First things first: cross-linguistic analyses of event apprehension

Master Thesis
presented by
Muqing Li 李沐晴

Supervisor:
Dr. Monique Flecken
Max Planck Institute for Psycholinguistics
Nijmegen, the Netherlands

Second reader:
Dr. David Peeters
Max Planck Institute for Psycholinguistics
Nijmegen, the Netherlands

Student number: 4765583
Programme: Research Master in Language and Communication
Faculty of Arts, Radboud University Nijmegen
Date of submission: 22 August, 2018

Acknowledgement

This thesis is the result of my internship at the Neurobiology of Language Department of Max Planck Institute for Psycholinguistics. First of all, I am extremely thankful for my supervisor Dr. Monique Flecken. Her supportive, efficient, and interactive guidance throughout the whole project allowed me to gain a magnificent amount of research experiences and skills. Without her help and guidance, I would not be able to find my research passion in psycholinguistics and achieve the tremendous improvement within two years. I hope we can further collaborate in the future for more interesting studies on event cognition engaging multiple languages. I would also like to thank Dr. Johannes Gerwien, our collaborator in Heidelberg, for the many valuable and creative ideas that are fundamental for this project.

This thesis also marks the end of my Research Master in Language and Communication at Radboud University Nijmegen. In addition to Monique, who also provided generous help on my courses, lab rotations, and PhD applications, I would like to thank Dr. Geertje van Bergen, and prof. Dr. Asli Özyürek and many other lecturers who offered critical advice and feedbacks that always kept me on the good track for proposing research ideas in courses and during the internship.

The amount of effort required for finishing this internship project as well as the master program within two years were extremely high, but luckily, I encountered many friends, from various countries in this world, for a lot of fun and support. Cheers to my dear international team Hande, Rehana, Christoph, Austin and Julia, we made it! I thank my supportive colleagues and friends from MPI and other institutes, Julia, Ksenija, Miguel, Yiyun, Cas, Jinbiao, Chen, James and NBL lunch group. Last but not least, to my dear Chinese friends in Nijmegen Yongxin, Feifei, Yingdi who were always open for conversations and helps, and made me feel like home. And mom, who chose to let me go for this adventure in a country that is ten-hour flight away from home.

Contents

1 Introduction 1

1.1 Message encoding in language production: the starting point debate 2

1.2 Cross-linguistic differences in language production 5

1.3 How to isolate the apprehension process: brief exposure paradigms 7

2 Aims of the present study 11

3 Experiment 1 18

3.1 Method 18

3.2 Data preprocessing, coding and analysis 21

3.3 Results 25

3.4 Discussion of Experiment 1 30

4 Experiment 2 34

4.1 Method 34

4.2 Data preprocessing and analysis 36

4.3 Results 37

4.4 Discussion of Experiment 2 41

5 General Discussion 44

5.1 Apprehension is a flexible process 46

5.2 First fixation locations and message encoding 47

5.3 Top-down effect of language on apprehension: more research is required 50

6 Conclusion 52

References i

Appendix 1. List of stimuli in Experiment 1 iv

Appendix 2 List of stimuli in Experiment 2 v

Appendix 3. A descriptive scatterplot of FFLs in Experiment 1. vi

Appendix 4. FFLs for individual stimulus vii

Abstract

Apprehension is the rapid visual process during which the gist of a scene can be extracted. This study investigates potential top-down effects of task demands (different language production tasks) and language background of the viewer (Mandarin Chinese and Dutch) on event apprehension. The tasks manipulate require information extraction from different elements of the scene (agent/action naming, event description). The manipulation of language background involves different degrees of saliency of agents in event perception: Mandarin Chinese allows subject omission when sufficient context is given, while Dutch obligatorily requires the encoding of the subject in a sentence. We ask whether these factors influence apprehension processes. In two experiments, we present causative event pictures, showing agents performing actions on objects, for a duration of only 300ms. Upon stimulus offset, Dutch and Chinese participants describe the agent and/or the action, or describe the entire event, following the different task demands. We measure the first fixation location for each task and language group, as an index of the information processed during apprehension, and as such, as a reflection of the result of this process. For the first time, we show that apprehension is a flexible process, in that it is modulated by task demands: first fixation locations differ depending on the requirements of the task. Furthermore, we find that the accuracy, specificity and the starting point of speakers’ verbal descriptions cannot be predicted by first fixation locations, indicating that this measure indeed reflects processes prior to linguistic formulation in language production. We observe mixed findings concerning cross-linguistic differences in first fixation patterns, inviting further exploration.

1 Introduction

Language production begins with the conceptualization and formulation of the message to be uttered (message encoding, Levelt, 1989). In everyday life, we talk about the dynamic events that we see in the environment around us (e.g., seeing and describing a person who is reading a book, or drawing something on paper). The planning of a message in this situation engages multiple complex mechanisms: it requires, first, the visual encoding of the scene, then the conceptualization of the event structure and contents, and finally the linguistic formulation of the message (Konopka & Brown-Schmidt, 2014). Surprisingly, although complex, extracting the *gist* from a visual scene is an extremely rapid process, which can be achieved within a single glance**.** This process is known as *apprehension* (Henderson & Ferreira, 2004), which is investigated in the current study to shed light on the early phases of language production.

Studies have shown that apprehension is a rapid and flexible process, during which multiple dimensions of information in a visual scene can be captured. People can already detect cued objects or pre-identified scenes within in as little as 30-50ms (e.g., Biederman, 1981; Hollingworth & Henderson, 1998; Potter & Levy, 1969). Basic-level category information (e.g., a park) and scene spatial layouts (e.g., objects along both sides of a street) can also be extracted within less than 100ms (e.g., Potter, 1976; Schyns & Oliva, 1994). In addition, scene coherence of an event can be correctly judged in an above-chance level within as short as 30ms (Dobel et al., 2007; Glanemann et al., 2016). Last but not least, people can successfully detect event roles and categories (i.e., answering what the event agent, patient or action is) to a great extent, already within 37ms (Hafri, Papafragou, & Trueswell, 2013). These findings suggest that during apprehension viewers can extract spatial, categorical as well as semantic information, i.e., the “gist” of the scene, *rapidly*, within only a few milliseconds (Henderson & Ferreira, 2004). In addition, studies in visual perception have suggested that apprehension is a *flexible* process. Visual perception can be modulated by top-down factors such as task demands (Henderson, 2003; Yarbus, 1967), attention (e.g., Treisman, 2006) and cultural backgrounds (e.g., Senzaki, Masuda, & Ishii, 2014).

The current study aims to explore the effects of two top-down factors, i.e., task demands and language, on event apprehension. First, we manipulate task demands in different linguistic description tasks, aiming to tap into the relation between apprehension and linguistic formulation in language production. Second, we compare apprehension in viewers with different language backgrounds. This factor is included to shed light on the role of cross-linguistic differences for core language processing theories (Jaeger & Norcliffe, 2009). Section 1.1 reviews theories and previous eye tracking studies that target the early phases in language production. Section 1.2 introduces the potential role of cross-linguistic differences on pre-linguistic message planning, and Section 1.3 discusses how the rapid apprehension process can be captured in experimental designs, and how it can provide further insights of language production.

1.1 Message encoding in language production: the starting point debate

One of the central debates concerning message encoding theories in language production is the "starting point question" (cf. Bock, Irwin, & Davidson, 2004; Bock et al., 2003; Gleitman et al., 2007; Konopka & Brown-Schmidt, 2014): During the process of linguistic formulation (i.e., grammatical and phonological encoding), following the conceptualization phase, conceptual knowledge has to be turned into linguistic forms. A linearization process must turn conceptual representations into a string of linearly ordered words, which means that a starting point has to be selected. Importantly, the starting point is a critical link between the preverbal message and the incremental formulation process, as it constrains both the content as well as the subsequent linguistic structure for the planning of utterances (Bock et al., 2004; Bock et al., 2003; Levelt, 2000).

However, it is notoriously difficult to investigate the "starting point question" in psycholinguistic experiments, because a measurement with a high temporal resolution and careful control of stimulus content is required to be able to isolate the message encoding phase from the consecutive processes. A reliable method is eye tracking, where eye movements can reflect cognitive processes in visual and linguistic processing (Griffin, 2004). Adopting eye tracking techniques, previous studies have put forward two alternative hypotheses accounting for the relationship between eye gaze and the selection of the starting point in language production processes (for a review see Bock et al., 2004; Bock & Ferreira, 2014; Konnopka & Brown-Schmidt, 2014).

The first hypothesis, *the linear incrementality account*, argues that the scope of message planning in preparation for linguistic formulation is linearly incremental in a "word by word" fashion. Speakers start building their utterance already after the conceptual and linguistic encoding of the initial starting point. In a visual environment, starting point selection is assumed to be mainly saliency-driven (which can also be interpreted as the "importance" or the "easiness" for processing, Bock & Ferreira, 2014) : speakers' eye gaze tends to be initially attracted toward the most perceptually salient element in the visual scene (e.g., an element that can capture the attention due to certain features, e.g., color, size, animacy etc.). The element that is fixated first will be anchored as the starting point of an utterance. The most convincing piece of evidence for this account comes from the experiments by Gleitman et al. (2007): A perceptual cue, which was briefly exposed (60-80ms) and was hardly noticed, was presented just before the visual stimuli that depicted various events (e.g., a dog chasing a man; two men shaking hands). The cue was designed to bias attention toward a specific event role (e.g., the dog), in order to test whether cueing certain visual elements to attract eye gazes could predict the starting point in sentence formulation. The eye tracking data and speakers' verbal production showed a clear pattern: perceptually cued elements were more likely to be fixated first as well as to be mentioned first in the event descriptions (e.g., if the perceptual cue appeared on the location where the dog would be shown in the ensuing picture, speakers were more likely to fixate on the dog first and utter "the dog is chasing the man" rather than "the man was chased by the dog."). The study suggests an impact of initial visual attention on the selection of the starting point of a sentence, which is taken as a support of a linear relationship between message encoding, subsequent eye movements and linguistic encoding: *where* people look first correlates with *what* is mentioned first. The "linguistic representations are *immediately* triggered" from the visual input, and an apprehension process during which an overall scene structure is extracted does not need to take place before linguistic formulation (Gleitman et al., 2007).

The alternative account, known as *the hierarchical incrementality account,* argues that message planning must include not only the starting point itself, but also a plan on how to proceed from this point. In other words, there should first be a phase, an apprehension phase, during which a rudimentary plan of the relational and structural information in the visual input is constructed. This "plan" guides the first fixation and determines the starting point of a sentence. In other words, the visual element that is fixated first does not directly *decide* the starting point, but it *reflects* the result of apprehension (e.g., Bock et al., 2003). The starting point of the to-be-produced utterance, then, does not necessarily correspond with the element that captures the initial eye gaze. For instance, Griffin & Bock (2000) recorded speakers' eye movements and verbal descriptions on line drawings of transitive events (e.g., a mailman chasing a dog). They observed that the first fixation location, registered within 400ms of stimulus onset, did *not* predict the starting point of the verbal descriptions (e.g., the first fixation did not always land on the mailman in the stimulus, when the utterance was "a mailman is chasing a dog."). This implies that, the starting point is not purely driven by visual saliency of certain elements in a scene. An apprehension phase should precede linguistic formulation and has to be finished in a very short time span during which speakers first encode the structural relationship in the event (e.g., who is the agent/patient**).** This "holistic process of conceptualization" of apprehending an event's gist guides the allocation of the first fixation, and later linguistic formulation processes (Griffin & Bock, 2000). Under this account, message encoding in language production should be tightly interrelated with apprehension.

However, amongst these studies that were the first to target the early phases in language production, little consensus has been reached on the two alternative accounts: Gleitman et al. (2007) also examined first fixation locations but found that the element that was mentioned first, was also the region that was fixated first within 200ms, even in the un-cued condition where no perceptual cue was presented prior to the stimuli. This result was contradictory with the data from Griffin & Bock (2000) where first fixation locations did not predict sentential subjects, indicating that linguistic representations are not "immediately" triggered by visual input. In sum, the mixed findings so far targeting the relationship of apprehension and starting point cannot clearly distinguish the two hypotheses (linear or hierarchical): Does a process of apprehending overall scene structures precede linguistic formulation?

One of the potential reasons that this question has not been answered to date lies in the fact that previous studies adopted a relatively long presentation duration of stimuli (i.e., free-viewing of stimuli during description, for 3-6 seconds), while the desired window for zooming into apprehension only lies within the initial 300 to 400ms. With free-viewing and longer exposure to visual scenes, researchers have less control over what exact processing phase they are tapping into, and what participants are doing within this initial phase. It is thus hard to control for the start of the actual language planning process, and to ensure that the moment of stimulus onset (and the first fixation) really tap into this process. A more suitable method for tapping into this process is a brief exposure paradigm, in which a stimulus is only presented for a few milliseconds. It does not only force participants to engage in starting the language production process immediately upon stimulus onset, but it also zooms into the rapid apprehension phase directly. A more detailed discussion of the brief exposure paradigm is presented in Section 1.3.

1.2 Cross-linguistic differences in language production

The languages spoken in the world differ widely in how they encode event segments in linguistic forms. Can message encoding theories derived from one language system be generalized into other languages? Studies have shown that language systems vary in what type of information must and must not be encoded in the message and expressed linguistically (e.g., Jaeger & Norcliffe, 2009), and this can impact the early message planning phase.

For instance, Myachykov et al. (2010) compared English and Finnish speakers using a similar paradigm as Gleitman et al. (2007). The eye tracking results were replicated in the English group where the first fixation locations were attracted by the perceptual cues and can predict the sentence starting point. However, this linear pattern between first fixations and starting points cannot be replicated in Finnish, a case-marking language (e.g., agent requires a Nominative case marker and patient with the Accusative marker): while the perceptual cue still captured the attention of the first fixation, it did not predict the starting point of the verbal descriptions for the Finnish group. Finnish speakers used SVO word order consistently, regardless of the position of the perceptual cue. The absence of the linear pattern between initial eye gaze and verbal descriptions in Finnish speakers implies that, in Finnish, the case-marking system may require a higher demand of on the messages encoded obligatorily in the early phase. Compared with English, which lacks grammatical cases, Finnish speakers need to assign case markers to nouns on the basis of event roles, which requires first an understanding of the event structure (i.e., who is the agent or the patient?). The structural information needs to be included within the apprehension process in order to allow the assignment of case markers onto the corresponding nouns in the later linguistic formulation phase, regardless of whether some elements are visually cued or not. Similar results were found in Korean, another case - marking language (Hwang & Kaiser, 2009).

Another piece of evidence supporting an initial extraction of structural information is from a set of eye tracking experiments conducted in verb initial languages, Tzeltal (Norcliffe et al., 2015) and Tagalog (Sauppe et al., 2013). Interestingly, the two languages require agreement markers on the initial verb to encode an argument's voice indicating whether the subject, which can be uttered in the middle or at final position in a sentence, is the agent or patient of the depicted event. In these languages, presumably, some structural knowledge must be obtained *before* the starting point, in order to decide initially which event role should be selected as the subject. Then , the corresponding verb and the appropriate agreement marker can be selected as the starting point of a sentence (Norcliffe & Konopka, 2015).

