Coordination of Vision and Language in Cross-Modal Referential Processing

Moreno I. Coco

School Of Informatics

Institute of Language, Cognition and Computation

University of Edinburgh

Doctor of Philosophy (PhD)

2011

Abstract

This thesis investigates the mechanisms underlying the formation, maintenance, and sharing of reference in tasks in which language and vision interact. Previous research in psycholinguistics and visual cognition has provided insights into the formation of reference in cross-modal tasks. The conclusions reached are largely independent, with the focus on mechanisms pertaining to either linguistic or visual processing.

In this thesis, we present a series of eye-tracking experiments that aim to unify these distinct strands of research by identifying and quantifying factors that underlie the cross-modal interaction between scene understanding and sentence processing. Our results show that both low-level (image-based) and high-level (object-based) visual information interacts actively with linguistic information during situated language processing tasks. In particular, during language understanding (Chapter 3), image-based information, i.e., saliency, is used to predict the upcoming arguments of the sentence, when the linguistic material alone is not sufficient to make such predictions.

During language production (Chapter 4), visual attention has the active role of sourcing referential information for sentence encoding. We show that two important factors influencing this process are the visual density of the scene, i.e., clutter, and the animacy of the objects described. Both factors influence the type of linguistic encoding observed and the associated visual responses. We uncover a close relationship between linguistic descriptions and visual responses, triggered by the cross-modal interaction of scene and object properties, which implies a general mechanism of cross-modal referential coordination. Further investigation (Chapter 5)

shows that visual attention and sentence processing are closely coordinated during sentence production: similar sentences are associated with similar scan patterns. This finding holds across different scenes, which suggests that coordination goes beyond the well-known scene-based effects guiding visual attention, again supporting the existence of a general mechanism for the cross-modal coordination of referential information.

The extent to which cross-modal mechanisms are activated depends on the nature of the task performed. We compare the three tasks of visual search, object naming, and scene description (Chapter 6) and explore how the modulation of cross-modal reference is reflected in the visual responses of participants. Our results show that the cross-modal coordination required in naming and description triggers longer visual processing and higher scan pattern similarity than in search. This difference is due to the coordination required to integrate and organize visual and linguistic referential processing.

Overall, this thesis unifies explanations of distinct cognitive processes (visual and linguistic) based on the principle of cross-modal referentiality, and provides a new framework for unraveling the mechanisms that allow scene understanding and sentence processing to share and integrate information during cross-modal processing.

Acknowledgements

The first big thanks goes to my supervisor Frank Keller, a brilliant and yet extremely modest scientist, who has continuosly followed my work with inspiring ideas, precious suggestions, constructive criticism and constant passion. I am very proud and grateful to have been a PhD of yours, and I could not thank you more for all of this. The second thanks goes to John Henderson whose ideas have given an unvaluable contribution to my work.

Also I would like to thank my examiners Gerry Altmann and Jon Oberlander for the interesting perspectives they gave me to look at my work. Many people have given me input throughout my PhD; among which, in random order, I would like to thank: Fernanda Ferreira, Manabu Arai, Matthew Crocker, Robin Hill, Frances Wilson, Helene Kreysa, Mark Steedman, Jeff Mitchell, Richard Shillcock, Tim Smith, George Malcolm, Antje Nuthmann, Parag Mital, Jens Apel and Ben Allison for the interesting discussions and valuable feedback at some point during the process. Any error or omission remain my own.

I would like to thank also my colleguaes Tom, Joel, Stella, Sasa, Trevor, Michael, and many more at the ILCC, that have kept me company all along these years. From the Italian community in Edinburgh, a thank goes to Daniele Sepe, Daniele Fanelli, Andrea, Antonella, Patricia, Massimo, Leonardo, Laura, Marco and Andrew for the various delicious dinners and discussions together. I would also thank Jamie, Gregor, Murdo, Al, Elena, Tim, Davide and Simone for the many fantastic live-music nights we have been in together, and for keeping the spirit of folk alive.

A final thank goes to Erida for her love, care and spiritual support, that has given me the stability to go through this journey without losing faith in myself.

Author's Declaration

I declare that this thesis was composed solely by myself in the School of Informatics at the University of Edinburgh, UK, during the period of October 2007 to January 2011, and that the work contained therein is my own. The copyright of this thesis belongs to the author under the terms of the United Kingdom copyright acts. Due acknowledgement must always be made of the use of any material contained in, or derived from, this thesis.

Moreno I. Coco

Contents

List of Figures

Chapter 1

Introduction

1.1 Cross-modal synchronous processing

The growing body of multi-modal data powered by new tools and technologies has brought forward questions about the role of different modalities and their interaction in cognition. A central challenge in cognitive science today remains the identification and quantification of cross-modal mechanisms that mediate the synchronous processing of multi-modal information.

Cognitive modalities are in synchronous processing during tasks requiring the coordination of multi-modal information. For example, when we are driving a car or simply walking, our visual system has to share and coordinate information with the associated motor-actions; thus looks to the road turn in our visual field have to be coordinated with steering movements of the wheel. It follows that access to certain aspects of the visual information occurs in synchrony with associated motor-responses.

Even if the majority of our daily tasks need the synchronization of different cognitive processes, very little is known about the mechanisms allowing such cross-modal interaction.

By exploring the temporal mechanisms regulating this cross-modal processing, we gain additional insight into the individual cognitive processes involved, while aiming at a unified explanation of the underlying cognitive architecture. We believe that this knowledge can be used to integrate multi-modal processes, in order to conceive models, computational tools and applications which can exploit such richness. Such

integrated understanding requires moving beyond mere correlation to identify plausible underlying dynamic mechanisms directly from experimental or behavioral data, that involve two or more modalities in synchrony. We must examine the emerging relationships between different cognitive processes at various temporal and spatial scales, and identify the links between top-down and bottom-up factors that influence their interdependence. A first step to achieve these goals is to discover general mechanisms of cognition which unify evidence gathered by different disciplines interested in cognition.

Cross-modal synchronous processing is a broad and largely unexplored topic which cannot be covered in a single thesis; thus, here, we focus on **referentiality**, a key mechanism that enables multi-modal interaction. Reference arises at the interface between different modalities. As such, it is subject to influences from various cognitive processes. Moreover, it is the core of all shared representations which form the building blocks of meaning. By investigating how referential information is shaped across modalities by various factors and mechanisms, we can begin to uncover the emergence of cross-modal integration.

1.2 Cross-modal referentiality in multi-modal interaction

Referents can be defined as cognitive entities with logical, visual, linguistic, and other components from distinct modalities, which unify the perception of a real world counterpart.

This thesis explores the claim that multi-modal perceptual processing occurs over a shared referential interface. In order to explain the multi-modal interaction of vision and language during tasks demanding synchronous processing, we assume the principle of cross-modal referentiality. We examine which visual and linguistic factors contribute to the formation and maintenance of cross-modal referentiality. Our aim is to unravel the shared cognitive mechanisms allowing visual attention and sentence processing to be coordinated during this synchronous interaction.

Insights about the existence of a shared referential interface between vision and

language have come, rather independently, from research in psycholinguistics and visual cognition.

On one hand, a psycholinguistic paradigm of eye-tracking research (**V**isual **W**orld **P**aradigm, VWP), where sentence understanding is investigated concurrently with a visual context, has shown clear links between visual and linguistic referential information (e.g. Altmann & Kamide 1999; Altmann & Mirkovic 2009; Crocker *et al.* 2010; Knoeferle & Crocker 2006; Spivey-Knowlton *et al.* 2002; Tanenhaus *et al.* 1995). A major conclusion reached by this paradigm is that the interpretation of linguistic information is mediated by the visual information in the context; and this interaction can be observed in the visual responses launched at the referents depicted in the visual context. By focusing on linguistic phenomena, however, these studies have largely underestimated mechanisms and factors attributable to visual processing, hence restricting their conclusions to linguistically motivated explanations.

On the other hand, research in visual cognition has gradually gathered evidence supporting the idea that the allocation of visual attention during goal-directed tasks is driven by object-based, referent-dependent, top-down processes (e.g. Findlay & Gilchrist 2001; Henderson & Hollingworth 1999; Henderson 2003; Nuthmann & Henderson 2010; Zelinsky & Schmidt 2009). The leading idea is that contextual information relating the visual referents of a scene (e.g. their semantics) is utilized to guide visual attention. Furthermore, this guidance seems to be modulated by processing of linguistic information, and by the nature of the task performed (Castelhano *et al.*, 2009; Schmidt & Zelinsky, 2009). Crucially, the tasks explored in the visual cognition literature were mainly visual, e.g. search. To have a thorough understanding of the mechanisms underlying visual attention, however, the investigation of visual tasks involving sentence processing cannot be neglected.

