English as 0.5).

Formula: lmer(elog~LANGUAGE*BIN+(1+BIN|Subject),data=data,subset=data$AOI=="1",weights=1/wts)

	AIC/BIC/ logLik/ deviance/ REMLdev		estimate	std. error	*t*-value
German vs. English	901.2/ 1016 −416.6/ 814.6/ 833.2/	(Intercept)	0.42712	0.05786	7.382
		LANGUAGE	0.03278	0.11572	0.283
		BIN0.3	−0.19675	0.06011	−3.273
		BIN0.6	−0.53689	0.08001	−6.71
		BIN0.9	−1.27792	0.08815	−14.497
		BIN1.2	−1.81732	0.08608	−21.112
		BIN1.5	−2.53802	0.11364	−22.334
		LANGUAGE:BIN0.3	0.30396	0.12021	2.528
		LANGUAGE:BIN0.6	−0.18447	0.16003	−1.153
		LANGUAGE:BIN0.9	−0.49872	0.1763	−2.829
		LANGUAGE:BIN1.2	−0.36509	0.17216	−2.121
		LANGUAGE:BIN1.5	−0.30133	0.22728	−1.326

APPENDIX D

Output of timecourse analyses of agent and action elogit values for each language separately (German was coded as −0.5, English as 0.5; 'Language' was a factor in this analysis).

AoI AGENT

Formula: lmer(elog~BIN*LANGUAGE-BIN+(1+BIN|Subject),data=data, subset=data$AOI==“-1”,weights=1/wts)

	AIC/BIC/ logLik/ deviance/ REMLdev		estimate	std. error	*t*-value
German vs. English	877.3 / 992 −404.6 / 791/ 809.3/	(Intercept)	−1.137368	0.10043	−11.325
		LANGUAGE0.5	−0.064963	0.136765	−0.475
		BIN0.3: LANGUAGE−0.5	0.504116	0.1133	4.449
		BIN0.6: LANGUAGE−0.5	0.346759	0.106557	3.254
		BIN0.9: LANGUAGE−0.5	0.343923	0.128044	2.686
		BIN1.2: LANGUAGE−0.5	0.275113	0.141821	1.94
		BIN1.5: LANGUAGE−0.5	−0.430505	0.168515	−2.555
		BIN0.3: LANGUAGE0.5	−0.004289	0.108134	−0.04
		BIN0.6: LANGUAGE0.5	0.243246	0.100369	2.424
		BIN0.9: LANGUAGE0.5	0.129808	0.121027	1.073
		BIN1.2: LANGUAGE0.5	−0.575284	0.142645	−4.033
		BIN1.5: LANGUAGE0.5	−1.452227	0.174124	−8.34

AoI ACTION

Formula: lmer(elog~BIN*LANGUAGE-BIN(1+BIN|Subject),data=data, subset=data$AOI==“1”,weights=1/wts)

	AIC/BIC/ logLik/ deviance/ REMLdev		estimate	std. error	*t*-value
German vs. English	901.2 /1016/ −416.6/ 814.6 / 833.2/	(Intercept)	0.41073	0.08487	4.84
		LANGUAGE0.5	0.03278	0.11571	0.283
		BIN0.3: LANGUAGE−0.5	−0.34873	0.08825	−3.952
		BIN0.6: LANGUAGE−0.5	−0.44466	0.11661	−3.813
		BIN0.9: LANGUAGE−0.5	−1.02856	0.12745	−8.07
		BIN1.2: LANGUAGE−0.5	−1.63477	0.12478	−13.101
		BIN1.5: LANGUAGE−0.5	−2.38738	0.16767	−14.239
		BIN0.3: LANGUAGE0.5	−0.04477	0.08163	−0.549
		BIN0.6: LANGUAGE0.5	−0.62912	0.10958	−5.741
		BIN0.9: LANGUAGE0.5	−1.52727	0.12181	−12.538
		BIN1.2: LANGUAGE0.5	−1.99986	0.11861	−16.861
		BIN1.5: LANGUAGE0.5	−2.68869	0.15343	−17.523