How would the eye movements of speakers of these languages differ from English speakers who prefer to encode the subject first in a sentence? Norcliffe et al. (2015) and Sauppe et al. (2013) adopted a similar design as Griffin & Bock (2000). Tzeltal, Tagalog and Dutch speakers viewed and described transitive event stimuli while their utterances and eye movements were recorded. The eye tracking data showed that in the early phase (0-600ms), fixations tended to be allocated toward the entity with the event role that was assigned as the subject of the sentence, which is preferably uttered at the final position in Tzeltal, and marked by a "Privileged Syntactic Argument (PSA)" in Tagalog. Here, early fixations did not correlate with the word orders following the initial verb: in both languages, the subject can be uttered at the sentence-final position (i.e., VOS), but the subject entity received early fixations (within 600ms). The eye tracking results from the studies on verb-initial languages further support that an apprehension phase preceding linguistic formulation is needed to extract rudimentary event structural information in order to decide the starting point for linguistic formulation.

These cross-linguistic studies indicate that perceptual saliency and first fixations do not directly correspond with the selection of the starting point of a sentence. In other words, these studies highlight that there is a phase preceding the formulation of the first word in the sentence, during which the information on the overall structure of an event is obtained, i.e., they are in favor of the hierarchical incrementality account. In addition, the specific language spoken by a viewer, varying in what must be explicitly encoded (e.g., case markers or PSA), may also modulate apprehension, as early fixation patterns differed between English and Finnish speakers in Myachykov et al. (2010), and between Tagolog and Dutch speakers in Sauppe et al. (2013). However, the early fixation data in these studies were obtained from a free-viewing paradigm. It is thus still unknown to what extent cross-linguistic variation can impact the apprehension process, which can happen within the allocation of the first fixation.

1.3 How to isolate the apprehension process: brief exposure paradigms

Scene apprehension is a rapid process that may even happen without a fixation, which, in experimental research, needs to be captured by a method with a high temporal resolution. A brief exposure paradigm, in which the duration of stimulus exposure is narrowed down to only a few milliseconds, can tap into the apprehension process. Brief exposure paradigm is a useful supplement to the eye tracking studies allowing free-stimuli viewing while speaking (Dobel et al., 2010).

Hafri et al. (2013) claimed that the apprehension of event roles and actions can even happen within 37ms. Stimuli depicting two-participant actions were briefly presented for only 37ms or 73 ms, after which English speakers were asked to answer explicit questions on event roles and actions (e.g., answering “did you see kicking?”, “is the girl performing?”). The results showed that viewers can already extract categorical as well as relational information of events within the shortest exposure condition. However, it is noteworthy that the experiments in Hafri et al. (2013) involved sentence comprehension, as participants were required to answer explicit questions containing information on event structure; this could have helped them in “filling in” what they had retrieved visually (e.g., a question such as “Is the blue boy being acted upon?” suggested that the event had involved a patient-role and that the boy could have served this role). Thus, although stimulus exposure (e.g., 37ms) is astonishingly short, it cannot be concluded that the information questioned is derived entirely from visual processes.

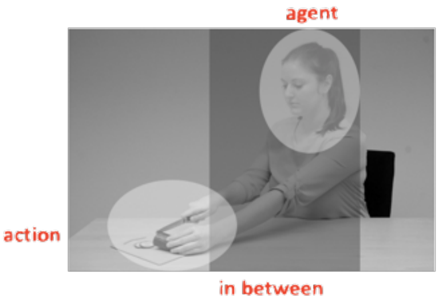
A brief exposure study that more plausibly captured the apprehension process is by Dobel et al. (2007). They presented coherent or incoherent transitive action scenes for 100 to 300ms (e.g., In a picture depicting “a hunter shoots an elephant with a bullet in between the two actors”, the scene was coherent when the hunter and the elephant face each other, but incoherent when the hunter was mirrored and faced back to the elephant). German participants were asked to judge scene coherence and to describe the scene by naming the agent, patient or the action. The verbal description data suggested, surprisingly, that participants could already accurately identify scene coherence in the shortest 100ms exposure condition. Within 200ms, they were also able to identify and name event actions and roles to a great extent (with an accuracy of 75% in agent naming). Involving event naming tasks in the brief exposure paradigm, Dobel et al. (2007) suggests that the apprehension of event structure can happen within 200ms. Bock et al. (2003) also adopted the brief exposure paradigm to investigate time expression across languages. In one condition, Dutch and English speakers described time on a clock that were presented for 100ms. Although 100ms is too short to plan and launch a fixation on the stimulus, speakers were fairly accurate in describing the time on the clocks. In addition, Bock et al. (2003) included a condition with an exposure duration of 3000ms. The eye tracking data suggested that early fixations did not predict the number that was uttered first for time naming. Bock et al. (2003) argues that sufficient information can be extracted in the initial saccade, which is responsible for directing the eyes toward a location where the information is needed for planning the utterance.

However, offline measurements alone (i.e., linguistic descriptions) cannot precisely dissociate apprehension from further linguistic formulation, as the only measurement is participants' final linguistic product. Online measures, such as eye tracking, are still needed to approach apprehension more directly. Gerwien & Flecken (2016) combined the brief exposure paradigm with eye tracking to examine the top-down effects of stimulus exposure durations and language backgrounds of viewers on apprehension. German and Spanish speakers described events in a full sentence after being exposed to causative event stimuli for 300, 500 and 700ms. German and Spanish vary in their prominence in the encoding and conceptualization of event agents and event actions: while German speakers emphasize event agents when conceptualizing events (Flecken et al., 2015), Spanish speakers, allowing subject omission for a sentence (i.e., pro-drop), tend to be action-oriented (Fausey & Boroditsky, 2010). The question is, to what extent the prominence of different event elements (i.e., variations in agent saliency) in the two languages can affect apprehension towards the corresponding visual scene.

Importantly, Gerwien & Flecken (2016) focused on *first fixation locations* on the visual stimuli to explore apprehension, which is considered the very first sign of overt attention allocation (also see Bock et al., 2003). They identified three main areas of interest (AOIs): Agent, Action and In-between AOIs (see Figure 1.1) and recorded the proportion of first fixations in each AOI. The results of first fixation locations and verbal descriptions showed that, first, when the stimulus exposure duration increased from 300ms to 700ms, the proportion of the first fixations locating in the Agent AOI increased, indicating that the time available for scene viewing can affect where people locate their eyes first. Second, a great proportion of first fixations did not predict the starting point of a sentence: only about 40% of the first fixations located on the agent AOI, while the verbal descriptions in the two language groups encoded the event agent exclusively as the starting point, i.e., the subject of the sentence. This result licensed first fixation locations as a direct online measure on apprehension, which can be isolated from the linguistic formulation phase: as the result of apprehension, first fixation allocation happens prior to linguistic formulation, since the first fixations and the starting point of utterance are not interrelated.

Regarding speakers' language backgrounds, the cross-linguistic differences between Spanish and German speakers were only found in the 300ms condition, but not in the 500 or 700ms condition. Within 300ms, where only one fixation can be registered, the Spanish group allocated more first fixations in the "In-between" AOI compared to German speakers, who first fixated more towards the Agent AOI. It is also noteworthy that Spanish speakers did not utter subject omission sentences in the experimental setting. The difference of speakers' language backgrounds on first fixation locations was interpreted as the impact of agent saliency on the conceptualization of event structures. Given the limited exposure time towards visual stimuli, the two language speakers choose their starting point differently: While the agent is preferred as the starting point in German (agent-oriented), Spanish speakers, possibly due to the flexibility in subject omission, tend to fixate on a location between the agent and the action AOI, in order to retrieve both event structural information. The in-between fixation pattern indicates a weaker emphasis on the agent compared with German speakers. However, this interpretation is not straightforward and requires further research (Gerwien & Flecken, 2016).

Figure 1.1 Example stimulus in Gerwien & Flecken (2016) with three Areas of Interests: Agent AOI (actor's face), Action AOI (actor's hands and the object), and "In-between" AOI (the dark grey area in between the Agent and Action AOI).



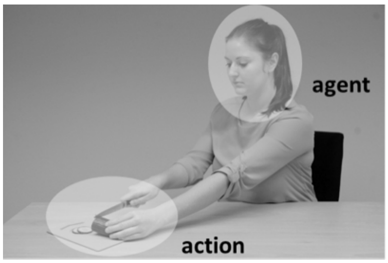
In addition, as all the participants only performed one description task (i.e., describing the picture in a full sentence), it is unknown whether the first fixation pattern in Gerwien & Flecken (2016) is driven by the specific task demand, i.e., an event description task, or whether it is a fixed pattern of first fixations in scene apprehension. In other words, it is unknown whether apprehension is a rigid process for each language speakers.

The current study extends the study by Gerwien & Flecken (2016) to further our understanding of the potential top-down effects, task demands and language backgrounds, on apprehension. Chapter 2 presents an overview of the aims of the present study. Chapter 3 and Chapter 4 report the methods and results of the two experiments we conducted. Finally, Chapter 5 discusses the present study within the larger context of language production and perception.

2 Aims of the present study

The current study investigates the top-down effects of linguistic task demands and language backgrounds on event apprehension, as a window onto the early phases of the language production process (Levelt, 1989). Apprehension is the rapid visual process of extracting the gist of a scene. The eye tracking methodology was used to capture this process. We employed real-world photographs of causative events (e.g., an agent performing an action on an object; a woman cutting a cucumber) as stimuli, allowing a clear spatial dissociation of the two main event elements: the upper half area encompassing the agent performer and the bottom half area depicting an action and the affected object, as exemplified in Figure 2.1.

Figure 2.1 Example stimulus that contains two distinct areas for two event elements: the agent area, locating on the upper half of the stimulus, and action/object area, locating on the bottom half of the stimulus.



Extending the cross-linguistic comparison from Gerwien & Flecken (2016), we compare Mandarin Chinese and Dutch which differ in the flexibility of the encoding of the subject of a sentence (i.e., the agents in our event stimuli). In Mandarin Chinese, omitting the subject is frequently allowed if sufficient contextual information is given (Li & Thompson, 1981)[1]. For instance, in answering a question from a conversion "Do you know Tom?", there are four options for a positive answer in Mandarin: 1) "I know Tom", in which the subject and the object are explicitly encoded, 2) "I know __.", in which the object "Tom" is dropped, 3) "__know Tom.", in which the subject "I" was omitted, and 4) "__know__.", where both the subject "I" and the

[1] It is noting that Mandarin also allows the omission of the object if sufficient context is given, however, our design in the present study does not license an object-drop context for Mandarin speakers. It is because in the encoding of causative events in Mandarin, verb and its object is tightly related and can become collocations: e.g., in Mandarin, "Xi-Pai", a verb-noun expression, means to shuffle cards, but the single verb "Xi" itself means "to wash", and "Pai" means the cards. The action of shuffling cards cannot be expressed if lacking any of the two elements. In Mandarin, an event action is not strictly represented by the verb, but also requires the object, which, thus, cannot be dropped freely.

object “Tom” are eliminated. All the four choices are grammatical and unambiguous as the context is sufficient to suggest the referent that is omitted. In addition, unlike alphabetic languages, Mandarin Chinese lacks verb inflections that encode person information, which means that the omitted information has to be retrieved from the context, and cannot be derived from the predicate, unlike other pro-drop languages such as Spanish (Hsiao, Gao, & MacDonald, 2014). By comparison, Dutch typically encodes agent information explicitly in the subject of a sentence. This cross-linguistic variation of the flexibility of subject encoding offers a contrastive case to explore whether the linguistic variation can result in perceptual differences in event structure, namely agent-saliency, during apprehension (see below for more details).

We measure and analyze first fixations, which we consider as the very first overt sign of attention allocation as a result of apprehension (Gerwien & Flecken, 2016). In a visual context, it is very difficult to disentangle apprehension from the message encoding process per se, as the two processes are presumably tightly interrelated, at least in language production tasks (also assumed in e.g., Dobel et al. (2007); Bock et al. (2003) and Bock et al. (2004), etc.). However, what can be disentangled is the relation between apprehension and linguistic formulations: First fixations**,** as the reflection of the result of the apprehension process, should precede linguistic formulations, which is evidenced by the fact that FFLs did not predict the starting point of the verbal description of an event scene (Gerwien & Flecken, 2016). In the present study, we are particularly interested in the location that a first fixation is registered on a visual scene after stimulus onset, namely, *First Fixation Locations* ("FFLs" below), to shed light on the relationship between apprehension and linguistic formulation.

In order to isolate the first fixation, and thus to target the apprehension process directly, a brief exposure paradigm is adopted. Native Dutch and Mandarin speakers are exposed to the visual stimuli for only 300ms. An exposure time of 300ms allows the participants to launch and place only one fixation on the stimulus at most (Gerwien & Flecken, 2016). The stimuli are presented randomly in one of the four corners of the screen, and the orientation of the agent (i.e., agent on the left or right side of a picture) is pseudorandomized, in order to prevent strategies that can predict a stimulus’ location. After brief exposure, participants have to verbally describe different event elements in different linguistic tasks. Participants’ eye movements, with a focus on FFLs, as well as their verbal responses are recorded, which capture the apprehension process using both online and offline measurements.

Two experiments are conducted using this brief exposure paradigm. In Experiment 1, four tasks are designed (see details in 3.1.2). Each participant is randomly assigned to three tasks, in three blocks. The tasks include a Nonverbal task (indicating whether a stimulus have been presented before), an Event description task (describing what is happening in the picture using a full sentence), and Agent or Action naming task (naming the agent or the action/object element in the stimuli). Detailed instructions to these tasks are given before each block. It is expected that the different task demands render different foci of attention towards the specific elements of the events depicted, i.e., the Agent and/or the Action/object element.

Experiment 2 adopts the same brief exposure procedure, and also employed the Agent, Action and Event description tasks in different blocks. In addition, we explore viewers' *memory* of the agent in the event scene to further compare the potential cross-linguistic differences in agent saliency, and we ask to what extent one's memory of the agent in a causative event is influenced by fixations under brief exposure, and by explicit linguistic encoding on certain event elements. The agent memory is tested after the Action Naming Task and the Event Description Task. Participants perform a surprise Recognition Memory Task in which they choose which picture they have seen before, amongst two alternatives that only differ with respect to the agent.

The present study adopts a novel and innovative measurement as the dependent variable in the analysis, namely, the Y-coordinates of the FFLs on the vertical dimension of the stimuli (see 3.2.2 for details). Previous studies analyzed fixation locations mainly by looking at the proportion of fixations in certain Areas of Interests (AOIs), which, however, are typically manually defined (e.g., Griffin & Bock, 2000). Analysis of AOIs can become problematic if fixations are placed on an undefined area, or if the fixations are not allocated accurately and precisely within the bounds of an AOI. For instance, Gerwien & Flecken (2016) defined an "In-between" area in the middle of the tested pictures, excluding the Agent and Action/object AOIs, which was fixated frequently under brief exposure among German and Spanish speakers. However, the boundary distinguishing the "In-between" AOI from the Agent or Action AOI was randomly defined (see Figure 1.1). In addition, given the high demands of brief exposure, participants may not always be able to locate their fixations precisely on the intended location. Rather, FFLs suggest the best attempts at fixating an intended location that a speaker can achieve within the time constraints given. Thus, a continuous dependent variable to analyze FFLs is more informative to observe attentional preferences.

Analysing the Y-coordinates of the FFLs as a continuous dependent variable, can avoid the potential problems caused by manually defining AOIs. In addition, it simplifies our handling of the stimulus-position variance that we introduce in the experiment. That is, regardless of agent orientation (on the right or left of the stimuli), the vertical layout of the event elements is consistent, with the agent element in the upper half of the stimuli, and the Action/object element in the bottom half of the stimuli, and this is reflected in the Y-coordinates of FFLs[2].