1.3 Background: Referential information

Referentiality allows the mapping from external world-entities to internal cognitive counterparts. The existence of such a mechanism enables different cognitive processes to happen. For instance, the ability of recognizing what is and is not edible, or whether we are confronting a prey or a predator, is strictly related to the capacity of linking the perception of real-world entities with their referential cognitive counterpart.

Referents are used to recognize, categorize and communicate information about the real world, and constitute a primary thread tying together perception, planning and action (Steedman, 2002). In fact, through the referents and their perceptual properties, we categorize, understand and perform actions afforded by them (Gibson, 1977). Imagine, for example, that we want to open a door. In order to perform this action, we need to know which referents compose this event, i.e. the DOOR[1] and the HANDLE, the perceptual properties involved, e.g. the DOOR and the HANDLE are movable objects, and the affordances inferred, i.e. by moving the HANDLE we open the DOOR.

Each modality sources different types of information about the referents, e.g. the door is red (vision) and heavy (haptic); nevertheless, this perceptual diversity is integrated within the same referential identity, a *door*. Moreover, it is interesting to note that a referent is conceptually unique (Wittgenstein, 1921), while the perceptual patterns by which an object is classified are fuzzy (Labov, 2004). For example, different MUGS can vary in size (small or tall) or color (red or blue) as objects, but they will be classified under the same type. Thus, reference can extend over different perceptual realizations of the same world-entity while preserving unique denotation. This implies that, even if the physical features of an object, or its spatial location, change it would afford the same set of actions, as long as its referential identity is maintained. Thus, if a MUG is on a COUNTER or on CHAIR, it would still afford the drinking action.

The integration of cross-modal referential information is essential to synchronize processing across modalities during tasks requiring the coordination of multi-modal information. The action of opening a door, for instance, requires the coordination between visual attention and motor-action on both referents the DOOR and the HANDLE: visual attention retrieves information about the HANDLE, while motor-actions adapt the hand position to perform grasping on it (Land, 2006). Even if the identity of referents is shared across modalities, the way the perceptual information is processed, and the mechanisms utilizing such information, have modality-specific effects. In this thesis, we explore the mechanisms of cross-modal referentiality underlying the interaction of vision and language. Thus, we review the concept of referentiality in both modalities, before passing onto the evidence about mechanisms that could relate their interaction.

[1] We use the following typography DOOR to indicate the real-world referent and *the door* for its linguistic counterpart.

1.3.1 Referentiality in language

The existence of referential identity makes it possible not only to recognize and categorize new instances of the same world-entity but also, and more importantly, to communicate about them. It is here that the notion of referentiality overtly interacts with language processing. Referents in language processing can be imagined as discourse pointers toward real-world counterparts. Interestingly, not everything that is linguistically processed has a clear referential counterpart in perception. For instance, concrete nouns usually refer to perceivable objects (e.g. cup, dog, mug), whereas verbs refer to events, where different aspects of perceptual information are conveyed (e.g. run, make, drive). However, despite these different levels of abstraction, it seems reasonable to frame language as the faculty for structuring referential information. By taking this approach, language is naturally contextualized within cognition as the interface organizing multi-modal perceptual information. The main challenge posed to this way of looking at language processing is the *grounding* problem (Barsalou, 1999; Gorniak & Roy, 2007; Roy, 2005; Stevan, 1990); i.e. finding mechanisms of correspondence between symbols of language and contents of perception. Neurophysiological research has shown that there are specific associations between linguistic processing and neural circuitry, a verb like *kick* elicits activity in the motor cortex (Pulvermuller, 2005). Moreover, Rizzolatti & Arbib (1998) have identified structures in the mirror neurons where this integration might take place. However, beside proving that language is indeed wired in our brain, the grounding problem remains largely unsolved.

In order to investigate language processing in terms of grounding, we have to break down the problem, at least, in three parts: the formation of referentiality, language grounding, and the mechanisms of cross-modal interactivity. The formation of a referential interface takes place during the first stages of cognitive development (Barsalou, 1999) and regards the question of dividing the fuzzy categories of perception into finite conceptual referents (Labov, 2004). After the referential mapping has been established, the referents are organized into event structures representing contextual expectation of our external world (Altmann & Mirkovic, 2009). During this period, language assigns the word/sound labels to the conceptual referents while building links to the different perceptual representations of that entity across modalities (Roy,

2005; Tomasello, 2003). The mechanisms of cross-modal interactivity, arising during language grounding, allow language processing to share and integrate grounded referential information across different modalities.

In this thesis, we assume the existence of a shared referential interface linking perceptual information across modalities, and, on the basis of that, we explore the mechanisms allowing cross-modal interaction. Especially, we explore how linguistic reference interacts with visual reference during situated language processing, while unraveling the visual and linguistic factors involved and the patterns emerging as a result of their interaction.

1.3.2 Referentiality in vision

A definition for the notion of reference in visual perception is more difficult to formulate than in sentence processing. The main reason is that in sentence processing, a referent is directly linked to a word, which can be uniquely identified, whereas in vision the debate about what an object is, and how it could be identified is still open. A main distinction emerging in the visual cognition literature to classify what is, from what is not referential information, is the difference between image-based (e.g. Itti & Koch 2000a) and object-based information (e.g. Nuthmann & Henderson 2010).

Image-based information, i.e. *saliency*, is made of primitive visual features, such as color, intensity and orientation, of the scene as a whole. Thus, this information does not imply referentiality, as it is purely based on the raw visual stimulus. Within this approach, there has been an attempt to use this information to derive an intermediate, referent-like, representation (i.e. proto-objects Walther & Koch 2006). However, even if proto-objects can be computationally found using image-based information, they do not seem to have any cognitive relevance for visual attention during performance of visual tasks (Nuthmann & Henderson, 2010).

Object-based information, as the name suggests, is the information carried by the objects composing a certain scene. In this approach, objects are recognized by integrating spatial, semantic and statistical information of the scene context in which they occur (Bar, 2004; Galleguillos & Belongie, 2010). Evidence of the important role played by contextual information comes from experimental, and computational research in visual cognition. Experimental work has shown that object-based, contextually driven,

information is utilized to efficiently allocate visual attention during goal-directed tasks, such as search (Brockmole & Henderson, 2006; Rayner *et al.*, 2009; Zelinsky *et al.*, 2008). Furthermore, computational models of visual attention, which integrate contextual information, outperform purely image-based models on such tasks (Ehinger *et al.*, 2009; Judd *et al.*, 2009; Torralba *et al.*, 2006).

In this thesis, since we are interested in cross-modal reference during situated language processing tasks, we adopt an object-based perspective, which deals with referential information processing more deeply than the image-based approach. Moreover, we take a more 'linguistic' view to interpret visual objects and contextual scene information. An object, for us, is defined with respect to a content word that can be used to refer to it; the contextual scene information, instead, is conceptualized as the spatial, semantic and statistical relations holding between the referent objects of a certain scene.

1.3.3 Situated language processing

Language processing does not occur in isolation. Usually, it depends on contextual information that is processed concurrently: linguistic, e.g. previous sentences of a discourse, or non-linguistic, e.g. an image we are watching. Research in formal linguistics has mainly focused on mechanisms of syntactic dependency, e.g. anaphora or co-referent resolution (Reinhart, 1983), happening at the level of syntactic clauses and trying to capture how linguistic information is selected (semantics), encoded (syntax) and maintained (discourse). The mechanisms found focused around formal observations based solely on linguistic information. However, when sentence processing is observed in relation to a non-linguistic context, e.g. visual information, linguistic explanations are not sufficient. Within a multi-modal framework, in fact, different types of contextual information concurrently interact and integrate to sentence processing.

A psycholinguistic eye-tracking paradigm of studies investigating linguistic phenomena situated in visual contexts is the Visual World Paradigm VWP (e.g. Altmann & Kamide 1999; Spivey-Knowlton *et al.* 2002; Tanenhaus *et al.* 1995). This paradigm is based on the assumption that eye-movements launched on a visual context during linguistic tasks (mainly comprehension) can show underlying mechanisms of linguistic processing. In a VWP study, the visual context, usually an array of objects or a clip-art

scene, has a direct correspondence with the ongoing linguistic processing. Typically, linguistic referential information of the sentence matches depicted information of the visual context. This correspondence makes it possible to investigate eye-movement responses time-locked to linguistic regions of interest (ROI). Thus, visual objects looked at, before and after a linguistic ROI, are interpreted as responses to ongoing linguistic processing.

Overall, these studies support the hypothesis that visual information is integrated during sentence processing, and utilized to situate (Crocker *et al.*, 2010) linguistic mapping. The integration has been observed, especially during sentence comprehension, at different levels of processing: from prosody (Snedeker & Yuan, 2008), where intonational patterns are combined with visual information to resolve syntactic referential ambiguity; to lexical semantics, where linguistic predictions based on verbal or thematic information (Altmann & Kamide, 1999; Knoeferle & Crocker, 2006), e.g. *eat* expects an edible direct object *the apple*, are anticipated as visual responses to contextually appropriate objects, e.g. looks to APPLE before the linguistic referent is mentioned.

A shortcoming of the VWP paradigm, however, has been the strict linguistic perspective used to investigate and interpret the integration.VWP experiments are mainly designed to test specific linguistic phenomena, e.g. syntactic priming (Arai *et al.*, 2007), thus manipulations are focused on the linguistic material. Within this approach, visual responses are understood as consequences to linguistic processing, hence assuming a uni-directional dependence (from language to vision) between the visual objects fixated and the linguistic material processed. However, a consistent body of studies in the visual cognition literature has shown that there are several factors, most of which independent from linguistic processing, influencing how visual attention is deployed. These mechanisms have to be understood in the context of situated language processing, in order to clearly interpret the combined output of linguistic and visual responses.

1.3.4 Mechanisms of visual attention

In the visual cognition literature, as already mentioned in section 1.3.2, the debate about the mechanisms guiding visual attention centers on the distinction between low

(e.g. Baddeley & Tatler 2006; Itti & Koch 2000b) vs high (e.g. Brockmole & Henderson 2006; Henderson 2007) level mechanisms; also distinguished as image vs object-based mechanisms of visual attention (Nuthmann & Henderson, 2010). Conceptually, low-level, or bottom-up, mechanisms are linked to physical properties of the image (e.g. color), whereas high-level, or top-down, mechanisms are based on cognitive structures attached to the individual objects (e.g. a mug is often in relation to kitchen counters) forming it. At the foundation for models of bottom-up driven visual attention is saliency (Itti & Koch, 2000b; Parkhursta *et al.*, 2002): an aggregated measure of visual information, based on primitive features of the image, such as color, intensity and orientation, computed at different spatial scales. The saliency map, in combination with a winner-take-all[1] network and the inhibition of return mechanism[2], can be used to predict the scanning sequence of fixations during free viewing tasks (Walther & Koch, 2006). The strength of models based on saliency is their reliance on visual primitive features, which have neurophysiological soundness (Moore, 2006). Within this approach, however, it is not clear how high-level cognitive knowledge about the scene is utilized. A scene, in fact, is not a meaningless combination of primary features, but rather it has a specific configuration consisting of individual objects, which are usually semantically and functionally related. The cognitive relevance of a scene becomes especially evident when vision is actively used to perform tasks requiring access to knowledge.

Experimental evidence, in support of a cognitive driven, object-based, approach to visual attention, comes from the active vision perspective (Findlay & Gilchrist, 2001; Land, 2006; Noton & Stark, 1971; Yarbus, 1967). The main assumption of this approach is that the deployment of visual attention is mainly bounded to the type of task performed (Castelhano *et al.*, 2009), and the knowledge structures activated in order to complete it (Malcolm & Henderson, 2009, 2010). For instance, if our task is to find a mug in a kitchen (visual search), we inspect only regions of the scene that are contextually relevant (Neider & Zelinsky, 2006; Schmidt & Zelinsky, 2009; Yang & Zelinsky, 2009), e.g. the COUNTER but not the CEILING, despite their saliency within the image (Henderson, 2007). Only objects that are cognitively relevant to the task are going to be fixated. The cognitive relevance approach (Henderson *et al.*, 2009a; Nuthmann &

[1] The highest value of the saliency map is selected.

[2] The likelihood of re-fixating the same location, after having just inspected it, is very low.

Henderson, 2010) is based on the assumption that we have cognitive knowledge about the visual objects, and this knowledge guides visual attention. Nevertheless, as already mentioned, it is unclear how representations about individual objects are, in the first place, generated from primary visual features, and which cognitive principles make categorical inference possible.

There is experimental evidence reconciling some aspects of the dichotomy, low vs high-level. For instance, objects and saliency are spatially correlated, thus the locations of a scene that are rich in objects have also high saliency (Einhuser *et al.*, 2008; Elazary & Itti, 2008); or that categorical search is also guided by low-level visual similarity[1] (Alexander *et al.*, 2010). Nevertheless, the mechanisms allowing the interaction and integration between low and high level information are still under debate.

The main aim of this thesis is to unify the evidence from psycholinguistic studies on situated language processing (VWP) with those found in visual cognition to uncover the connections and cross-modal mechanisms allowing the formation, maintenance and sharing of multi-modal referential information between visual and linguistic processing during tasks requiring their synchronous interaction.

1.4 Central Claims

Neither of the fields discussed above has managed to provide an exhaustive theory of cross-modal referentiality; where the mechanisms of integration, the factors involved, and the type of task performed during synchronous visual and linguistic processing are accounted for by the same framework.

In order to provide a unified theory of cross-modal referentiality, we must integrate empirical evidence gathered in both fields: psycholinguistics and visual cognition. To do so, we put forward four main claims.

The first claim is that visual attention and sentence processing interact bi-directionally during tasks demanding synchronous processing. So, mechanisms known to influence the response of one modality are expected to modulate the response of the other modality. We demonstrate this dependence in a series of situated language understanding eye-tracking experiments. We show that low-level, image-based, scene infor-

[1] An effect that has also been found within the VWP (Huettig & Altmann, 2007).

mation, i.e. *saliency*, is actively used to make predictions about upcoming arguments of the sentence. When the linguistic information accumulated during the understanding process is not sufficient, sentence processing resorts to saliency information in the scene.

The second main claim is that object-based information modulates the type of sentences observed and the associated pattern of visual responses. We investigate this hypothesis in two language production experiments situated in photo-realistic scenes. By looking at production, rather than comprehension, we give more prominence to visual attention which sources referential information for the sentence processor. Moreover, by situating the language generation task in photo-realistic scenes, we are able to analyse more realistic visual responses. We demonstrate that the *clutter* of the scene and the *animacy* of the objects described can modulate the cross-modal interaction of sentence processing and visual attention.

The third central claim is that cross-modal integration is established through *coordination* of referential information. We demonstrate a consistent positive correlation of similarity between sentences and associated scan patterns. This effect is robust across the different phases of language generation, both within and across scenes. The cross-modal similarity found across different scenes highlights an object-based, rather than image-based, interaction between sentence processing and visual attention.

The fourth claim is that the nature of the task performed affects cross-modal interaction. We distinguish between *single* modality task (e.g. visual search) where only one modality (e.g. visual attention) is actively involved, and *multi*-modal task (e.g. scene description), where different modalities demand synchronous processing. We find that a single-modality task requires less visual processing than a multi-modal task. Moreover, we show that in a single modality task participants display a lower scan pattern similarity than during a multi-modal task. The synchronization between visual and linguistic referential information leads to a stronger scan pattern coordination. Coordination arises when multi-modal information needs to be synchronized across different cognitive processes.

1.5 Overview of the thesis and Contributions

In Chapter 2, we present our methodology of investigation situated in the context of previous approaches. We also give details on the experimental procedure, the response measures utilized and the reasons for choosing linear mixed effect models to perform inferential statistics.

In Chapter 3 (Experiments 1-3), we investigate the interaction between low-level visual (saliency) and linguistic (intonational break) information during the resolution of syntactic ambiguity. Our approach differs from previous work in the VWP, which has investigated sentence understanding situated in a visual context (e.g. Tanenhaus *et al.* 1995), without, however, questioning whether visual mechanisms are actively involved during sentence processing. Through our findings, we demonstrate that mechanisms of visual attention are important during situated language processing, as they directly interact with linguistic factors. The results show that when the linguistic material processed is not sufficient to generate a full prediction, (e.g. around verb site), saliency is utilized to predict upcoming arguments. This suggests that image-based low-level information can actively inform the sentence processor when the linguistic material is unable to guide visual attention towards precise visual referents of the scene, on its own.