Two research questions are addressed: The first research question concerns the influence of task demands on apprehension. Depending on the task, one or more of the event elements is required to be focused and mapped onto a linguistic representation: The agent element is likely to be focused during the Agent naming task. Similarly, the Action/object elements are likely to be attended during the Action naming task. In addition, all the elements are relevant for the Event description task: The agent element will be encoded as the subject of a sentence, the action depicted will be mapped onto the predicate (the verb), and the patient element (i.e., the object in the event) will be encoded as the object of a sentence. The four tasks designed in the study aim to elicit verbalizations that require different foci of attention on these event elements for language production. The research question is, whether apprehension is influenced by the different linguistic task demands that focus attention on different event elements.

The second research question concerns the effect of language background of the viewer on apprehension. In linguistic theory, the frequent subject omission in Mandarin Chinese is

[2] The concern that only engaging the Y-coordinates of FFLs may include fixations that locate on the blank areas should be ruled out, because it is also known that fixations tend to cluster around informative areas of a stimulus, and it is rare for people to fixate on the blank area of a stimulus, such as the blank areas around the agent and action shown in Figure 2.1 (cf. Buswell, 1935). Thus, it is highly unlikely that the Y-coordinates would reflect fixations that were *intended* to land on the blank spaces in the stimuli. Rather, a FFL registered on e.g., the upper part of the screen will reflect a fixation that was launched in the direction of the agent's face in the stimuli. Appendix 3 also depicts a scatterplot that directly plots the recorded fixations in an absolute x-y dimension, as a side evidence to support that only Y-coordinates of the FFLs are sufficient to indicate fixation patterns in our design.

assumed to contribute to its nature as being a topic-prominence language (e.g., Huang & Yang, 2013; Paul, 2017). However, what is less known is whether cross-linguistic variation can also affect the conceptualization of event structure. Mandarin speakers do not have to rely on explicit linguistic encoding to refer to agents in events, which means that reference to the agent needs to be tracked implicitly but may also happen more carefully compared with Dutch, a language encoding the subject obligatorily. It is noteworthy that similar to the Spanish group[3] in Gerwien & Flecken (2016), the experimental design of the present study does not provide enough context for Mandarin speakers to actually produce a pro-drop expression, as participants were instructed to formulate one sentence only. What is interesting for the present study is whether the habitual use of pro-drop for Mandarin speakers could affect their conceptualization of events and their first fixation locations. The research question is, whether the cross-linguistic differences in pro drop between Mandarin and Dutch can influence the early apprehension of event structure.

We outline two hypotheses. First, we hypothesize that FFLs can be modulated by the demands of the different production tasks employed. If apprehension is a flexible process and FFLs are the result of the apprehension process in which the first overt attention is allocated towards the most informative region for the task at hand, the distribution of the FFLs should be centered around different event elements, under different task demands: For the Agent naming task, FFLs should cluster around the upper region of the stimuli, closer to the actor’s facial area, whereas in the Action naming task, FFLs should be targeted more towards the lower half of the stimuli, closer to the action/object depicted. Alternatively, if FFLs show similar patterns across different tasks, apprehension will be a rigid process, during which various foci on event elements do not influence the initial fixation pattern in a visual scene.

Second, we aim to use the FFLs patterns in Event description task, in which participants were required to describe the stimuli in a full sentence, to shed light on the “starting point” debate. Two alternative outcomes are possible, based on the two accounts for the “starting point

[3] Subject omission in Mandarin is different from Spanish in that there is no verb inflection or any other marking system to help the speakers to retrieve reference on the person information. They have to track the omitted information from the context in order to “check” whether an agent is continued across events. So, the direction of the hypothesis for the effect of pro-drop on apprehension (i.e., whether the agent or the action element is more focused) is not necessarily aligned with the result for Spanish in Gerwien & Flecken (2016).

question” (for a review, see Section 1.1): First, if the linear incrementality hypothesis is true, meaning that initial fixations are saliency-driven and their transition to linguistic representations is immediate (Gleitman et al., 2007), FFLs would cluster around the upper half of the stimuli, i.e., the *agent* element, because the event descriptions in both Dutch and Mandarin are dominated by subject-first word order. The subject, i.e., the agent element in our stimuli, should be apprehended and formulated first in sentence production. Alternatively, if the hierarchical incrementality hypothesis is true, meaning that a holistic conceptualization on event structure is set first to guide later linguistic formulation processes (Griffin & Bock, 2000), FFLs will not necessarily cluster around the agent element, but rather towards the region in between the agent and action/object elements, enabling the extraction of both agent and action/object information. Gerwien & Flecken (2016) reasoned that this pattern reflected speakers’ attempt to extract the entire event structure.

Furthermore, the potential cross-linguistic effect of subject omission on apprehension should be considered exploratory, given that there are no prior studies analyzing the Y-coordinates of FFLs as an index of apprehension. Based on previous studies (e.g., Gerwien & Flecken, 2016; Norcliffe et al., 2015; Sauppe et al., 2013), a potential outcome is that the two language groups would differ in their FFLs patterns in the Event description task. If pro drop affects apprehension in a similar pattern found in Gerwien & Flecken (2016), where Spanish speakers fixated more on the “In-between” AOI, the Y-coordinates of the FFLs for Mandarin speakers may be closer to the action/object element compared to the Dutch group. However, if the effects of pro-drop in Mandarin follow theories on topic-prominence (e.g., Huang & Yang, 2013), the pattern would show that FFLs cluster closer to the agent element compared to Dutch speakers. Another alternative hypothesis is that there is no differencce between languages, which would suggest that there is no effect of pro-drop on apprehension.

Another hypothesis concerns the memory task in Experiment 2. We expect an effect of task demands on the accuracy of memory of the agent. If explicit agent encoding can enhance memory of the agent element, memory in the Event Description Task should be better compared to the Action Naming Task. In addition, we explore to what extent cross-linguistic differences in pro drop may affect agent memory.

By analyzing FFLs in different language production tasks and in Dutch and Mandarin speakers, the study provides various insights in language production theories, as well as in the relation between visual and linguistic processing more generally. First, we will shed light on whether the FFL registered in a brief exposure paradigm (300ms) is an appropriate index for the apprehension process. If so, an effect of task demands can provide evidence for the flexible nature of apprehension that can be modulated by the top-down factor. Second, the FFLs patterns in the Event description task under the brief exposure paradigm directly disentangle apprehension from the linguistic formulation phase in language production, which will add value to the debate of the “starting point question”, namely, whether the relation between apprehension and linguistic formulation is linear-ordered and saliency-driven, or whether an overall structural conceptualization of the event message is needed before deciding on a starting point for formulation processes. Third, the cross-linguistic comparison on pro drop between Mandarin Chinese and Dutch will further test to what extent language production theories can be generalized or varied given the cross-linguistic variations.

3 Experiment 1

3.1 Method

3.1.1 Participants

The Dutch group included 30 participants recruited from the participant pool of the Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands (mean age = 28.53, SD = 14, male N= 7 and female N= 23). All participants were students at Radboud University. Out of the original group, six participants had to be excluded due to technical errors. The final Dutch group consisted of 24 participants.

The Chinese group included 26 participants recruited from Radboud University Nijmegen in the Netherlands (N= 18) and Heidelberg University in Germany (N=8) (mean age = 26.67, SD=2.38, male N=12, female N=14). Chinese participants were international students and employees currently enrolled at Radboud University or Heidelberg University. Out of the original group, three participants had to be excluded due to technical errors. The final Chinese group consisted of 24 participants.

All the participants had normal or corrected-to-normal vision. All the participants received a payment of 6 euros.

3.1.2 Task and List Design

In total, Experiment 1 consisted of four tasks, varied across blocks:

Non-verbal task: In this task, participants were instructed to say “yes” when they saw a picture that had been shown in previous trials.

Agent Naming task (“Agent task” below): In this task, participants were instructed to name the actor aloud when they saw a photo that was performed by one of the actors they had been introduced to at the beginning of the block.

At the beginning of this task, four actors’ names and their photos were introduced to the participants by the experimenter. To ensure they memorized the agents, a picture naming test was conducted in which participants had to write down the names of the agents under the respective photograph. The eye tracking task would only commence once the naming was correct.

Action Naming Task ("Action task" below): In this task, participants were instructed to describe the action in the picture only, e.g., "cut a cucumber".

Event Description Task ("Event task" below): In this task, participants were instructed to describe what happened in the picture using a full sentence, e.g., "a girl is cutting a cucumber."

To counterbalance the sequence of tasks, the experiment included four lists with different combinations and ordering. Each list was assigned to an equal number of participants (i.e., N=6 for each list and each language group). Each list contained three blocks: All the four lists contained the Non-verbal task as well as the Event task in two of the blocks, while the Non-verbal task always appeared as the first block. Half of the group performed the Action Naming Task and the other half performed the Agent Naming Task (i.e., N=12 for each task in each language group). Task sequence in the second and the third block was randomized (See Table 3.1)

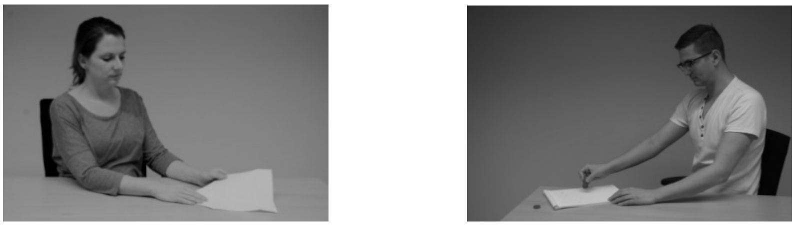
Table 3.1. List and block design for Experiment 1 (N=6 for each list and language group).

	List 1	List2	List 3	List 4
Block 1	Non-verbal Task	Non-verbal Task	Non-verbal Task	Non-verbal Task
Block 2	Action Naming Task	Agent Naming Task	Event Naming Task	Event Naming Task
Block 3	Event Naming Task	Event Naming Task	Agent Naming Task	Action Naming Task

3.1.3 Materials

The critical stimuli were photographs in black and white colors, shot for the purpose of this study at the Max Planck Institute for Psycholinguistics. In total, 48 causative event photographs depicted four actors (3 female, 1 male) performed actions on objects. Each task contained 16 stimuli, in which the stimuli for Agent and Action Naming Tasks were identical, given that each participant only performed one of the two tasks. Among the 16 stimuli in each task, half of them were presented in agent right orientation, half in agent-left orientation (see Figure 3.1 as an example). In addition, each block also included 16 filler pictures depicting a stative scene (e.g., a jar of coffee beans on a table; a person standing next to a tree). The sequence of the stimuli presented in each task was randomized.

Figure 3.1 Example stimuli with performers of different genders agent-left or agent right orientation. A full list of the content of the critical stimuli is attached in Appendix 1

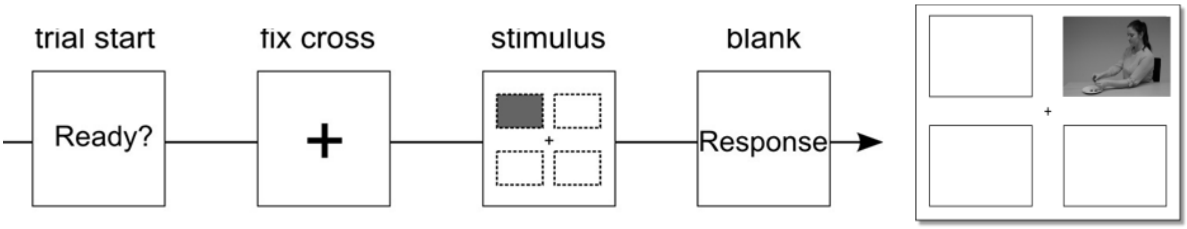


3.1.4 Procedure

The participants were asked to sign the consent form first. They were then asked to sit still in front of the remote SMI RED250m Eye Tracker (SensoMotor Instruments, sampling rate 250 Hz) at a distance of approximately 65 cm. The eye tracker was attached to the lower part of a laptop with a display resolution of 1920*1080. A masked webcam was attached for audio recordings. The experiment was run on the software package Experiment Center, which controlled the eye tracker, the presentation of the stimuli, button presses and speech recordings for the experiment.

Four point calibrations were performed four times throughout the experiment in a semi-automatic fashion: the first calibration was presented at the very beginning of the experiment, and the other three calibrations were performed before each task, after task instructions were presented.

Figure 3.2 Trial procedure (left) and stimulus display (right). The frames and the fixation cross were not presented in the experiment.



Participants were guided by a native language experimenter (Dutch or Mandarin Chinese) and the written instructions were also presented in their native language. The instructions of the four tasks explicitly aimed at eliciting the required utterances (naming the agent, action or the whole event). Each participant performed the assigned list and the corresponding tasks. The experimental session lasted approximately 30 minutes.

Each trial started with a fixation cross presented in the center of the screen, which the participants were required to fixate. The stimulus would only appear if a fixation on the cross was registered. Each photo appeared (pseudo-)randomly in one of the four corners of the screen for 300ms. This exposure time guaranteed that participants had sufficient time to plan, launch, and place one fixation. The order of the stimuli and their presentation location on the screen were randomized together with the filler pictures, in order to prevent the participants from predicting the location of the stimuli that would appear, and the content to be uttered. The number of agent-left and agent-right photos was counterbalanced within each task in order to counterbalance a left-to-right preference in scene perception (Buswell, 1935)**.** After stimulus offset, a blank screen was shown where the participants uttered aloud the required information in their native language. Participants could proceed to the next trial by pressing the space bar, indicating that they had finished the current trial (see Figure 3.2 depicting the trial procedure).

3.2 Data preprocessing, coding and analysis

3.2.1 Verbal production data: description accuracy and specificity

Experiment 1 analyzed the verbal production data in the Action and Event Tasks, but not in the Agent task. Previous work has suggested that the successful identification of an agent happens rapidly (e.g., identifying the man when apprehending a picture depicting “a man shoots an elephant”), within 200ms of stimulus exposure, and performance does not seem to further improve with longer stimulus exposure (maintained at an accuracy of 75% in Dobel et al., 2007). Our design used a stimulus exposure of 300ms (i.e., above 200ms), plus a straightforward agent naming task, which involved identifying the only animate component in the stimulus. This ensured that agent naming performance was at ceiling, and thus is not of the main interest for the present experiment.

By comparison, in Dobel et al. (2007), the accuracy of action and patient recognition (e.g., identifying the action of “shooting” and the patient “elephant” in the previous example) is not as good as agent recognition: The accuracy maintains around 60% for patient identification and only 46% for action identification, with 300ms exposure, which suggests that naming the action and patient given the brief exposure duration may require a more comprehensive understanding towards the event structure (e.g., identifying the agent at the first place). The accuracy difference in event roles and action identification observed in Dobel et al. (2007) motivates our study to focus on the performance in the Action and Event tasks, which involves action and patient naming. But still, we will analyze first fixation locations in the Agent naming task to explore the effect of task demands on event apprehension.

Data coding

The utterances that were recorded in the Action and Event Tasks were transcribed by a native Dutch and Mandarin Chinese speaker respectively. The transcribed data were then coded based on two criteria separately: the accuracy of the overall response, and the specificity of the reference to each event element. The coding was carried out by the same Dutch and Mandarin Chinese native speakers. Ambiguous cases only existed in a few cases and they were solved after discussion with a third researcher.

For the accuracy coding, the responses were marked as "correct" if they correctly and concretely represented the content of the event stimuli, and answered the question posed in the specific task. The rest of the utterances with mismatched event contents were marked as "incorrect" otherwise. Answers indicating a failure of capturing the content (e.g., No idea/I did not see it clearly, etc.) were marked as missing data.