In situated language understanding tasks, visual attention plays only a marginal role. Thus, in Chapter 4 (Experiments 4-5) we decided to investigate situated language *production* tasks, where visual attention is expected to play a more active role. Through the use of photo-realistic scenes, we can explore more realistically the impact of scene referential information on cross-modal interaction between visual and linguistic responses. In particular, we show how the visual density of the scene, i.e., *clutter* (Rosenholtz *et al.*, 2007), and the *animacy* of the object cued for description (McDonald *et al.*, 1993) influence the type of sentence encoded and the associated visual responses. One main hypothesis is that scene clutter should facilitate language generation. Regarding object animacy, we expect that an inanimate object will be described in spatial relation with another object, whereas an animate object will be primarily chosen as the subject of an action. Beside these independent effects, we expect interactions to emerge between clutter and animacy. In particular, the linguistic encoding of animate objects should benefit from higher clutter, as more visual information can

be used to contextualize their role/action in the scene. Our results show more looks on the scene during mention of the animate referent in cluttered scenes, thus confirming that the description of animate objects benefits from the density of visual information. In contrast, in minimal scenes, participants faced more difficulties in encoding animate referents; and visual attention had to source additional information from the animate object itself to feed the ongoing sentence. The convincing evidence of cross-modal interaction gathered in Chapters 3-4 strongly motivate us to quantify more precisely the relation between visual and linguistic referential information. Inspired by literature in visual cognition showing coordination of gazes (represented as scan patterns), during multi-modal tasks, such as dialogue (Richardson *et al.*, 2007) and motor-action (Land, 2006), we test whether coordination emerges also between sentence processing and visual attention. In Chapter 5 (Experiment 6), we demonstrate cross-modal coordination between visual attention and sentence processing. Our main hypothesis is that similar sentences are associated to similar scan patterns, and our results confirm it. Across different phases of the language generation task, we observe a positive correlation of similarity between sentences and scan patterns.

Throughout the thesis, we investigate situated language processing tasks, where cross-modal interaction is a requirement to perform the task. However, tasks can differ by the level of cross-modal interaction they require. Some tasks, such as search, demand the active engagement of a single modality. On the other hand, more complex tasks, e.g. scene description, require the involvement of more than one cognitive modality. So, in Chapter 6 (Experiments 7-8), we investigate how the cross-modal nature of the task affects visual responses. We analyze three tasks: visual search, object naming and scene description. They vary by the degree of cross-modal interactivity demanded. Our main hypothesis is that naming and description, being tasks that demand explicitly cross-modal interaction, should be characterized by longer visual responses than search, which is an intrinsically visual task. This difference in temporal processing is confirmed by our empirical results, and seems to be due to cross-modal referential *integration*, which is a key feature of synchronous processing. The comparison of the scan patterns across tasks reveals that the multi-modal tasks have higher scan pattern similarity than the single modality task. This effect is due to the synchronization of referential information across modalities (visual objects and linguistic referents). In

search, instead, more variability can arise as the cognitive control of visual attention decays after an early phase of contextually driven guidance (e.g. Torralba *et al.* 2006).

Finally, in Chapter 7, we summarize our findings and discuss the implications that cross-modal referential interaction entails. We also propose future research to expand and strengthen our understanding about the role of cross-modal referential interface during synchronous processing.

1.6 Collaborations and Publications

The experiments presented in Chapter 4 were published in Coco & Keller (2009, 2010b), The experiment presented in Chapter 4 have published in Coco & Keller (2010a). The thesis has also benefited from the comments of the audience of CogSci-09/10, CUNY-09/10, AMLaP-08/09 and HSI-09 which have refined and strengthened the content and form of my ideas. I am grateful to Jeff Mitchell to have provided the code to calculate LSA similarity of sentences for Experiment 6 in Chapter 5, and George Malcolm for designing and running Experiment 7 in Chapter 6.

Chapter 2

Methodology, Tools and Analysis

2.1 Introduction

In this chapter, we describe the methodology of investigation used to support the hypotheses discussed in the thesis. In section 2.2, we begin by illustrating the importance of eye-movements for the study of cognitive processes, while presenting components, definitions and measures commonly adopted in eye-tracking research. Since this thesis focuses on the interaction of vision and language, in section 2.3 we contextualize eye-movements analysis during language processing, and we discuss theoretical and technical implications of this approach. Then, in section 2.5 we focus on cross-modal referential alignment with particular emphasis on its temporal implications. Here, sequence analysis techniques, used to measure similarity between visual and linguistic referential information, are described. All the measures presented are used in inferential statistical analysis. In the last section 2.6, we review statistical methods commonly used to analyze eye-movements data showing advantages and disadvantages. On these premises, we motivate and describe the choice of a linear mixed effects modeling approach.

2.2 Eye-movements: a window on cognitive processes

With our eyes, we actively gather and process visual information from the surrounding world. During this activity, the eyes constantly move from one location to another of

the visual space. This motion can be used as an explicit indicator of how visual attention distributes in **space** over **time** while performing a task, e.g. reading (Rayner, 1984; Yarbus, 1967). Eye-tracking technology makes it possible to quantify visual attention by capturing two main components of the oculomotor signal: the **saccade**, usually measured as a distance[1], which is the movement covered by the eye when moving in the visual space, from one location to another; and the **fixation**, which is the time, between saccades, during which the eye remains relatively still[2] on a spatial location for about 200-300ms[3]. A fixation is usually directly interpreted as an indicator of ongoing cognitive processes. It provides information regarding the identity of the **spatial region** where attention is allocated, e.g. for reading studies, it can be an ambiguous word, and of the **processing load**, i.e. the time spent in the region (Rayner, 1998). A saccade, instead, is a more implicit index of visual attention[4], especially informative of spatial inspection. Different visual tasks, e.g. memorization or search, trigger different saccadic lengths, with search having longer initial saccade than memorization (Castelhano *et al.*, 2009). A search task requires, initially, a wider sampling of the scene to identify the likely locations where the cued target might be found, whereas during memorization, the goal is to remember as many objects as possible regardless of their position within the scene. In this thesis, we focus on fixation data. The main reason is that a fixation explicitly tells us about which visual referent is currently processed, when and for how long. This piece of information is especially crucial when eye-movements are interpreted in the context of other ongoing processes, e.g. sentence processing. In such case, the responses of the visual system are modulated by mechanisms of sentence processing; thus, the only way to interpret this interaction is by looking at how the relation between visual referents fixated and processed linguistic information changes over time.

[1]Expressed in degree of visual angle. The distance is correlated with time: the longer the distance, the more time saccading takes.

[2]The eye makes always micro-movements (e.g.,*nistagmus*).

[3]Fixation duration varies according to the task performed (Castelhano *et al.*, 2009; Rayner, 1984).

[4]An open debate is whether during saccading, there is partial suppression of cognitive processes, e.g. Van Duren & Sanders 1995.

2.3 Eye-movements in situated language processing

In reading (e.g. Clifton *et al.* 2007; Demberg & Keller 2008; Tinker 1958), or purely visual tasks (e.g. search Malcolm & Henderson 2009; Schmidt & Zelinsky 2009), eye-movements are observed with respect to a Region Of Interest (ROI), which can be a *word* of the sentence, or a *target object* embedded in the scene. At the ROI, eye-movements are expected to carry crucial information about the process investigated. For example, in reading, the fixation duration on words reflects processing complexity; or in search, the time spent inspecting the scene before finding the target object can inform on the complexity of the scene. Notice that, in reading or visual search, the experimental conditions manipulated in a trial are time independent. In reading studies of garden-path sentences, e.g. *the horse raced past the barn fell* (Ferreira *et al.*, 2001), the ambiguous ROI *past* can be read by different participants at different times; what is crucial is the eye-movement information at that particular ROI, e.g. first pass gaze duration (Sturt, 2002). Thus, the measure of fixation observed at the ROI is independent from the time-course of the trial.

More recently, however, the use of eye-movements has extended to more interactive tasks where language is situated in a visual context (**V**isual **W**orld **P**aradigm, e.g. Tanenhaus *et al.* 1995). In these studies, participants are asked to perform language processing tasks, such as understanding spoken stimuli (e.g. (Spivey-Knowlton *et al.*, 2002)), or producing sentences (e.g. (Griffin & Bock, 2000)) while concurrently viewing a visual context; this context, to some extent, corresponds to the linguistic information processed, e.g. a depicted APPLE matching the spoken phrase *the apple*. The presence of a visual context required new ways of analyzing eye-movement data. In fact, during synchronous processing of visual and linguistic information, eye-movements patterns on the visual context are conditioned by the linguistic stimuli processed. Thus, at different time-points during a trial, which usually correspond to the phrases of a sentence, the eye-movement patterns change in response to linguistic processing. Moreover, depending on the complexity of the visual context, each change in the linguistic stimuli can involve more than one visual ROI; the most evident case being when linguistic and visual information are referentially ambiguous[1], i.e. to one linguistic referent (*the apple*) might correspond several visual objects (APPLES).

[1] Referential ambiguity will be more deeply discussed in chapter 4.

A solution to treat time in its continuity is achieved by **time-locking** the visual responses to the linguistic stimuli. Time-locking means aligning the eye-movements data to the linguistic window of interest. In a typical situated language comprehension experiment the phrases of a sentence are analyzed with respect to the visual regions of the context. Once the linguistic and visual ROIs have been decided, fixations are first aligned[1] to the onset of linguistic ROI and then aggregated by visual ROI. After aligning fixations to the onset of the linguistic ROI, we have to decide on the size of the temporal window to be considered, and its position with respect to the onset of the linguistic ROI, i.e. before or after. This decision depends on whether the task is language comprehension or production, and on the issues concerning the research question. In general, for language comprehension the effects are usually expected to emerge after the linguistic ROI is mentioned, thus the window considered starts at the onset of the critical region and ends after it. For production instead, effects can be observed while the referent is mentioned, thus the window considered starts before the referent is mentioned (Qu & Chai, 2008), and ends shortly after it (Coco & Keller, 2010b). Often, however, the decision about which window of analysis has to be considered strictly depends on the theoretical expectations based upon the experimental design. For example, in studies investigating the phenomena of visual anticipation during sentence comprehension, looks on the visual ROI are expected to increase before the corresponding linguistic ROI is mentioned: the verb *eat* triggers anticipatory looks to edible objects (CAKE) before the referent *the cake* is actually mentioned (Altmann & Kamide, 1999). Once the window of analysis is set, we have to decide its size. Sentences usually vary both within and across different experiments. For example, the time elapsing between phrases of the same type of sentence might be considerably different across sentences. Thus, in order to avoid a misleading interpretation of the data, it is important to correctly center the window around the linguistic ROI and choose a size that does not overlap with neighboring regions. Each window can be further sliced into smaller units, e.g. 10ms, where the presence or absence of a fixation is represented by a binary response[2] (0,1). On each slice, then, proportions[3] are

[1] In an experiment there are several trials, each with different onsets for the linguistic regions; alignment is done considering this information.

[2] Larger slices, however, can contain more than a single fixation. A 50ms slice might contain 2 fixations.

[3] Other measures can be computed, e.g. log-ratio (Arai *et al.*, 2007) or empirical logit (Barr, 2008),

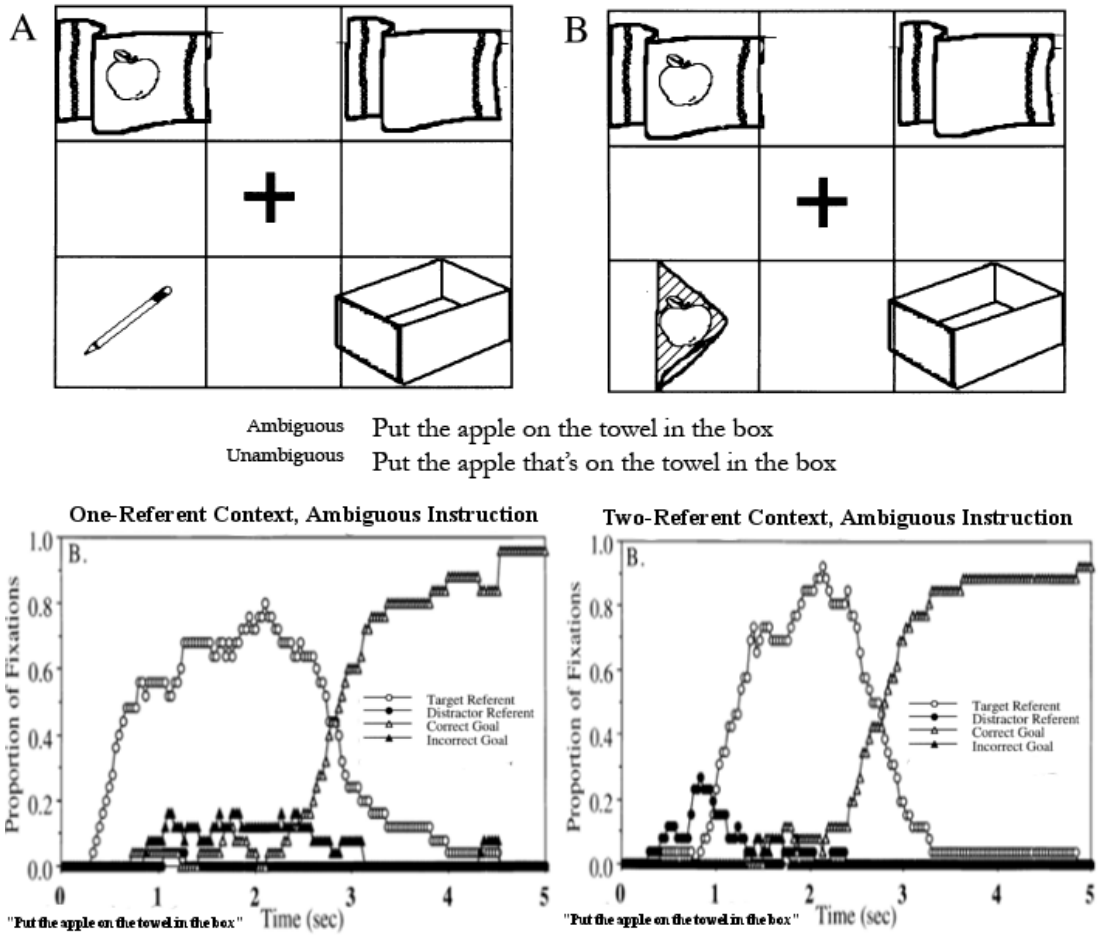


Figure 2.1: Example of a VWP experimental trial and proportion of fixations plot taken from Spivey-Knowlton *et al.* 2002. Visual ROI: *Target Referent* (APPLE ON TOWEL); *Distractor Referent* (PENCIL — APPLE ON NAPKIN) *Correct Goal* (EMPTY BOWL); *Incorrect Goal* (EMPTY TOWEL).

computed, across the different conditions, aggregating by participants and trials (more details about statistical analyses of eye-movements can be found in section 2.6).

In order to exemplify how the analysis of eye-movements during situated language processing is done, we are going to briefly walk through an experiment by Spivey-Knowlton *et al.* 2002.

In this study, the authors asked how prepositional phrase (PP) attachment ambiguity of sentences such as *Put the apple on the towel in the box*, is resolved in different

depending on the specific research hypotheses, and the statistical analysis performed.

visual contexts: one-referent context which supports an ambiguous goal-location reading of the prepositional phrase *on the towel* (A: a single apple on the towel), and the two-referent context that does not (B: two apples depicted); see Figure 2.1 for an example trial. Their hypothesis is that in one-referent context, *on the towel* is ambiguously interpreted as goal location: i.e. the APPLE ON TOWEL has to be moved in the EMPTY TOWEL; whereas in a two-referent context, the referential ambiguity of APPLE (single and on a towel) resolve the syntactic ambiguity by making *on the towel* interpreted as noun-modification *the apple that's on the towel*.

This hypothesis is tested by looking at proportion of fixations over time across the different visual ROIs compared on the same condition[1], e.g. unambiguous vs ambiguous. During the first PP *on the towel*, proportion of fixations are mainly expected to change on two visual ROIs: the EMPTY TOWEL (*Incorrect Goal*), and APPLE ON TOWEL (*Target Referent*); whereas at the second PP *in the bowl*, effects are expected only on the BOWL (*Correct Goal*). Their prediction is that there should more looks on the EMPTY TOWEL during linguistic ROI *on the towel* for one-referent context (A: APPLE ON TOWEL) compared to two-referent context (B: SINGLE APPLE, APPLE ON TOWEL). For one-referent context, if the first PP *on the towel* is interpreted only as noun-modification, we would expect looks only to APPLE ON TOWEL; instead if it is also misinterpreted as goal-location, we would expect looks also on EMPTY TOWEL. In Figure 2.1, we show proportion of fixations for one-referent and two-referent context across the different visual ROI when sentence is ambiguous. The fixations are aligned at the onset of direct object *the apple* and proportions calculated across the four different visual ROIs in slices of 33 ms, over a total time course of 5 sec. By comparing the two plots, we can observe that in one-referent context, there are more looks, i.e. higher proportion of fixation, on EMPTY TOWEL (*Incorrect Goal*) during the first PP *on the towel* compared to two-referent context. This implies that in one-referent context, participants are interpreting the first PP *on the towel* as goal location for direct object *the apple*; whereas in two-referent context, the referential ambiguity resolves syntactic ambiguity through visual competition. The statistical significance of these results is assessed using ANOVA on proportion of fixations aggregated over trials and time. In

[1] In our work, we compare proportion of fixations on the same object across conditions. In this way, the effect of conditions emerge more clearly on individual objects.

section 2.6, we discuss the shortcomings of this approach while motivating our use of linear mixed effect modeling.