The specificity of each event element was coded: The Agent element was coded as "specific" if the utterance in relation to the agent was gender specific (e.g., *a man/ a woman*), and "unspecific" if the reference was gender neutral (e.g., *someone/a person*). The Action element was "specific" when the utterance contained a concrete action verb (e.g., *to cut, to paint*), compared with an "unspecific" action verb (e.g., *to do, to hold*) or stative verbs (e.g., *to sit at a table*). Similarly, the Object references were coded as "specific" if the utterance mentioned the concrete item (e.g., *a cucumber, a bottle*) and as "unspecific" if the object was described generally (e.g., *something*) or not mentioned at all. Similar to Accuracy coding, answers indicating a lack of capturing any relevant content were marked as missing data.

Analysis

Accuracy and the specificity of each event element was analyzed separately using logistic mixed effect regression with R (version 3.4.2) and package *lme4* (Bates et al., 2015). The fixed factors were *language* (Dutch and Chinese), *task* demands (Action and Event task) and their interaction. Both factors were treatment coded. Random factors in the regression model followed the maximal structure justified by the design, which considered the random intercepts for *participant* and *stimulus*, as well as a by-*participant* random slope for the effect of *task.* The analysis was run after excluding the missing data.

3.2.2 Eye movement data: First fixation locations

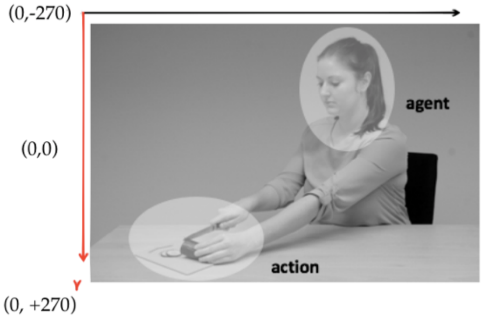
Data preprocessing

Participants' fixations were computed and tracked online with SMI BeGazeTM software, adopting a "two-pass" saccade detection algorithm (Holmqvist et al, 2011, p173). The data was computed twice based on 1) the velocities to detect saccades and 2) the onset and offset of the saccades. Fixations are typically assumed and identified when the detected event is not saccades or blinks.

We were primarily interested in the FFLs and the corresponding latency of fixation projections. The FFLs were defined following Gerwien & Flecken (2016), which refers to the first eye gaze registered by the eye tracker after the stimuli onset. Each fixation location was registered in the eye tracker as a pair of X- and Y- coordinates, together with the latency of the fixation projection after stimulus onset. Only the data for the right eye were analyzed.

Because the locations of agent and action/object were mirrored and randomized across trials, the analyses focused on the Y-axis only. On a screen with a 1920*1080 resolution, pictures were shown either on the upper half (a y-coordinate smaller than 540) or bottom half (larger than 540) of the screen. Fixation locations were transformed as to fit onto the same dimension by subtracting 540 if the stimuli were shown on the bottom part of the screen. Data were then centered by subtracting 270 pixels, i.e., the origin of the y-axis was the midline of the vertical dimension of a stimulus.

Figure 3.3 Data transformation: the Y-coordinates of FFLs were centered by moving the original point to the horizontal midline of the stimuli. A Y-coordinate below zero indicated that the first fixation located on the upper half of the picture, i.e., closer to the Agent element. A Y-coordinate above zero stood for a first fixation locating on the lower half of the picture.



A y-coordinate that was smaller than zero suggested that the FFL was located on the upper part of a stimulus, which was closer to the agent in the stimuli (i.e., head and upper body). Similarly, a y-coordinate that was larger than zero represented that the FFL was located on the lower part of a stimulus, which was closer to the area depicting the action and the object (See Figure 3.3 as an example)**.** All following analyses were based on the transformed Y-coordinates.

Data points were excluded on the basis of the following criteria: First, first fixation latencies smaller than 150ms were excluded, as they cannot have been launched upon stimulus onset (Holmqvist et al, 2011) and may be caused by technical error. In total, 140 data points (7.55% of all data) were excluded based on this criterion.

Second, the number of registered first fixations for each participant that was lower than 60% of the total trials were excluded (i.e., less than 28.8 trials of registered first fixations). One Dutch participant (with 22 recorded first fixations) was excluded in this step. In total, 23 Dutch and 24 Chinese participants were included in the final analysis.

Analysis

FFLs on the y-axis were analyzed using linear mixed effect regression models. The fixed factors were *language* (Dutch and Chinese), *task* demands (Non-verbal, Agent, Action and Event task) and their interaction (sum coded). Random factors in the regression model followed the maximal structure that justified the design, which contained random intercepts for *participant*, *stimulus* and *picture locations* (i.e., stimuli showing on the upper or lower half of the screen)[4], as well as a by-*participant* random slope for the effect of *task.*

3.3 Results

3.3.1 Verbal production data

Accuracy

Table 3.2 reports the accuracy of the verbal output, and Table 3.3 reports the results of the logistic mixed effect regression. Overall, there was no significant difference in the accuracy of verbal responses, across tasks and language groups.

Table 3.2 Frequency (proportion) of correctness of verbal production in the two language groups in Action Naming and Event Description Task

		Correct	Incorrect	NA	Total
Action Task	Dutch	65 (33.16%)	81 (41.33%)	50 (25.51%)	196
	Chinese	47 (24.10%)	94 (48.21%)	54 (27.69%)	195
Event Task	Dutch	152 (40.00%)	215 (56.58%)	13 (3.42%)	380
	Chinese	147 (38.58%)	201 (52.76%)	33 (8.66%)	381

Table 3.3. Output for the logistic mixed effect regression model for verbal production accuracy. The fixed effects are *language* and *task. Language* Chinese and *Task* Action condition was coded as the reference level. Coefficient estimatesβ, standard errors *SE, z*-values and significant levels are reported. **p*<.05

	β	SE	z
Intercept	-1.484	0.621	-2.390*
Dutch	0.924	0.560	1.650
Event Task	0.484	0.713	0.678
Dutch: Event	-0.919	0.603	-1.524

Specificity of verbal descriptions

Figure 3.4 depicts the proportion of specific encodings of each event element in each task and language group. Table 3.4 and Table 3.5 report the frequency of specific and unspecific

[4] The effect of picture location relative to the fixation cross was also reported in Dobel et al., (2007). Participants were more easily to identify the actor that was closer to the fixation cross. We observed a similar effect in our study: participants' first fixations tend to locate towards a closer area to the fixation cross. For instance, when a picture was presented on the upper right of the screen, participants' fixations tend to cluster around the bottom left area of the picture (see Appendix 4 for a demonstration of the effect of picture locations in a scatterplot).

utterances in the Action naming task and Event description tasks, respectively. Table 3.6 - 3.8 report the results of logistic mixed effect regression analyses on Agent, Action and Object specificity.

Figure 3.4 Proportion of specific encodings of the agent (left), action (middle) and object (right) elements. Error bar: mean +/- 2*SE. Missing data was excluded before the analyses and plotting.

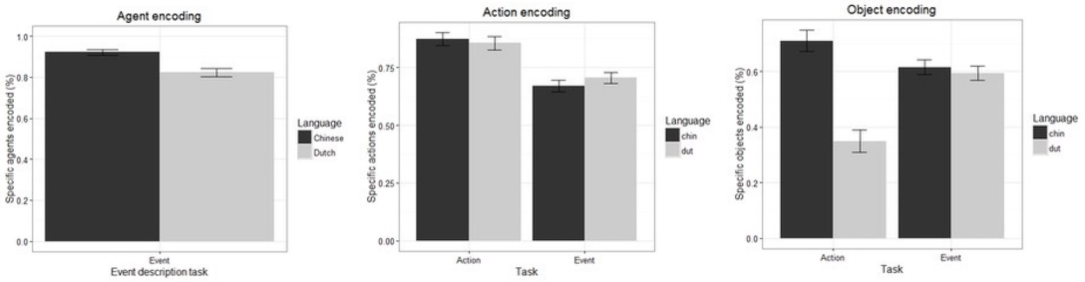


Table 3.4 Frequency (proportion) of the specificity of the verbal production in Action Naming Task for Dutch and Mandarin speakers.

Language	Event elements	Specific	Unspecific	NA	Total
Chinese	Action	123 (63.08%)	18 (9.23%)		
	Object	100 (51.28%)	41(21.02%)	54(27.69%)	195
Dutch	Action	125 (64.10%)	21 (10.71%)		
	Object	51 (26.15%)	95 (48.47%)	50 (25.51%)	196

Table 3.5 Frequency (proportion) of the specificity of the verbal production in Event Description Task for Dutch and Mandarin speakers.

Language	Event element	Specific	Unspecific	NA	Total
Chinese	Agent	321 (84.25%)	27 (7.09%)		
	Action	233 (61.15%)	115 (20.18%)		
	Object	214 (56.16%)	134 (35.17%)	33 (8.67%)	381
Dutch	Agent	302 (79.47%)	65 (17.11%)		
	Action	259 (68.16%)	108 (28.42%)		
	Object	218 (57.37%)	149 (39.21%)	13 (3.42%)	380

Table 3.6 Output for the logistic mixed effect regression model for Agent specificity. The fixed effect was *language*. *Language* Chinese was coded as the reference level. Coefficient estimatesβ, standard errors *SE, z*-values and significant levels are reported.

	β	SE	z
Intercept	0.158	1.215	0.130
Dutch	-1.587	1.084	-1.464

Table 3.7 Output for the logistic mixed effect regression model for Action specificity in Action Naming Task and Event Description Task. The fixed effect was *language* and *task. Language* Chinese and *Task* Action condition was coded as the reference level. Coefficient estimatesβ, standard errors *SE, z*-values and significant levels are reported. **p*<.05, **p<.001, ****p*<.0001

	β	SE	z
Intercept	2.178	0.449	4.847***
Dutch	-0.058	0.447	-0.129
Event Task	-1.244	0.551	-2.258*
Dutch: Event	0.339	0.504	0.674

Table 3.8 Output for the logistic mixed effect regression model for Object specificity in Action Naming Task and Event Description Task. The fixed effect was *language* and *task. Language* Chinese and *Task* Action condition was coded as the reference level. Coefficient estimatesβ, standard errors *SE, z*-values and significant levels are reported. **p*<.05, **p<.001, ****p*<.0001

	β	SE	z
Intercept	1.089	0.484	2.250*
Dutch	-2.090	0.544	-3.842***
Event Task	-0.636	0.557	-1.141
Dutch:Event	2.063	0.549	3.760***

There was no significant difference in Agent specificity between the two language groups (see Table 3.6 and Figure 3.4 left). For Action specificity, the proportion of specific encodings of the action element in the Event task was significantly lower than the Action task. No language effect was found (see Table 3.7 and Figure 3.4 middle). For object specificity, the interaction between *task* and *language* was significant: object specificity for Dutch speakers was significantly lower than Mandarin speakers only in the Action task. (see Table 3.8 and Figure 3.4 right).

3.3.2 Results of First fixation locations

Figure 3.5 depicts the mean and the distribution of FFLs in each task (Figure 3.5 left) in the Dutch and Chinese language groups (Figure 3.5). Table 3.9 presents the mean of FFLs and the corresponding standard error in each task and language group. Qualitatively, the data show different FFL patterns across different task conditions. The distribution in the two language groups also show a small numerical difference: for Dutch speakers, the difference between the mean of the Action and Event task is larger (ca. 20 pixels), compared with Mandarin speakers (ca. 3 pixels). Statistically, Table 3.10 reports the statistical output of the model.

There was a significant main effect of *task*, no main effect of *language,* and no interaction effects between *language* and *task*.

Figure 3.5 Distribution of the centered Y-coordinates of FFLs modulated by task demands (left) and language backgrounds (right). Error bar: mean +/- 2*SE

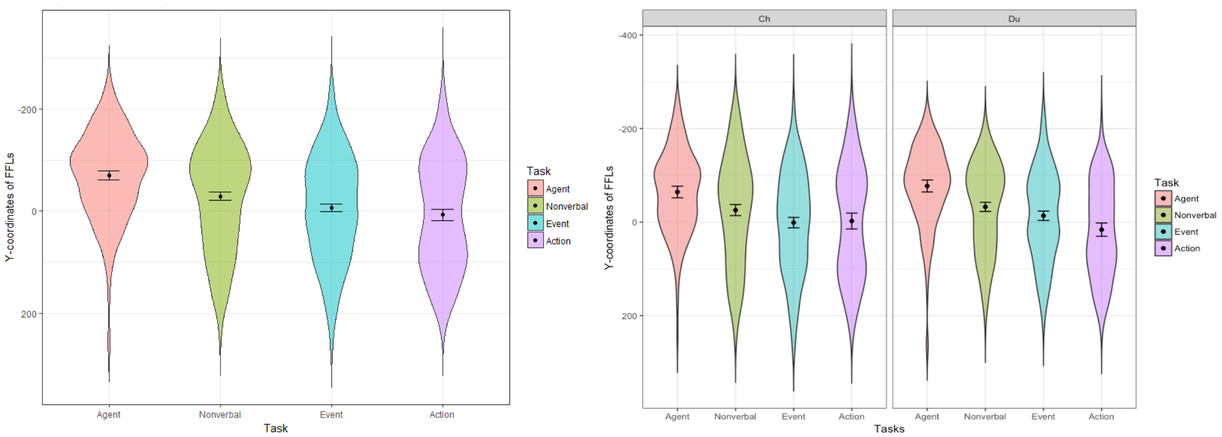


Table 3.9 Mean of the centered Y-coordinates of FFLs and its standard error for each task in each language group

Language	Task	Mean	SE
Chinese	Agent	-64.387	6.20
	Nonverbal	-25.709	5.93
	Event	1.090	5.60
	Action	-2.225	8.44
Dutch	Agent	-77.201	6.40
	Nonverbal	-32.577	4.89
	Event	-13.543	5.01
	Action	16.152	7.12

Table 3.10 Output for the linear mixed effect regression model on the Y-coordinates of FFLs. The fixed effects were *task* and *language* and their interaction (sum-coded). Coefficient estimatesβ, standard errors *SE, t*-values and significant levels are reported. **p*<.05, **p<.001, ****p*<.0001

	β	SE	t
Intercept	-23.015	66.676	-0.345
Dutch	2.521	5.521	0.457
Action Task	26.965	5.349	5.041***
Agent Task	-36.973	4.340	-8.520***
Event Task	16.703	4.897	3.411**
Dutch: Action	-3.651	4.876	-0.749
Dutch: Agent	-1.388	3.744	-0.371
Dutch: Event	5.700	3.571	1.596

Post-hoc analysis: effect of language

A post-hoc analysis was conducted to further explore the hypothesized cross-linguistic differences in the Action and Event Tasks. The data was split up into two language groups, and two linear mixed effect regression models were conducted to compare FFLs across tasks, for each group separately. The fixed effect was *task* (treatment coded), and the random effect structure included the maximal structure justified by the design, which contained random intercepts for *participant*, *stimulus* and *picture locations*, as well as a by-*participant* random slope for *task.*

Table 3.11 reports the statistical output of the model on the Mandarin Chinese dataset. There was a significant difference between FFLs in the Agent and Action Task, as well as between the Nonverbal and Action Task. The difference between the Event and Action Tasks was not significant.