By analyzing proportion of looks aggregated in visual objects, the sequentiality of individual eye-movement record is lost, thus failing to utilize the information about temporal relations of looks across the different objects: i.e. looking at the APPLE ON TOWEL just after having looked at SINGLE APPLE. In experimental designs which make use of object arrays, and therefore the total number of objects is small and balanced across trials, this is not problematic. However, when sentence processing is investigated within naturalistic scenes (Coco & Keller, 2009), and the number of objects, together with their visual properties, varies considerably across scenes, focusing on individual objects detached from their sequential context might be oversimplistic.

2.4 Situated language processing in naturalistic scenes

Psycholinguistic research based on the VWP has mainly used very simple visual contexts, such as objects arrays or clip-art pseudo-scenes where the number of objects is controlled, and the visual complexity, both in terms of low (i.e. color, intensity etc.) and high (i.e. spatial layout, contextual information etc.) level features, is minimal.

Figure 2.2: Comparison between photo-realistic scenes used by Coco & Keller (2010b) and standard VWP visual material used by Knoeferle & Crocker (2006).

In Figure 2.2 we compare a standard VWP visual context with a photo-realistic scene. The first noticeable difference is the number of objects. In a pseudo-scene the number of objects is much smaller than in a photo-realistic scene. This referential

simplicity makes the match between linguistic and visual referents easier in a pseudo-scene, where very few objects can be named or referred to. This also reduces significantly the noise in the eye-movements data, in that visual attention will be artificially drawn by the few objects depicted compared to a photo-realistic scene.

The referential simplicity of pseudo-scenes often implies the absence of ambiguity[1] in that each visual object can be uniquely referred to. In a photo-realistic scene, instead, many objects could share the same linguistic referent. Moreover, when dealing with photo-realistic scenes, there is the clear advantage of having more natural visual responses; hence, we can also observe the impact of image-based visual factors, e.g. clutter (Rosenholtz *et al.*, 2007), on sentence processing mechanisms (see Chapter 4). Some of the differences between pseudo and photo-realistic scenes can be accounted for using methods already applied in VWP studies. The issue of multiple referentiality, for example, can be solved by simply comparing the log-ratio of fixations between ambiguous objects: where $\log(\frac{p(O_1)}{p(O_2)}$, and $p(O_1$ is the proportion of fixation on an object compared to fixating another $p(O_2$; see Arai *et al.* 2007 for an application. However, the introduction of photo-realistic scenes imposes a more radical re-interpretation of eye-movements. The fixations on objects are now dependent on contextual relationship, which are implicit in the scene layout: a FORK and a PLATE are both spatially close and semantically connected, thus also the order of fixations on them is expected to be temporally related (Hwang *et al.*, 2009). In contrast to pseudo-scenes, where the objects do not have contextual relationship, both in terms of visual layout, i.e. objects float, and semantics, i.e. a BALLERINA is expected in a THEATER dancing rather than pouring water on a CELLIST. Therefore, there is no contextually implicit ordering, determining the way visual attention is displayed. When monitoring visual attention during situated language processing in a photo-realistic scene, instead, it becomes crucial to keep track of sequential order when analyzing fixations.

[1] Unless ambiguity is experimentally manipulated.

2.5 Sequentiality during referential information processing

When we observe a scene, our eyes move sequentially from one object to another. This sequence of fixations forms a **scan pattern** (Noton & Stark, 1971). We interpret a scan pattern as an explicit representation of the referential information visually attended to during a task by a viewer. However, during situated language processing, a scan pattern is paired with a corresponding sentence, which is a sequence of words. Thus, the objects visually attended have a sequential relation with the words uttered (in production) or listened to (in comprehension).

In Chapter 4, we report a situated language production eye-tracking experiment, where participants are asked to describe photo-realistic scenes. The sentences generated are paired to the scan pattern that followed (see Figure 2.3 for an example of scene and data obtained, sentence and scan pattern). A sentence is already in the form of a sequence, e.g. *the man is signing in*; however, in order to have also eye-movement in sequential form we need to map fixations generated during the course of the trial into a scan pattern. To do that, we use the LabelMe Matlab toolbox (Russell *et al.*, 2008), which allows us to annotate the scene with polygons, drawn around the edges of a recognized object. In the scene of Figure 2.3, MAN or COUNTER are examples of LabelMe annotations. The information of polygons is saved in XML format, and beside the name of the label given by the annotator, it contains all coordinates (x,y) for the points of the vertexes forming the polygon. In order to map a fixation to the label of polygon visually attended, we calculate whether the fixation's coordinates fall within the area covered by the polygon. In case of embedded polygons, i.e. a HEAD is part of the BODY, we assign to the fixation the names of all embedded polygons ordered by their area calculated in pixel square, from smallest to largest, e.g. HEAD $<$ BODY $<$ MAN. A final scan pattern is then a sequence of fixated objects represented as labels.

Once sentences and scan patterns are in the form of sequences, we can investigate their synchronous relation as a problem of referential information alignment. However, before being able to create a model of cross-modal alignment[1], it is crucial to explore the conditions influencing this alignment.

[1] A goal which goes beyond the purposes of this thesis.

Figure 2.3: An example of scene used in experiment 6 (Chapter 5) annotated with polygons. A visualization of scan-pattern information for two different participants.

A first question that needs to be addressed is whether similar sentences correlate with similar scan patterns, and if so, which linguistic and visual factors are involved in this coordination. To answer this question, in Chapter 5 we look at the pairwise

similarity between these two types of sequences, and derive a measure of cross-modal coordination which can be used to predict the strength of association of a pair sentence/scan pattern.

The problem of computing similarity between sentences and scan patterns can be treated as a problem of sequence analysis. Finding similarity between sequences is a well known problem in the field of bio-informatics (Durbin *et al.*, 2003), where genetic codes have to be compared to unravel underlying similarities. A guiding principle used to capture these similarities is **alignment**. The more elements two sequences share, the more similar they are. Alignment presents, however, two major issues: sequences differ in length, and the elements composing the sequences, even if identical, can be positioned differently. We implement three measures[1], Needleman-Wunsch (NW, Durbin *et al.* 2003), Longest Common Subsequence (LCS, Gusfield 1997) and Ordered Sequence Similarity (OSS, Gomez & Valls 2009), all of them able to solve these issues.

NW is a classic and simple method of sequence alignment, which has recently been applied to eye-movements[2] data in a study conducted by Cristino *et al.* 2010. NW is an iterative dynamic programming algorithm which performs a global alignment between two sequences using a substitution matrix and a gap-penalty term. A substitution matrix returns a similarity score between two aligned data-points (i.e. aligning MAN-L with MAN-R may return a score of 0.8, whereas aligning MAN-L with CLIPBOARD returns a score of 0.1; and aligning MAN-L with MAN-L may return the maximum similarity of 1), and the gap-term, e.g. -1, penalizes this score every time a gap has to be introduced in order to allow a matching between two sequences. The algorithm scores and saves values of local alignment of sub-sequences using the substitution matrix and the gap penalty. Then, the best alignment is found by backtracking the optimal alignment path within the matrix. The similarity between the two sequences corresponds to the score obtained by this optimal path: the higher the score, the more the similarity.

A simpler method of sequence alignment similar to NW[3] is LCS. In LCS, the goal

[1] For application, see Chapter 5.

[2] Implemented as Matlab toolbox (ScanMatch): http://eis.bris.ac.uk/ psidg/ScanMatch/index.html.

[3] In Chapter 5 we show that LCS gives highly correlated result with NW, when the substitution matrix has 1 along the diagonal, indicating a perfect match, 0 otherwise; and the gap-penalty is 0, i.e. no penalty.