Table 3.11 Output for the linear mixed effect regression model on the Y-coordinates of FFLs for Mandarin Chinese group. The fixed effect was *task* (treatment-coded, the Action task is the baseline). Coefficient estimatesβ, standard errors *SE, t*-values and significant levels are reported. **p*<.05, **p<.001, ****p*<.0001

	β	SE	t
Intercept	2.714	72.759	0.037
Agent Task	-61.397	11.409	-5.382 ***
Event Task	-0.543	12.805	-0.042
Nonverbal Task	-30.678	11.419	-2.687 *

For the Dutch group, Table 3.12 reports the statistical output of the model. The random effect structure here only included random intercepts for *participant*, *stimulus* and *picture locations*. The random slope was excluded due to a convergence problem. The results show that FFLs in each task were significantly different from each other.

Table 3.12 Output for the linear mixed effect regression model on the Y-coordinates of FFLs for Dutch group. The fixed effect was *task* (treatment-coded, the Action task is the baseline). Coefficient estimatesβ, standard errors *SE, t*-values and significant levels are reported. **p*<.05, **p<.001, ****p*<.0001

	β	SE	t
Intercept	0.0226	61.518	0.000
Agent Task	-56.323	8.066	-6.983 ***
Event Task	-15.065	7.080	-2.128*
Nonverbal Task	-31.984	7.056	-4.533 ***

3.4 Discussion of Experiment 1

Experiment 1 investigated the effect of task demands and language background on event apprehension, measured by FFLs. Dutch and Chinese participants described stimuli that were briefly presented for 300ms, according to different linguistic tasks.

3.4.1 Verbal production data

Given the high task demands associated with brief stimulus exposure, the results of the production data showed a below-average performance on accuracy for correctly and concretely describe event stimuli, across the two language groups and task conditions. The specificity of verbal encoding showed generally similar patterns across groups: both language groups can produce specific descriptions at a similar level in the two tasks, except the object specificity difference in the Action task between the two language groups. The significant group difference in Object specificity in the Action Task can be explained by the different degree of naturalness of producing a single verb as a description of an action in the two languages: In Dutch, it is common to adopt an infinite verb form referring to an action (e.g., “tekenen”, an example from the transcription); In Mandarin, by comparison, verb-object collocations are more common and natural (e.g., 画画 hua-hua, “draw a painting”, an example from the transcription). In addition, the denotation of a large number of verbs in Mandarin will be underspecified if no object is followed. For instance, “schudden kaarten" in Dutch, meaning of shuffling cards, strictly correspond to the action (“shudden”) and the object (“kaarten”) in a word-by-word fashion, but in Mandarin, the expression is 洗牌 Xi-Pai, where 洗 Xi, as the verb, only means “to wash”, if no object is followed. Thus, many action descriptions have to adopt a Verb + Object structure in Mandarin to encode a specific action, compared with Dutch, where the verb per se already encode the action specifically. Object specificity was higher in Mandarin due to this linguistic constraint.

There was also a small shrink on the proportion of Action specificity in Event task compared with Action task (see Table 3.7 and Figure 3.4), which should be accounted by the exclusion of missing data. According to Table 3.4 and 3.5, the frequency of missing data was higher in Action task than in Event task. Due to the limitation of statistical analysis, missing data were not taken into account. If missing data were included, the proportions of specific action elements are similar in the two tasks (ranging from 60%-70% in both task and language groups).

Overall, the accuracy of the verbal production, and the description specificity of event elements showed no difference in the two language groups and the two description tasks.

3.4.2 First fixation locations

Regarding the fixation data, the results suggested a significant main effect of task demands, where in the Agent identification task, the FFLs clustered on the upper part of the stimuli showing agent-identifying information (face and upper body) of the visual scene. Similarly, in the Action or Event task, the FFLs clustered at the bottom half of the stimuli, showing the object- and action-related information. The result suggests that the FFLs, reflecting the event apprehension process, is modulated by linguistic task demands.

Regarding the effect of language background, a difference was found in the post-hoc analysis only when the data from the two language groups were split up (However, it is noteworthy that there was no overall interaction between task demands and language backgrounds). The FFLs in the Action Naming and Event Description Tasks showed a different pattern in the two language groups: For Dutch speakers, the mean of the FFLs for the two tasks clearly dissociated from each other where FFLs in the Action task were significantly lower than FFLs in the Event task; For Mandarin speakers, by contrast, the mean of the FFLs of the two tasks were similar and cluster in the middle of the Agent and Action/object element. This difference indicated that for Mandarin speakers, the apprehension of the Action element of an event was similar to the apprehension of the whole event scene, which included the agent element even when explicit naming was not required.

3.4.3 General discussion of Experiment 1

Linking the verbal production data with the fixation data, Experiment 1 showed that FFLs do not correlate with verbal output. Specifically, although speakers' first fixations can reasonably be located on the most informative visual area needed for the task at hand, the verbal descriptions of the visual scene, due to the high demands of the brief exposure paradigm, were low in terms of accuracy and specificity.

More interestingly, in the Event Task, FFLs were not located towards the agent-region that locates on the upper part of the stimuli, as a strictly linear incrementality hypothesis would predict (e.g., Gleitman et al., 2007). Namely, if message encoding followed the same linear order of the language structure, the FFLs in Event task should have been clustered on the Agent element, as a subject need to be uttered in the initial position of a sentence in both languages given the stimuli setting.

Alternatively, the distribution of the FFLs in the Event Task supported the hierarchical incrementality hypothesis (in line with Dobel et al., 2010; Dobel et al., 2007; Griffin & Bock, 2000). The FFLs distribution, located near the center of the stimuli, suggests an "overall" fixation pattern, where the FFLs in the Event task clustered *in between* the Agent and Action/object elements. This "in-between" location was considered as an attempt of apprehending the overall event structure by combining the two event elements (Gerwien & Flecken, 2016) and has been reported previously in an event description task. We replicated this pattern here.

Regarding cross-linguistic differences, different fixation patterns were found in the Action and Event Tasks across the two language groups, but only when the data were split by language. The language effect reported above could be explained by Mandarin being a pro-drop language: In order to license the omission of the reference to the subject in a sentence, the entity needs to be identified as part of the context knowledge. Our hypothesis is thus that Mandarin speakers typically keep track of the agent element in events (i.e., the subject of a sentence in this study), even when mentioning the subject is not obligatory, as the case of Action Naming Task. However, when interpreting data for cross-linguistic differences, it is noteworthy that the sample sizes in the two tasks were relatively small and unbalanced (i.e., only half of the participants performed the Action Task while the whole sample participated in the Event Task), and the random effect structure also differed due to the convergence problem in the Dutch group.

To further explore the cross-linguistic differences we designed a follow-up experiment. Experiment 2 focused mainly on the Action Naming and Event Description Task in order to examine whether the FFLs patterns found in Experiment 1 could be replicated. In addition, a surprise Recognition Memory Task was designed to assess the extent to which participants' memory of the encoded events, in particular the *agents*, may differ across tasks and/or languages. The memory task was performed after both the Action and the Event task. Participants selected the stimulus they had seen in the preceding apprehension task, from two stimuli that only differed in the agent element.

We expected, based on the effect of task demands in Experiment 1, a better recall accuracy towards the agent element after performing the Event description task, compared with the Action description task. The hypothesis was that the explicit encoding of the agent in language production as well as a rudimentary fixation location that integrate the Agent and Action/object element may enhance the encoding of agent information in memory. This is different from the Action task where only the action element is verbally encoded and was found to be quite precisely fixated.

Regarding cross-linguistic differences, if the different FFLs patterns across the two language groups reported in Experiment 1 are reflecting a true language-specific apprehension process, we expected the recall accuracy in relation to the agent element in the Action naming task for Mandarin speakers to be higher compared to Dutch speakers. The reason was that, if FFLs indeed are affected by the different degrees of agent saliency in the two languages, in the fashion speculated in Experiment 1, Mandarin speakers may keep track of and store the salient, but sometimes implicit agent information more habitually compared with Dutch speakers, and thus may obtain a better memory of the agent information even when explicit encoding is not required.

4 Experiment 2

4.1 Method

4.1.1 Participants

The Dutch group included 29 participants recruited from the participant pool of the Max Planck Institute for Psycholinguistics (Age mean=22.34, SD=3.68, N=22 female and N=7 male). Out of the original group, six participants had to be excluded due to technical error. The final Dutch group consisted of 23 participants.

The Chinese group included 24 participants recruited (Age mean=28.29, SD=5.28, N=17 female and N=7male). Participants are international students or employees enrolled at Radboud University Nijmegen. Out of the original group, one participant had to be excluded due to technical error. The final Chinese group consisted of 23 participants.

All the participants had normal or corrected-to-normal vision. All the participants received a payment of 8 euros.

4.1.2 Task and list design

Experiment 2 adopted a block design with three apprehension tasks, identical to Experiment 1: namely, Agent Naming Task, Action Naming Task and Event Description Task (see 3.1.2 for a detailed task description).

A surprise Recognition Memory Task (“Memory task” below) was added: after the second and the third blocks, participants were required to make a choice from two photos that only differed in the agent actors. The object and the action elements were hold the same (see an example trial in Figure 4.1). The two stimuli were presented side by side. The positions of the original picture (i.e., correct choice on the left or right) were pseudo-randomized.

Two lists were created to counterbalance the order of tasks, and each list was assigned to an equal number of participants. Each list contained three blocks: The first block in the both lists was the Agent Naming Task. The order of the Action Naming and Event Description Tasks in the second and the third block were counterbalanced in list 1 and list 2 (Table 4.1)

Figure 4.1 Example stimuli for Recognition Memory Task. The two pictures differed in the action performers while the action and object elements were hold constant.

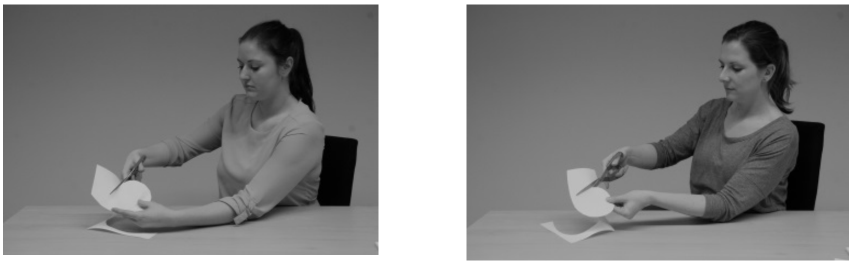


Table 4.1 List and Block design for Experiment 2

	List 1 (Chinese N=12, Dutch =11)	List 2 (Chinese N=11, Dutch N=12)
Block 1	Non-verbal Task	Non-verbal Task
Block 2	Action Naming Task	Event Naming Task
Memory	Recognition Memory task	Recognition Memory task
Block 3	Event Naming Task	Action Naming Task
Memory	Recognition Memory task	Recognition Memory task

4.1.3 Materials

Stimuli were causative event photographs, identical to Experiment 1 (Appendix 2 for a full list). 48 causative event photographs depicted one of the four actors (3 female, 1 male) performs an action on a single object. Each task contained 16 stimuli, half of which were presented in agent-left orientation, and the other half in agent-right orientation. The Memory Task included 64 stimuli, half of which were identical as the stimuli in the Action and Event task, while the other half of the stimuli showed the same action/object element, but with a different agent.

4.1.4 Procedure

The participants were asked to sign the consent form first. The apparatus, software package as well as the calibration procedure was identical with Experiment 1. What was different was that the eye tracker was attached to the lower part of a desktop computer screen with a display resolution of 1920*1080, and was situated in a sound-proof booth. The desktop computer screen for testing was controlled by an SMI laptop situated outside the booth.

Chinese participants were instructed by a Chinese native language experimenter. Dutch participants were tested by the same experimenter but the oral instructions were given in English. Written instructions to the tasks were presented in their native language (Dutch and Mandarin). Each participant performed the assigned list and the corresponding tasks. The experiment session lasted approximately 30 minutes.

For each apprehension task, the trial procedure was identical with Experiment 1 (see Section 3.1.4). At the end of the second and the third block, instructions on the Memory Task were presented. This task was not introduced prior to the apprehension tasks in order to guarantee that the participants did not intentionally spare more attention on the agent elements during apprehension tasks.

A trial in the Memory Task started with the word "Ready?" appearing on the screen. The participants were instructed to click the mouse to proceed. Then, the two stimuli were presented side by side. The participants clicked on the picture they had seen in the previous apprehension task. This was followed by a blank screen and participants clicked the mouse to proceed to the next trial. Participants' eye movements and mouse clicks were recorded.

4.2 Data preprocessing and analysis

4.2.1 Memory data

We analyzed the accuracy of agent memory. All participants (N=46) were included in the accuracy analysis, given that every participant performed the task properly. Missing Data were marked as NA when the participants clicked on the blank space surrounding the pictures (2.82% of the data was missing in total, among which 1.60% was from Mandarin speakers and 1.2% from Dutch speakers).

Memory accuracy was analyzed with logistic mixed effect regression, with the fixed effects of *task, language* and their interaction (sum coded), as well as a random effect structure of *participants* and *stimuli* that maximally justify the design, which contained random intercepts for *participant* and *stimuli*, as well as a by-*participant* random slope for the effect of *task.*

4.2.2 Eye movement data: First fixation locations

The analysis focused on the FFLs recorded in the three tasks. Data transformation and exclusion criteria were identical with Experiment 1 (see 3.2.2). In total, 21 Dutch and 20 Chinese participants were included in the final analysis. The analytical method and regression model structure were also identical with Experiment 1.

Given that Experiment 1 have shown that FFLs did not correlate with verbal production (see Section 3.4.3) and similar results from previous studies (e.g., Gerwien & Flecken, 2016), the verbal data of this experiment was not analyzed.

4.3 Results

4.3.1 Recognition Memory Task

Figure 4.1 presents the accuracy of agent memory for each condition (left) and in the two blocks (middle and right). Table 4.2 reports the accuracy of memory task in each task and block. Table 4.3 reports the statistical output of the model. There was a significant main effect of *language*, no main effect of *task,* and no interaction effects between *language* and *task*.

Figure 4.2 Proportion of correct choices in the Memory Task. Overall, Dutch participants' accuracy was significantly higher than the Mandarin group (left). A significant effect of task was found in Mandarin Chinese speakers in Block 1 (middle) but not in Block 2 (right) in the post-hoc analysis. Missing data was excluded before the analysis and plotting. Error bar: mean +/-2*SE.

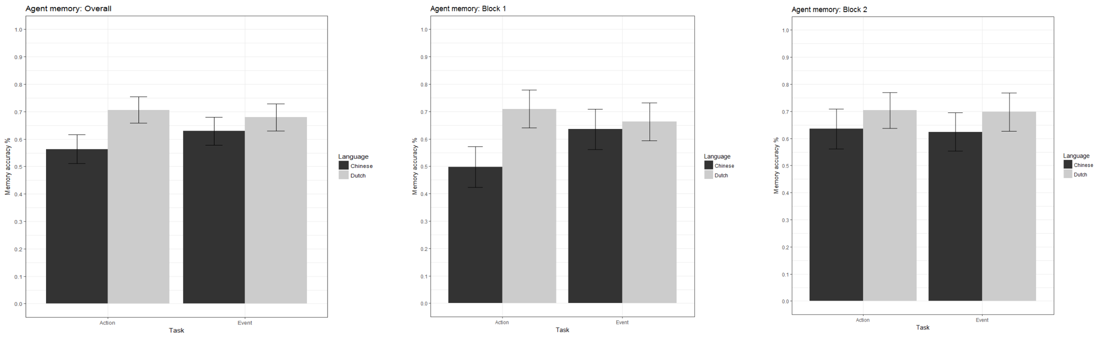


Table 4.2. Frequency (proportion) of the correct choices in the Memory Task. The proportion of correct choices was calculated based on the total number of trials in the single task in each block and language group.