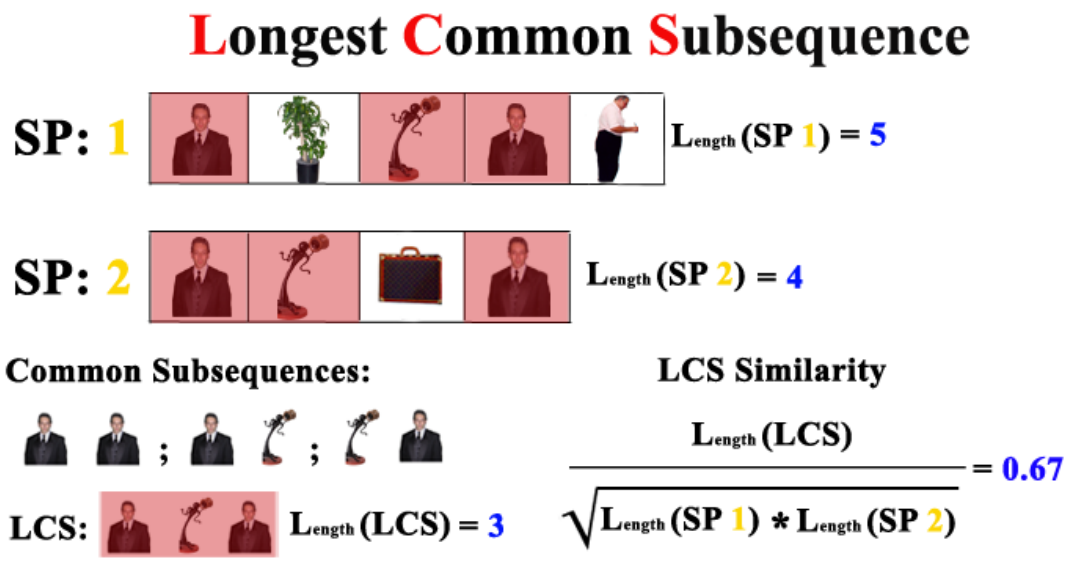


Figure 2.4: Longest Common Subsequence is a measure of similarity based on ordered subsequences. Between two sequences, it explores the space of all common subsequences seeking for the longest. SP-1 and SP-2 share several common subsequences of length 2 (e.g. man-man). In this example, the LCS is of length 3.

is to find the longest subsequence common to two, or more, sequences. Conceptually, the algorithm searches the space of all combinations of ordered subsequences, looking for the alignment which maximizes the number of common elements. The algorithm follows a dynamic programming approach, where the final solution (the longest alignment) is iteratively built up, based on solutions of subproblems (looking for all common subsequences). Once we find the longest subsequence, we calculate the similarity score as the ratio between the length of LCS and the geometric mean of the two sequences. For example in Figure 2.4, SP-1 and SP-2 share several common ordered subsequences, e.g. man-man or man-statue, with a length of 2. The algorithm is designed to explore all possible combinations trying to find the common subsequence with the longest length. In this example, the longest ordered common subsequence is *man-statue-man* with a length of 3. Often, LCS finds more than an unique solution, i.e. two sequences can have two LCS of the same length. However, even if more than one LCS is found, they will have the same similarity score.

The second method used to compute sequence analysis is Ordered Sequence Similarity, shown to be more effective than established measures such as edit distance

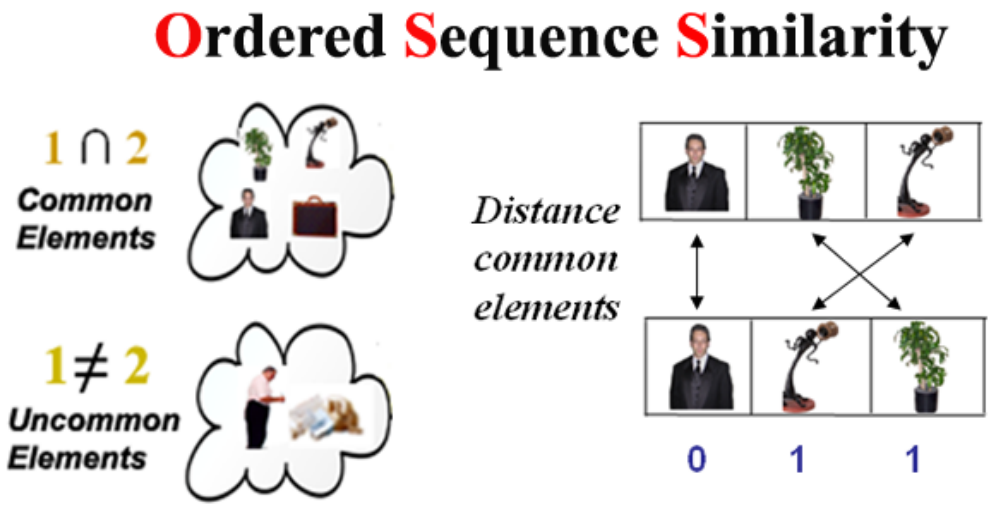


Figure 2.5: Ordered Sequence Similarity is a dissimilarity measure which integrates the information about which elements are common or uncommon between 2 sequences while taking into account the relative distance between those elements that are common.

(Gomez & Valls, 2009). OSS is based on two aspects of sequential data: the elements the sequence is composed of, and their positions. When comparing two sequences, it divides the elements into common (shared) and uncommon (unique); and on the shared elements, it takes into account the relative position. The first step is to separate target objects that are common in both scan patterns, from those that are unique. For example in Figure 2.5, four objects are shared by the two scan patterns (man-R, plant, statue, suitcase); whereas two objects (telephone, man-L) are unique respectively in SP-1 and SP-2. For each common element, we calculate the distance between the two sequences, e.g., statue of scan pattern 1 is two units distant from statue in scan pattern 2. Distances between common elements, and number of uncommon elements are integrated into a unique metric, which is normalized on the basis of sequence lengths (for details refer to Gomez & Valls 2009). Despite its name, OSS gives a dissimilarity measure, which we convert into similarity, to allow easier visualization, by simply subtracting distances from 1. We use sequence analysis in Chapter 5 to quantify patterns of similarities between cross-modal (visual and linguistic) referential information. In Chapter 6, sequence analysis is used to compare scan pattern similarity within and between different tasks, which allows us to unravel shared mechanisms of referential

information processing underlying the different tasks performed.

2.6 Inferential Analysis

In the previous sections, we have discussed our methodology of investigation, contextualized within the literature, along with descriptions of the measures that we will be using to explore our hypotheses. However, in order to test the statistical validity of our experiments, we need to have a general framework for our inferential analysis. Statistical inference is a crucial step to extract significance from the data observed while refining, in the light of its application, the explanations of the hypothesis investigated. Therefore, it is extremely important to select a method which correctly adapts to the data observed. First, we review statistical methods commonly used to analyze eye-movements data showing advantages and disadvantages; then we motivate and describe the use of the linear mixed effect modeling approach.

2.6.1 Traditional Vs Modern methods of statistical inference

Eye-movements data used to quantify linguistically situated visual attention are specified both spatially and temporally. Fixations are, in fact, spatially bounded to visual objects and temporally distributed over a precise time-course. Moreover, fixations are repeatedly sampled (longitudinal data) at a high resolution, over the same subject during the same trial in a hierarchically nested design. A comprehensive analysis of eye-movements needs, therefore, to take into account the spatial and temporal components of the data while accounting for the way in which the data is sampled.

A standard method used to analyze visual-world data is ANOVA. ANOVA tells us whether the mean of a certain response variable, e.g., proportion of fixation, is significantly different between different explanatory variables, e.g. number of visual *Referents* (**One** Vs **Two**) corresponding to a linguistic referent *the apple*, by looking at their variance. So, if the proportion of fixations to a certain target object changes by including more than a visual referent (e.g. **Two** referents lead to a smaller proportion of fixations on target object than **One**): ANOVA compares the means of different experimental conditions and determines whether to reject the hypothesis that the conditions have the same population means given the observed sample variances within

and between the conditions. ANOVA compares the means of different experimental conditions, e.g. One or Two referents, and determines whether it is or not statistically different by looking at their variance[1].

The first problem of using ANOVA in this example stems from the fact that we are dealing with proportions calculated over categorical variables. On a continuous variable we can have a clear interpretation of mean, variance and confidence intervals, whereas on a categorical variable, the confidence intervals can extend beyond the interpretable values of 0 and 1 (refer to Jaeger (2008); Richter (2006) for a more detailed discussion). The application of ANOVA might, therefore lead to spurious results.

A second problem encountered by using ANOVA is the assumption of independence between observations. Namely, the proportion of fixations at time t is assumed to be independent from that at time $t+1$. Eye-movements are sampled over time at a very high resolution (e.g. every 10ms), thus observations at time $t+1$ are certainly dependent on those observed at time t (Barr, 2008). An ad-hoc solution to overcome this limitation was to calculate proportions on fixations aggregated in large temporal windows (e.g. 200ms), and then run separate ANOVA over the different windows (Kamide *et al.*, 2003; Novick *et al.*, 2008). This solution is redundant, i.e. as many ANOVA as there are windows; incomplete, i.e. we know the variance of explanatory variables within a specific window but we cannot estimate how it changes over time[2], and more importantly prone to Type II errors, i.e. by multiple testing we might accept the null hypothesis when it is in fact false.