Block	Preceded Task	Language	Correct	Total
Block 1	Action naming task	Chinese	91 (49.73%)	183
		Dutch	122 (70.93%)	172
	Event description task	Chinese	108 (63.53%)	170
		Dutch	124 (66.31%)	187
Block 2	Action naming task	Chinese	108 (63.53%)	170
		Dutch	133 (70.37%)	189
	Event description task	Chinese	118 (62.43%)	189
		Dutch	120 (69.77%)	172

Table 4.3 Output for the logistic mixed effect regression model for the accuracy of Recognition Memory Task. The fixed effects were *language, task and their interaction* (sum-coded). Coefficient estimatesβ, standard errors *SE, z*-values and significant levels are reported. **p*<.05, **p<.001, ****p*<.0001

	β	SE	z
Intercept	0.659	0.124	5.307***
Dutch	-0.229	0.061	-3.760***
Event Task	-0.030	0.061	-0.498
Dutch: Event	-0.111	0.060	-1.844

Post-hoc analysis: data split by blocks

Due to the fact that the second Memory task ("block 2" below) was no longer a surprise compared to the first one ("block 1" below), a post-hoc analysis split the data by the two blocks, looking at the first and the second memory task separately. Logistic mixed effect regression analysis was performed to compare the accuracy of agent memory in the two tasks in each block. The fixed effects were *task* and *language* (sum coded) as well as their interaction, and the random effect structures included the maximal structure that justified the design, which contained random intercepts for *participant*, *stimulus* and *picture locations*, as well as a by-*participant* random slope for the effect of *task.*

For block 1 (also see Figure 4.2 middle), Table 4.4 reports the results the statistical output of the model. There was a significant main effect of *language*, as well as a significant interaction effect of *language* and *task.* The effect of *task* was not significant. For block 2 (also see Figure 4.1 right), Table 4.5 reports the results the statistical output of the model. There was a significant main effect of *language,* no effect of *task* and their interaction.

Table 4.4 Output for the logistic mixed effect regression model for the accuracy of Recognition Memory Task in block 1. The fixed effects were *language, task* and their interaction (sum-coded). Coefficient estimatesβ, standard errors *SE, z*-values and significant levels are reported. **p*<.05, **p<.001, ****p*<.0001

	β	SE	z
Intercept	0.52037	0.11425	4.555***
Dutch	-0.26696	0.08036	-3.322***
Event Task	-0.09178	0.08037	-1.142
Dutch: Event	-0.19804	0.08017	-2.470*

Table 4.5 Output for the logistic mixed effect regression model for the accuracy of Recognition Memory Task in block 2. The fixed effects were *language* and *task* and their interaction (sum-coded). Coefficient estimatesβ, standard errors *SE, z*-values and significant levels are reported. **p*<.05, **p<.001, ****p*<.0001

	β	SE	z
Intercept	0.879	0.231	3.807***
Dutch	-0.180	0.089	-2.016*
Event Task	0.028	0.090	0.090
Dutch: Event	0.002	0.089	0.025

4.3.2 Results of first fixation locations

Figure 4.2 presents the mean and the distribution of FFLs in each task (Figure 4.2 left) and in each Dutch and Chinese language group (Figure 4.2 right). Table 5.6 presents the mean of FFLs and the corresponding standard error in each task and language group. Qualitatively, the data show different FFL distributions across the three task demands, similar as in Experiment 1, but no obvious cross-linguistic differences can be observed.

Statistically, Table 4.7 reports the statistical output of the model. The final model included fixed effects of *language*, *task*, and their interaction (sum coded). The random effect structure contained random intercepts for *participant*, *stimulus* and *picture locations*, as well as a by-*participant* random slope for the effect of *task.* There was a significant main effect of *task*, no main effect of *language,* and no interaction effects between *language* and *task*.

Figure 4.3 Distribution of FFLs modulated by task demands (left) and language backgrounds (right). Error bar: mean +/- 2*SE

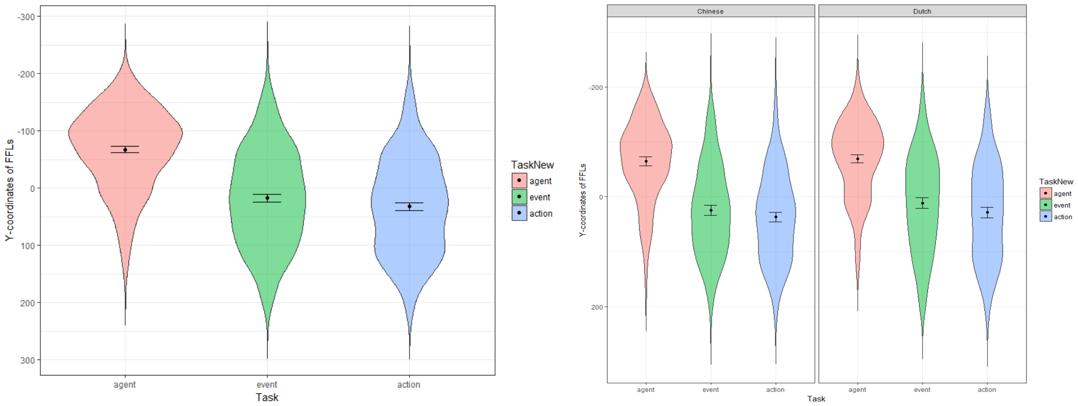


Table 4.6 Mean of the Y-coordinates of the FFLs and its standard error for each task in each language group

Language	Task	Mean	SE
Chinese	Agent	-64.650	4.056
	Event	24.223	4.654
	Action	36.809	4.563
Dutch	Agent	-69.546	3.840
	Event	11.164	4.816
	Action	28.319	4.787

Table 4.7 Output for the linear mixed effect regression model on the Y-coordinates of FFLs. The fixed effects were *task* and *language* and their interaction (sum coded). Coefficient estimatesβ, standard errors *SE, t*-values and significant levels are reported. **p*<.05, **p<.001, ****p*<.0001

	β	SE	t
Intercept	-3.3037	49.2136	-0.067
Dutch	5.4165	4.6497	1.165
Action Task	38.4465	3.3819	11.368***
Agent Task	-59.5268	5.0128	-11.875***
Dutch: Action	-0.2563	2.7261	-0.094
Dutch: Agent	-1.9023	3.0166	-0.631

Post-hoc analysis: data split by language

A post-hoc analysis was conducted to check for cross-linguistic differences in fixation patterns (similar to Experiment 1). The data was split up by language group and linear mixed effect regression models were performed to compare the FFLs in each task for each group. The fixed effect was *task* (treatment coded), and the random effect structures included the maximal structure that justified the design, which contained random intercepts for *participant*, *stimulus* and *picture locations*, as well as a by-*participant* random slope for the effect of *task.*

Table 4.8 and Table 4.9 reports the statistical output of the model for the Mandarin and Dutch group respectively. The results suggested that for both language groups, FFLs distributions in each task were significantly different from each other.

Table 4.8 Output for the linear mixed effect regression model on the Y-coordinates of FFLs for Mandarin Chinese group. The fixed effects are *task* (treatment-coded, the Action task is the baseline). Coefficient estimatesβ, standard errors *SE, t*-values and significant levels are reported. **p*<.05, **p<.001, ****p*<.0001

	β	SE	t
Intercept	39.877	44.013	0.906
Agent Task	-100.005	9.914	-10.087***
Event Task	-14.760	6.710	-2.200*

Table 4.9 Output for the linear mixed effect regression model on the Y-coordinates of FFLs for Dutch group. The fixed effects are *task* (treatment-coded, the Action task is the baseline). Coefficient estimatesβ, standard errors *SE, t*-values and significant levels are reported. **p*<.05, **p<.001, ****p*<.0001

	β	SE	t
Intercept	30.025	55.002	0.546
Agent Task	-96.854	8.718	-11.110***
Event Task	-19.624	7.143	-2.747 *

4.4 Discussion of Experiment 2

Experiment 2 continued to investigate the effect of task demands and language background on event apprehension indexed by FFLs. It focused on comparing the Action Naming Task and Event Description Task, with an equal distribution of the sample sizes. Moreover, a surprise Memory Task was added after the two apprehension tasks in order to test participants' memory of the agents shown in the stimuli.

4.4.1 Memory data

In general, the accuracy on the Memory task was above chance level, which indicated that, despite the high demands of the brief exposure task, participants stored information presented in the pictures, including the agent elements, in memory, at least to some extent.

However, the results did not support our hypothesis (see 3.4.3 for details). Instead, we found a significant main effect of language, with Dutch participants' performance overall higher than Mandarin participants. There was no effect of task.

One of the speculative reasons that we propose for the better performance of Dutch participants is that they (mainly, university students) are more experienced in taking part in psycholinguistic experiments. Looking at the data, we found that FFLs in the Dutch data were clearly modulated by the location of the stimuli on the screen (see plot in Appendix 4). Dutch participants fixated the region closer to the fixation cross, with the least scan path distance: For stimuli that were shown in the lower part of the screen, the FFLs had a tendency to cluster in the upper part of the stimuli, which was closer to the fixation cross. Similarly, stimuli shown on the upper part of the screen obtained more FFLs that clustered in the lower part of the stimuli, which suggested that Dutch participants tended to use least effort to move their eyes to fulfill the tasks. By comparison, this pattern was less clear in Mandarin speakers. However, even when the Dutch used this economic strategy of placing the first fixation on the area that was closest to the fixation cross, their accuracy on the agent memory task was significantly higher than Mandarin speakers. This indicated that, even when Dutch participants put less effort in moving their eyes, they were able to retrieve the required information more successfully than the Chinese, suggesting that their peripheral vision is more enhanced or trained.

In addition, a post-hoc analysis split the data by blocks. The results suggested that participants' memory of the agent element was generally higher in block 2 in the Mandarin group (See Table 4.2). One plausible reason is that the memory task in block 2 did no longer come as a surprise to the participants, given that they had already participated in one memory task in block 1. Therefore, we analyzed data from the memory task after block 1 and block 2 separately. In block 1 and for Mandarin speakers, agent memory was significantly lower in the Action task compared with that in the Event task. However, in block 2, this task effect was no longer presented. The difference in the agent memory accuracy between the two tasks, found in block 1, could be driven by the influence of explicit agent naming in the Event task, which was not required in the Action task. Naming the agent could attract more attention and enhance the memory of the agent element for the participants. However, this speculation has to be interpreted with caution, as no similar difference was found in block 2, nor in the Dutch group.

4.4.2 First Fixation locations

The effect of task demands on FFLs was replicated in Experiment 2: in the Agent identification task, the FFLs clustered on the upper part of the stimuli showing agent-identifying information of the visual scene. In the Action or Event task, the FFLs clustered at the lower part of the stimuli, showing the object- and action-related information. However, the effect of cross-linguistic differences found in the exploratory analysis in Experiment 1 was not replicated. The FFL distributions were similar in the Dutch and Mandarin group in all the tasks. Due to the inconsistent results in the two experiments, more research is required to understand the top-down effects of language backgrounds on apprehension. The potential reasons accounting for the mixed results in Experiment 1 and 2 are discussed in Section 5.3 in General Discussion.

4.4.3 General discussion on Experiment 2

Experiment 2 replicated the effect of task demands on event apprehension, which was indexed by the distributions of the FFLs in the different linguistic tasks: the FFLs cluster around the visual area that was the most informative and needed for the linguistic task at hand. More importantly, the FFLs distribution in the Event Task supports the hierarchical incrementality account for message encoding (e.g., Griffin & Bock, 2000): the rudimentary "in-between" fixation pattern (Gerwien & Flecken, 2016) indicates that within 300ms, apprehension involves more than deciding just on a starting point. Speakers need to understand the event structure (e.g., what is the agent or the patient), at least to some extent, to be able to both select the starting point and to decide how to precede the sentence. We discuss this effect further in the context of language production in General Discussion.

5 General Discussion

The current study set out to investigate the effect of task demands and speakers' language backgrounds on event apprehension. Apprehension refers to the rapid visual process during which the gist of a scene can be extracted. It was measured through an analysis of *first fixation locations* on causative event visual stimuli that were presented for only 300ms. First fixation locations are considered the result of the apprehension process: it is based on the information extracted within initial visual processing; during apprehension a first fixation is computed and a saccade is launched toward the most relevant or informative region (Holmqvist, 2011). Importantly, the first fixation is not randomly placed. Instead, it is directed towards the most informative region for the task at hand, guided by the information obtained during apprehension (Bock et al., 2003; Gerwien & Flecken, 2016). Three main findings from Experiment 1 and 2 are summarized and further discussed in this chapter.

First of all, apprehension is a *flexible* process. Within a stimulus exposure of 300ms, the first fixation is already modulated by the demands of different linguistic tasks, varying in the attention that the tasks drive towards different elements in an event scene. The linguistic tasks as well as the information extracted from apprehension direct the first fixation to locate on the most informative visual area required by the task at hand (e.g., focusing on the Agent element in the Agent naming task).

Second, the first fixation location serves as an index that can isolate the apprehension process from linguistic formulation. Furthermore, the fixation pattern based on the data from the Event Description task is in support of the hierarchical incrementality account in language production. Our data show that first fixation locations do not predict the specificity and accuracy of the verbal output, nor do they correlate with the starting point of the verbal descriptions of the stimuli (i.e., the agent element) in the Event Description Task. This implies that where viewers look first, does not necessarily decide the linear order of the sentence to be produced. Rather, it reflects the result of pre-linguistic visual processes, i.e., apprehension. The absence of a linear relation between first fixations and the linguistic starting point further supports the hierarchical incrementality account, assuming that a separated and holistic conceptualization process precedes linguistic formulation. In the Event description task in Experiment 1, participants exclusively produce a full sentence with a subject-initial word order (i.e., the Agent element is mentioned first in the sentence), but the first fixations in this task are not directed solely towards the Agent. Rather, they are located in between the areas of the Agent and Action/object elements, which we interpret as an attempt to combine the two elements and to understand the overall event structure during apprehension, before engaging in linguistic formulation processes.

Third, regarding the top-down effect of cross-linguistic variation on apprehension, the current study obtains mixed results in the two experiments. In Experiment 1, the fixation patterns in Action Naming Task and Event description task differ across the two language groups: For Dutch speakers, while first fixation in the Event Description Task tends to locate in the middle of the Agent and Action/object element, the first fixations in the Action Naming Task locate significantly lower than those in the Event task. They tend to cluster at the lower bottom half of the stimuli, directly associated with the Action/object element. By contrast, first fixation data in the Mandarin group obtain a similar pattern in the two linguistic tasks. The first fixations in the two tasks both are located close to the midline of the stimuli. However, this cross-linguistic differences on first fixation locations in the two linguistic tasks is not replicated in Experiment 2. In both language groups in Experiment 2, the first fixation locations in the Action task are significantly lower than the first fixations in the Event task.

In addition, the memory of the Agent element in the Action and Event tasks, measured by the accuracy of agent memory in the surprise Recognition Memory task, cannot provide conclusive evidence on cross-linguistic differences in agent saliency. The memory data only show a small difference between tasks: The memory of the Agent in the Event task is better than that in the Action Task for Mandarin speakers. In other words, Mandarin speakers' memory of the Agent is better when they are required to utter a full sentence explicitly naming all the event elements in the stimuli (i.e., the Event task), compared with only naming the Action/object of the event (i.e., the Action task). However, this difference is only found in the first Recognition Memory task but not in the second memory task, nor among Dutch speakers. Indeed, the first recognition task is more surprising than the second recognition task after the third block, because participants may have been informed by the first Memory task to pay special attention to the agent element in the following experiment, given that the picture alternatives only differed in the Agent element. Thus, the effect should be more relevant and robust in the first Memory task. However, the Dutch group shows no difference between tasks nor blocks, and their agent memory is significantly higher than Mandarin speakers throughout the two blocks. The current evidence is based on a small sample size (i.e., Mandarin speakers and block 1 only) and thus cannot provide a definite conclusion on the effect of apprehension and linguistic encoding on memory. Further research is required for exploring the effect of cross-linguistic differences on agent saliency both in apprehension as well as the memory of event structures.

5.1 Apprehension is a flexible process

The current study is the first empirical demonstration of how specific linguistic task demands modulate first fixation patterns, resulting from the apprehension process. The different FFL patterns in the Agent and Action naming tasks in both Experiment 1 and 2 suggest that, during the apprehension phase, participants have extracted information on event structure (i.e., event agent, patient and action) and subsequently computed a saccade towards the relevant region within 300ms, which lead to the first fixation on the task-relevant agent or action/object elements of the visual scenes.

These results suggest that the apprehension process is flexible, which is in line with previous free-viewing eye tracking studies that also manipulate task demands prior to visual exposure. For instance, studies adopting object-search tasks have shown that fixations tend to be placed on task-relevant objects, rather than the most perceptually salient objects (e.g., Hayhoe et al., 2003; Henderson, 2003; Henderson et al., 2007). It has also been shown that participants are faster to fixate on semantically consistent objects (e.g., a cocktail in a barroom), compared with semantically inconsistent objects (e.g., a microscope in the barroom) (Henderson et al., 1999).

The current study, extending these findings, adopts first fixation locations obtained in a brief exposure paradigm, as a direct measurement of apprehension to show that the result of apprehension can be detected within 300ms. A top-down effect of task demands is captured during apprehension; this factor directs the very first fixation towards the goal of the task at hand (following theoretical proposals by Dobel et al., 2010; Henderson & Ferreira, 2004).

5.2 First fixation locations and message encoding

One of the aims of the present study is to isolate the process of apprehension from linguistic formulation in language production, in order to shed light on the “starting point” debate in language production that concerns the relation between the initial fixation and the initial word in utterance. A traditional view for this debate is the linear incrementality account, which argues that the element selected as the starting point is based on the extra attention the element may require, or the importance or perceptual saliency it represents (e.g., Osgood & Bock, 1977). It predicts that the starting point of speakers’ utterances is driven by what captures attention first. This was indeed observed in the experiments by Gleitman et al. (2007), in which a briefly presented perceptual cue was used to capture speakers’ attention on certain event roles in line drawings. However, the linear relation between first fixations and utterance starting points as suggested in Gleitman et al. (2007) is not replicated in our study. The first fixations in the Event description task from both Experiment 1 and 2 do not cluster around what is mentioned first, i.e., the Agent element.

The present study, together with those from e.g., Griffin & Bock (2000), Sebastian et al. (2013) and Gerwien & Flecken (2016), do not favor the linear incrementality account. Rather, we argue that an apprehension process should happen prior to linguistic formulation in the language production process. The first fixation data obtained here indicate that although the information extracted during apprehension can successfully correspond to the linguistic task demands, first fixations do not predict the accuracy and specificity of the verbal descriptions. Furthermore, first fixations do not predict what is mentioned first in event descriptions (i.e., the agent in all cases), but rather, they show an “In-between” pattern locating in the middle of the Agent and Action/object of the stimuli, which we interpret as an attempt to combine the two elements and to extract the overall event structure during apprehension, before engaging in linguistic formulation processes.

Put differently, the “in-between” FFL pattern in the Event description tasks in both experiments provides a direct empirical answer in support of the alternative hypothesis in the “starting point” debate, namely, the hierarchical incrementality account: the apprehension phase prior to linguistic formulation involves the extraction of relational information and event structure, which guides the location of first fixations. As such, the “in-between” first fixation pattern in the Event description task reflects an attempt to integrate information on event actions and event roles at the same time. This pattern is In line with the early fixation locations reported in Bock et al. (2003), Griffin & Bock (2000) and Gerwien & Flecken (2016). Here, we do not deny that perceptual cues *can* affect utterance starting points, but our results suggest that speakers do not necessarily only follow perceptual prominence to start their utterances (Bock et al., 2004).

The present study thus supports previous evidence in line with the hierarchical incrementality account (e.g., Griffin & Bock, 2000; Sauppe et al., 2013). In the current study, we zoom into the brief apprehension process by adopting a brief exposure paradigm and measuring first fixation locations. Given that first fixation locations can vary across different exposure times (Gerwien & Flecken, 2016), the fixation patterns obtained in the present study should not be considered as equal to the first fixations reported in the previous free-viewing eye tracking studies. Our data on first fixation locations should be more informative concerning the apprehension process, since they are achieved in a more demanding brief exposure condition (i.e., within in 300ms), with the aim of isolating apprehension from the language formulation phase in language production.

One unanswered question is how first fixations and apprehension relate to the message encoding phase in language production. Message encoding is the initial stage of language production (Levelt, 1989). The root of the hierarchical incrementality account can be dated back to Wunt (1900) and Lashley (1951), who argue that there is a holistic conceptualization phase that precedes linguistic formulation. However, this assumption does not necessarily refer to a situation involving visual perception, where speakers have to describe what they see. With the emergence of eye tracking techniques, message encoding is largely monitored by and linked with eye movements, mainly fixations, assuming that where people look can reflect what they think, which, to some extent, reliably reflect the cognitive processes of language production (Griffin, 2004), However, the link between fixations and message encoding is imperfect (Konopka & Brown-Schmidt, 2014). Indeed, in the visual world, it is very difficult to strictly tease apart apprehension from message encoding. Our analysis on first fixations as well as the previous studies on apprehension mostly alter the question regarding message encoding and linguistic formulation into testing the relationship between apprehension and linguistic formulation (also as in e.g.,Bock et al., 2003; Dobel et al., 2010;

Gleitman et al., 2007; Griffin & Bock, 2000, etc.). What we can conclude based on first fixation locations is that the first fixation, as the result of apprehension, does precede linguistic formulation, and thus apprehension is an isolated visual process prior to linguistic processing. However, it is unclear to what extent first fixations also reflect message encoding. In other words, is the preverbal message constructed simultaneously with the visual apprehension process? If so, to what extent does visual processing during apprehension and message conceptualization interact? What is the temporal relationship between message encoding and apprehension? The current study cannot offer definite answers to these questions. A potential method to dissociate apprehension and message encoding is to present a mask immediately after briefly exposing the stimuli, in order to "constrain visual information uptake to the duration of prime presentation" (Zwitserlood et al 2018). In other words, it will block further processing. A mask may be able to force the participants to engage in an apprehension process only, even without a fixation. Based on a more stringent control on apprehension, further research can shed light on the interrelation between apprehension and message encoding, e.g., when and under which conditions people finish the construction of a message.

Another direction for future studies is to target the scope of message planning under the hierarchical incrementality account. Although the current study shows evidence supporting an "overall" message planning, a critical question is, to what extent the scope of message planning on a visual scene can be counted as "structural" and "holistic" (termed in Griffin & Bock, 2000)? According to Konnopka & Brown-Schmidt (2014), the scope of message planning refers to the unit of conceptual information speakers can construct before passing the information to linguistic formulation phase. Previous studies show that the message should contain at least a unit that cannot be segmented into smaller function units (e.g., a "functional phrase" such as *the flower*/*above the house* in Allum & Wheeldon, 2009) or even a larger unit of conceptual information for one clause (Smith & Wheeldon, 1999). Our fixation and verbal data imply that speakers can manage to plan event roles and the action, i.e., an entire causative event structure, within the apprehension stage. However, if the complexity of the visual scene increases, can speakers still manage to apprehend the entire visual scene within one glance? If not, what would be the scope for conceptual information that can be captured within apprehension? To answer these questions, more complex visual scenes are required in the future studies. For example, an event that includes more event elements can be designed as stimuli: in addition to causative events or transitive events that normally involve only two event roles (i.e., an agent and a patient), a more complex event can also include event elements such as instruments, locations and event goals, etc.

5.3 Top-down effect of language on apprehension: more research is required

The results of the cross-linguistic comparisons in Experiments 1 and 2 show mixed findings. Whilst Mandarin speakers, unlike the Dutch group, do not differentiate the apprehension for the entire event (i.e., the Event task) from the apprehension for event action (i.e., Action task) in Experiment 1, no cross-linguistic difference is found in Experiment 2. A potential reason for inconsistent results could be accounted by the changes of experimental design in the two experiments.

Specifically, the addition of the memory task after the first block could have raised participants the awareness of the spatial layout as well the relational information in the stimuli depicted. Stimuli in the memory task do not have limited presentation duration, so participants have ample time to inspect them. Participants in both language groups could have become familiarized with the location of the agent and the action/object in the scenes: the agent is always present in the upper part of the picture and the object/action element in the lower half. The apprehension task following the first block of memory trials could have then been affected by the previous longer exposure, leading to more predictable first fixation patterns. This could have been the case especially for Mandarin speakers as they no longer show the dissociation between the first fixation locations in the Action naming and Event description tasks in Experiment 2: Their enhanced awareness of the action being presented always in the lower parts of the stimuli may have cancelled out any potential agent-saliency effects, implied in Experiment 1. In addition, Experiment 2 has a lack of fillers which could have led to participants becoming more aware of the structure and the spatial layout of the event scenes. the spatial information of the agents, actions and objects in the stimuli could have been acquired prior to stimulus onset because of prior experience with very similar stimuli in the experiment. Further research needs to take these potential influences into account to gain a more systematic view on cross-linguistic effects.

Another possibility is that the apprehension pattern in Mandarin and Dutch speakers does not differ, which means that the pattern in Experiment 1 is “false positive” (also see the critical remarks concerning the analyses in section 3.4.3). This would mean that differences related to pro-drop (subject omission) do not affect the initial apprehension of event structure for different language speakers. More research with careful experimental design is warranted.

Another critical note concerns the methodology relates to the Recognition Memory task. Given the unexpected overall higher performance of Dutch speakers compared with Mandarin speakers, and the results based on between-subject data (for each block the memory data in the two tasks are from participants performing two different experiment lists, which differ in the sequence of tasks in block 2 and 3), it is hard to draw a definite conclusion on the agent memory from the current results. Indeed, there was evidence reporting cross-linguistic differences in memory in relation to variations in subject-encoding before. For instance, Fausey & Boroditsky (2010) compared English and Spanish speakers’ memory after viewing accidental events videos, while no verbal description task was involved. In the two languages, accidental events are encoded differently with respect to the agents (e.g., “She broke the vase accidentally” in English vs. “the vase broke accidentally” in Spanish. The latter leaving the agent implicit), suggesting different focus on the agent elements in accidental events. English speakers showed better memory of agents after viewing accidental event videos compared with Spanish speakers, which was interpreted as an effect of cross-linguistic differences in subject encoding on the non-linguistic memory of event agents, even in the condition where explicit verbal description was not required.

However, unlike the significant differences reported in Fausey & Boroditsky (2010), in our study, only a small difference in agent memory is found in the Mandarin group and only in the first memory task: Mandarin speakers’ memory of the agent element after performing the Event description task is better than the memory after the Action task. In addition, methodologically, using memory as an offline measure may overall not be suited for capturing any effect based on the early visual input under brief stimulus exposure, as many factors may have an effect on memory. First, the two apprehension tasks preceding the memory task differ in their requirement on the linguistic encoding on the agent element: Participants are not required to name the agent in the Action task. Second, the fixation patterns also imply a different focus of the visual uptake of event information under the two task demands: Speakers focus more on the Action/object element in the Action naming task. Both factors, i.e., whether to explicitly name the agent and where to allocate attention on the visual scene, could have affected the memory of the stimuli that are only presented for 300ms. The memory result cannot be directly correlated with cross-linguistic differences in agent saliency in our study. Further research can explore the relationship between first fixation locations and memory of the visual input during apprehension, checking whether they correlate at all as the first step.

6 Conclusion

The present study measures first fixation locations under 300ms brief stimulus exposure to gain insights into potential top-down effects of task demands and language background on the event apprehension process. For the first time, it shows that linguistic task demands requiring the encoding of different event elements can directly affect the locations of speakers' first fixations. In addition, first fixation locations do not predict the specificity and accuracy of speakers' verbal descriptions, nor do they correlate with what is mentioned first in the verbal descriptions of the stimuli. Thus, we argue that first fixations are the result of the initial apprehension of the gist of a visual scene, rather than the starting point guiding linguistic formulation processes. Finally, we obtain mixed effects regarding an influence of the language background of the viewers. Cross-linguistic differences in relation to pro-drop are only found in Experiment 1 showing that Mandarin speakers keep track of the agent information even when explicit linguistic encoding of the agent is not required. Further research is warranted to explore top-down effects of cross-linguistic variations on scene apprehension.

References

Allum, P. H., & Wheeldon, L. (2009). Scope of lexical access in spoken sentence production: Implications for the conceptual–syntactic interface. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(5), 1240–1255.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Biederman, I. (1981). On the semantics of a glance at a scene. In M. Kubovy & J. R. Pomerantz (Eds.), *Perceptual organization* (pp. 213-253). Routledge.

Bock, K., & Ferreira, V. S. (2014). Syntactically Speaking. In M.Goldrick, V. Ferreira & M. Miozzo (Eds.), *The Oxford Handbook of Language Production* (pp. 21-46). New York, NY: Oxford University press.

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

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

Buswell, G. T. (1935). *How People Look at Pictures: A Study of the Psychology and Perception in Art.* Chicago, Illinois: The 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.

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

Fausey, C. M., & Boroditsky, L. (2010). Who dunnit? Cross-linguistic differences in eye-witness memory. *Psychonomic bulletin & review*, *18*(1), 150-157.

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

Gerwien, J., & Flecken, M. (2016). First things first? Top-down influences on event apprehension. In A. Papafragou, D. Grodner, D. Mirman, & J. Trueswell (Eds.), *Proceedings of the 38th Annual Meeting of the Cognitive Science Society (CogSci 2016)* (pp. 2633–2638). Austin, TX: Cognitive Science Society.

Glanemann, R., Zwitserlood, P., Bölte, J., & Dobel, C. (2016). Rapid apprehension of the coherence of action scenes. *Psychonomic bulletin & review*, *23*(5), 1566-1575.

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

Griffin, Z. M. (2004). Why Look? Reasons for eye movements related to language production. In J.M.Hederson & F. Ferreira (Eds), *The interface of language, vision, and action : eye movements and the visual world* (pp. 213–248). New York, NY: Psychology Press.

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

Hafri, A., Papafragou, A., & Trueswell, J. C. (2013). Getting the gist of events: recognition of two-participant actions from brief displays. *Journal of Experimental Psychology. General*, *142*(3), 880–905.

Hayhoe, M. M., Shrivastava, A., Mruczek, R., & Pelz, J. B. (2003). Visual memory and motor planning in a natural task. *Journal of Vision*, *3*(1), 6.

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences*, *7*(11).

Henderson, J. M. & Ferreira, F. (2004). Scene perception for Psycholinguistics. In J. M. Henderson & F. Ferreira (Eds.), *The interface of language, vision, and action : eye movements and the visual world*. New York, NY: Psychology Press.

Henderson, J. M., Brockmole, J. R., Castelhano, M. S., & Mack, M. (2007). Visual saliency does not account for eye movements during visual search in real-world scenes. In R. P. G. van Gompel, M. H. Fischer, W. S. Murray, & R. L. Hill (Eds.), *Eye movements: A window on mind and brain* (pp. 537-562). Amsterdam, Netherlands: Elsevier.

Henderson, J. M., Weeks, P. A. . J., & Hollingworth, A. (1999). The effects of semantic consistency on eye movements during complex scene viewing. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(1), 210–228.

Hollingworth, A., & Henderson, J. M. (1998). Does Consistent Scene Context Facilitate Object Perception? *ournal of Experimental Psychology: General, 127(4), 398.*

Holmqvist, K. (2011). *Eye tracking : a comprehensive guide to methods and measures*. Oxford: Oxford University Press.

Hsiao, Y., Gao, Y., & MacDonald, M. C. (2014). Agent-patient similarity affects sentence structure in language production: evidence from subject omissions in Mandarin. *Frontiers in Psychology*, *5*, 1015.

Huang, B., & Yang, C.Y. (2013). Topic Drop and MCP. In *The 87th Annual Meeting of the Linguistic Society of America*. Boston, MA.

Hwang, H., & Kaiser, E. (2009). The effects of lexical vs. perceptual primes on sentence production in Korean: an on-line investigation of event apprehension and sentence formulation. In *Talk presented at the 22nd CUNY conference on sentence processing, Davis, CA*.

Jaeger, T. F., & Norcliffe, E. J. (2009). The Cross-linguistic Study of Sentence Production. *Language and Linguistics Compass*, *3*(4), 866–887.

Konnopka, A. E., & Brown-Schmidt, S. (2014). Message Encoding. In M.Goldrick, V. Ferreira & M. Miozzo (Eds.), *The Oxford Handbook of Language Production* (pp. 21-46). New York, NY: Oxford University press.

Lashley, K. S. (1951). *The problem of serial order in behavior (Vol. 21). Bobbs-Merrill.*

Levelt, W. J. M. (1999). Producing spoken language: A blueprint of the speaker. In C. M. Brown, & P. Hagoort (Eds.), The neurocognition of language (pp. 83-122). Oxford University Press.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation.*Cambridge, MA, US: The MIT Press.

Li, C. N., & Thompson, S. A. (1981). *Mandarin Chinese : a functional reference grammar*. University of California Press.

Myachykov, A., Garrod, S., & Scheepers, C. (2010). Perceptual priming of structural choice during English and Finnish sentence production. *Language & Cognition: state of the art,* 54-72.

Norcliffe E., Konopka A.E. (2015) Vision and Language in Cross-Linguistic Research on Sentence Production. In: Mishra R., Srinivasan N., Huettig F. (Eds) *Attention and Vision in Language*

Processing. Springer, New Delhi

Norcliffe, E., Konopka, A. E., Brown, P., & Levinson, S. C. (2015). Word order affects the time course of sentence formulation in Tzeltal. *Language, Cognition and Neuroscience*, *30*(9), 1187–1208.

Osgood, C. E., & Bock, J. K. (1977). Salience and sentencing: Some production principles. In *Developments in research and theory* (pp. 89-140). Lawrence Erlbaum.

Paul, W. (2017). *Null subject, null topics and topic prominence in Mandarin Chinese and beyond*. *Studies on Syntactic Cartography* (Vol. 1). China Social Sciences Press.

Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory*, *2*(5), 509–522.

Potter, M. C., & Levy, E. I. (1969). Recognition memory for a rapid sequence of pictures. *Journal of Experimental Psychology*, *81*(1), 10–15.

Schyns, P. G., & Oliva, A. (1994). From Blobs to Boundary Edges: Evidence for Time- and Spatial-Scale-Dependent Scene Recognition. *Psychological Science*, *5*(4), 195–200.

Sebastian, S., Norcliffe, E., Konopka, A., Van Valin, R. D., & Levinson, S. C. (2013). Dependencies First: Eye Tracking Evidence from Sentence Production in Tagalog. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society* (pp. 1265–1270).

Senzaki, S., Masuda, T., & Ishii, K. (2014). When Is Perception Top-Down and When Is It Not? Culture, Narrative, and Attention. *Cognitive Science*, *38*(7), 1493–1506.

Smith, M., & Wheeldon, L. (1999). High level processing scope in spoken sentence production. *Cognition*, *73*(3), 205-246.

Treisman, A. (2006). How the deployment of attention determines what we see. *Visual Cognition*, *14*(4–8), 411–443.

Yarbus, A. L. (1967). Eye Movements During Perception of Complex Objects. In *Eye Movements and Vision* (pp. 171–211). Boston, MA: Springer US.

Zwitserlood, P., Bölte, J., Hofmann, R., Meier, C. C., & Dobel, C. (2018). Seeing for speaking: Semantic and lexical information provided by briefly presented, naturalistic action scenes. *PLOS ONE*, *13*(4).

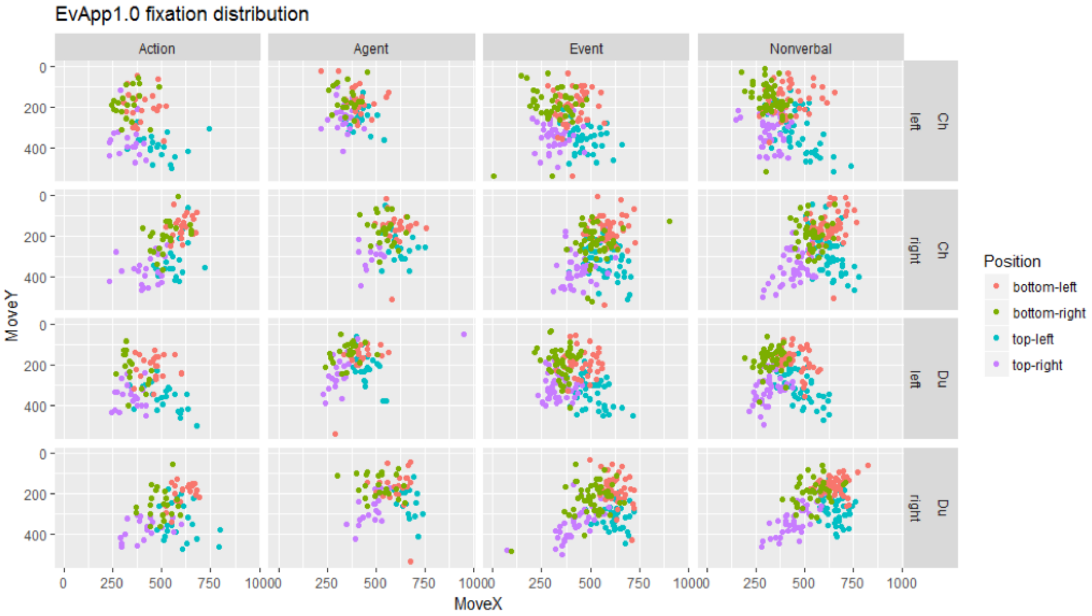
Appendix 1. List of stimuli in Experiment 1

Nonverbal task	Event description task	Action / Agent naming task
A man cuts paper in half	A woman cuts circle	A woman breaks chocolate
A woman draws a flower	A woman cuts an apple	A woman cuts cucumber
A woman peels a mandarin	A woman peels banana	A woman puts a puzzle together
A man peels a potato	A man rolls wool	A man opens jam jar
A woman fills out a form	A woman tears paper	A woman stamps paper
A woman folds towel	A woman grinds spice	A man grates cheese
A man makes paper plane	A man highlights text	A man plays a drum
A woman puts toothpaste on toothbrush	A woman opens coke can	A woman stirs soup
A woman cleans glasses	A man paints cup	A woman builds lego tower
A man cleans a knife	A woman punches paper	A woman opens a can
A woman cleans a mirror	A woman lights candle	A woman pours coke
A woman knits a scarf	A woman beats cream	A woman opens a wine bottle
A woman salts soup	A man polishes glass	A woman spread Nutella
A woman staples paper	A woman pours water from flask	A woman measures a box
A woman beads a necklace	A woman opens a letter	A man mixes cards
A woman reads a book	A woman opens a can opener	A woman wipes table

Appendix 2 List of stimuli in Experiment 2

Agent naming task	Action naming task	Event description task
A woman breaks chocolate	A woman draws flower	A woman cuts a circle
A woman builds Lego	A woman opens a can	A woman cuts an apple
A woman cuts cucumber	A woman pours coke	A man cuts paper
puts a puzzle together	A woman tears paper	A woman peels a banana
A woman rolls wool	A woman folds towel	A woman peels a mandarin
A woman grinds spice	A woman highlights text	A woman fills a form
A woman punches hole	A woman opens a jam	A woman makes a plane
A man spreads Nutella	A woman paints a cup	A man opens coke
A man stamps paper	A woman toothpaste	A woman peels a potato
A woman lights a candle	A man beats cream	A woman opens a wine
A woman opens a letter	A man cleans a knife	A woman opens a can opener
A man grates cheese	A man knits scarf	A woman beads necklace
A woman measures a box	A woman salt soup	A man cleans glasses
A woman plays drum	A woman staples papers	A woman cleans a mirror
A man polishes glass	A woman stir soup	A man mixes cards
A woman wipes table	A woman reads a book	A woman pour water

Appendix 3. A descriptive scatterplot of FFLs in Experiment 1.



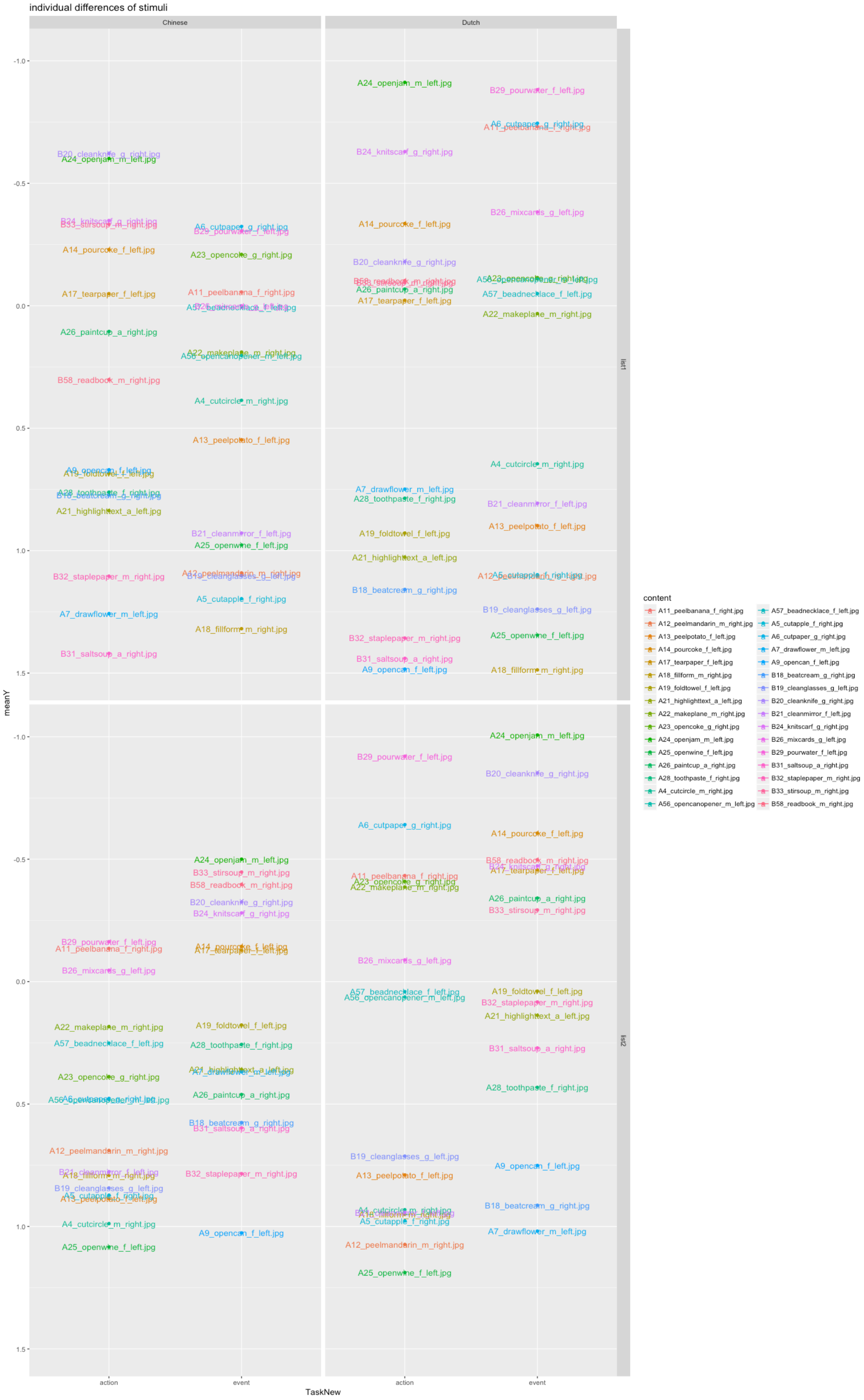
The scatterplot demonstrates the distribution of FFLs on the X-Y dimension of stimuli, separated by stimuli that are presented on one of the four corners of the screen (bottom-left, bottom-right, top-left and top-right) and by two agent orientations (agent on the left and on the right). This plot further supports the several ideas mentioned in the main body of the thesis:

Idea 1. The presentation location of a stimulus influences the FFLs: participants tend to fixate on an area on a stimulus that is closer to the fixation cross presented in the middle of the screen. For instance, for a picture showing on the bottom left of the screen, fixations tend to cluster at the top-right corner of the picture (the red dots in the scatterplot).

Idea 2. Fixations tend to cluster on informative regions, and it is rare for an intended fixation to locate on the blank area. For example, for agent-left stimuli (row 1 and 3 in the scatterplot), it is rare to find fixations located on the right side of a picture, which is the blank area. It further licenses that our analysis focusing on the Y-coordinates of FFLs is sufficient as an index for FFLs and further, for apprehension.

Idea 3. The variations modulated stimuli presentation location further boost the robust effects of task demands we find, because even when the presentation location modulates FFLs in a relatively consistent manner (Idea 1), participants are still paying extra efforts to meet the task demands. For instance, the FFLs in the Agent Task (column 2 in the scatterplot) tend to cluster towards the Agent element (higher on the Y-coordinates) compared with e.g., FFLs in the Nonverbal Task. It further indicates that apprehension is a flexible process that is modulated by task demands: participants paid extra effort (i.e., longer scan path distance than default) to fixate on the area that is needed for the task at hand.

Appendix 4. FFLs for individual stimulus



This scatterplot demonstrates the mean Y-coordinates of FFLs (normalized) for each stimulus in Action and Event Task in Dutch and Mandarin group. Here, a clear dissociation of two groups of pictures can observed (e.g., for Dutch Action condition in List 1, a group of pictures cluster above zero but the other half cluster below zero). A *post hoc* check matches this pattern with stimuli presentation locations: the stimuli group on the upper cluster in the scatterplot are stimuli that are presented on the bottom half of the screen in the experiments. Similar to Appendix 3, it shows that participants tend to fixate on a stimulus area that is closer to the fixation cross (e.g., fixations cluster on the bottom half when the pictures are presented on the upper half of a screen).

More interestingly, the dissociated pattern of FFLs modulated by stimulus presentation location is more obvious in the Dutch group than Mandarin group (e.g., the gap in list 1 for Dutch group), which indicates that Dutch people pay less effort in saccades and visual encodings compared with Mandarin group. However, the memory accuracy for Dutch group is significantly higher than the Mandarin group, indicating that they pay less effort in saccade movements but are more capable of retrieving the visual memory.