The third problem concerns the presence of random variance due to a nested experimental design. Observations are sampled on different trials and subjects, which are both nested within the explanatory variables (e.g. **One** or **Two** *Referent)*.
The solution proposed to deal with the random variance was to calculate proportions aggregated by the design random effects, i.e. subjects and trials, and then get ANOVA F-scores for both (Clark, 1973). This method discriminates the random variance by independent random groups, either subject or trial, but without handling both simultaneously. Also, by computing the proportions twice, aggregated by subjects and trials, we redundantly duplicate results over the same dataset.

[1] ANOVA partitions the total sum of square information into components related to the effects used in the model. Then, an F-test is used to assess the total deviation among these components.

[2] ANOVA is thus temporally underspecified.

2.6.1.1 Linear Mixed Effect Regression Models

For the reasons presented above, extensively discussed in the *Journal of Memory and Language special issue on Emerging Data Analysis* edited by Forster & Masson 2008, we perform our inferential analysis under the **G**eneralized **L**inear **M**ixed Effects Regression **M**odels (GLMM) framework, using the **L**inear **M**ixed **E**ffect (LME) class of models (Baayen *et al.*, 2008; Pinheiro & Bates, 2000).
The simplest model of regression is a *linear model* which assumes a linear relationship between the observed response variable, e.g. proportion of fixations, and the explanatory variables, e.g. *One*/*Two* referents. The explanatory variables are expressed in terms of regression coefficients, β, which inform on the nature and strength of the linear relationship with the response variable y.
A linear model is then a selection of coefficients β_i, one for each explanatory variable i (and one for each of their interactions): $y = \beta_0 + \beta_1 x_1 + ... + \beta_i x_i$. The coefficient β_i expresses the contribution of the i_{th} variable to the probability of the outcome event, that is, in our case, proportion of fixation (Agresti, 2007). So, the explanatory variable *Two* will have a negative coefficient if it contributes negatively to the proportion of fixation on our target object. Moreover, the strength of this negative relation will be expressed by the size of the coefficient: e.g. small coefficient, small contribution.

A GLM is a generalized case of linear model, where the linear relationship between the response variable and the explanatory variables can be established for a variety of distributions through the use of link functions (e.g. logistic, Poisson, etc.). In eye-movement data, the measure of fixations is a binary response variable (e.g. presence/absence of fixation on target object), not normally distributed, which we can transform into a probability distribution through the *logit link* function. The logit function is created by taking the logarithm of the *odds* of the response variable: $logit(y) = \log(\frac{y}{1-y})$. The odds is the ratio of the probability that the event of interest occurs, y, to the probability that it does not, $1-y$[1]. In our case, odds is the ratio of the number of times that an object (e.g. BOWL) has been fixated to the number of times that it has not been. Then, the logit is obtained by scaling the odds logarithmically. The logit transformation allows binary responses to be normally distributed.

[1] If we chose a random day for a week, the odds that it would be a Sunday is 1/6. Instead, the probability that by choosing a random day it will be a Sunday is 1/7.

The limit of GLM is that it can't distinguish between effects due to explanatory variables *fixed*, and those, instead, related to sampling (*random*): both in terms of variables (e.g. subjects) or method (e.g. longitudinal data). LME can overcome this limitation by explicitly discriminating in the formula between fixed β_i and random b_i effects: $y = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + b_1 z_1 + ... + b_q z_q + \varepsilon$; where p and z denote fixed and random regressors respectively, and ε is the error term. In summary, we adopt LME because it allows us to quantify the relative contribution of each explanatory variable to the outcome of our binary response variable through logistic regression, while discriminating the relative impact between the different multilevel components, fixed and random, of the model.

2.6.1.2 Model Selection

A common problem encountered when doing model based inference is the process of selection.
Complex experimental designs with many explanatory variables can lead to different models, all equally good in explaining and fitting the data observed.
At a superficial glance, the optimal model should set the explanatory variables in a way that is consistent with the hypothesis under investigation. However, by building such an ad-hoc model we would deductively test an unique hypothesis wrongly assuming that no other explanations of the phenomena under observation are possible. In a more inductive approach to model selection, we can instead assume that multiple hypotheses (Anderson, 2008) are linked to the explanatory variables, and through a bottom-up exploration of the data we bootstrap the best hypothesis/model. The assumption is that the experimental design is deductively built to contain a certain universe of hypotheses, expressed in terms of explanatory variables. Then, through an inductive exploration of the data, i.e., model selection, we decide the hypothesis that best models the data observed.
There are two main approaches to perform model selection: *backward* (Crawley, 2007; Whittingham *et al.*, 2006) and *forward* (Baayen, 2008; Burnham & Anderson, 2002). The terms refer to the direction of inclusion/exclusion of explanatory variables.
In backward selection the explanatory variables, main effects and interactions, are all initially included in a fully specified model. Then, the model is reduced by iteratively

excluding non-significant variables.
A forward selection operates in the other direction. It starts with an empty model, and then variables are included, one at time, if statistically significant.
The decision on which variables have to be excluded or included is made on the basis of model fit.

A statistical method which allows us to compare the fit between two models is the ***Log-L****ikelihood* test. The likelihood tells us how well a certain model is fitting the data; the natural logarithm of the likelihood is taken because it is computationally more convenient. We utilize the Log-likelihood test to test nested models. Given a specific model, we calculate the likelihood of observing the actual data. The log-likelihood of this model is compared to the log-likelihood of a nested model (i.e. one in which fewer parameters are allowed to vary independently). The model with the best log-likelihood is retained. Comparison between models is performed pairwise. Thus, at every iteration two models are compared: a new model containing one more or less parameter (depending on whether the selection is backward or forward) and an old model coming from the precedent iteration; we compare their Log-Likelihood, and we keep the model that has the higher Log-Likelihood (better fit).

When dealing with LME another aspect that further complicates the process of model selection is the distinction between fixed and random effects. For example, a fixed effect can have a stronger impact on the fit than another: how do we decide the order of inclusion of fixed effects? It can also have quite different intercepts for the different groups of our random effect: how do we treat the relation between fixed and random effect?
We implement a step-wise forward selection algorithm that iteratively finds the best model containing both fixed and random effects. We start with an empty model, then we add the random effects, e.g. $(1|subject) + (1|item) + (1|...)$[1] until no further improvement is possible. Supposing that after iterating over the random effects we have a model where $(1|subject)$ improved the fit, whereas a model containing also *item*, $(1|subject) + (1|item)$, didn't; we keep the first, simpler model.
With a model containing random effects, we pass on to add fixed effects. Fixed and random factors are included ordered by their log-likelihood improvement. Every time we include a new fixed variable $(1|subject) + referent$ we calculate whether the inclusion

[1] The example uses R lme4 pseudo-code.

of a random slope improves model fit $(1|subject) + referent + (0 + referent|subject)$. Notice that we assume the random slope on *referent* to be independent $(0|...)$ from another random slope on a different fixed effect $(0 + fixeff|subject)$. In our experimental design, the different explanatory variables are unrelated, thus we do not expect them to have similar randomness $(1 + referent + fixeff|subject)$ on the different groups of the random effect. We iterate over fixed effects and related random slopes until no further improvement is possible.

The next step is to include the interactions. We generate the interactions which do not violate the subset criterion by considering only the fixed effects present in the final model selected. At this point, we do not calculate random slopes on the interactions as it makes computation intractable. We include interactions up to three ways. The final model obtained is guaranteed to have only those fixed and random effects that significantly improve model fit. Moreover, the forward selection guarantees parsimony on the number of parameters included. This algorithm of model selection has been used to generate the inferential results reported along the different chapters.

2.6.1.3 Comparison with alternative analyses

Data can be analyzed in multiple ways, each with advantages and disadvantages. Traditionally, fixation data is represented in terms of probability of fixation[1], which being 'raw', is assumed to be a more credible quantification of the data; see section 2.3 for a discussion in the context of Spivey-Knowlton *et al.* (2002). However, several transformations could be applied to the data to fit the requirements of modeling strategies, as shown in the section above. These transformations may, or may not disrupt what is actually seen in the raw data. In this thesis, we plot and use empirical logit of fixation, rather than proportion; and we count a fixation when the eye lands, and remains still, on the object. Thus, we don't include the time saccading between objects as fixation for the landing object. Planning a saccade towards an object implies the intention of fixating, which can prove useful to find anticipatory effects. So, in order to justify our definition of fixation, and assess the validity of our empirical logit transformation, in the next part of the section I will compare methodologies on a subset of the data pre-

[1] Equivalently proportion.

sented in 3. Then, I will walk the reader through a simple made-up experiment on how linear mixed effect results are interpreted.

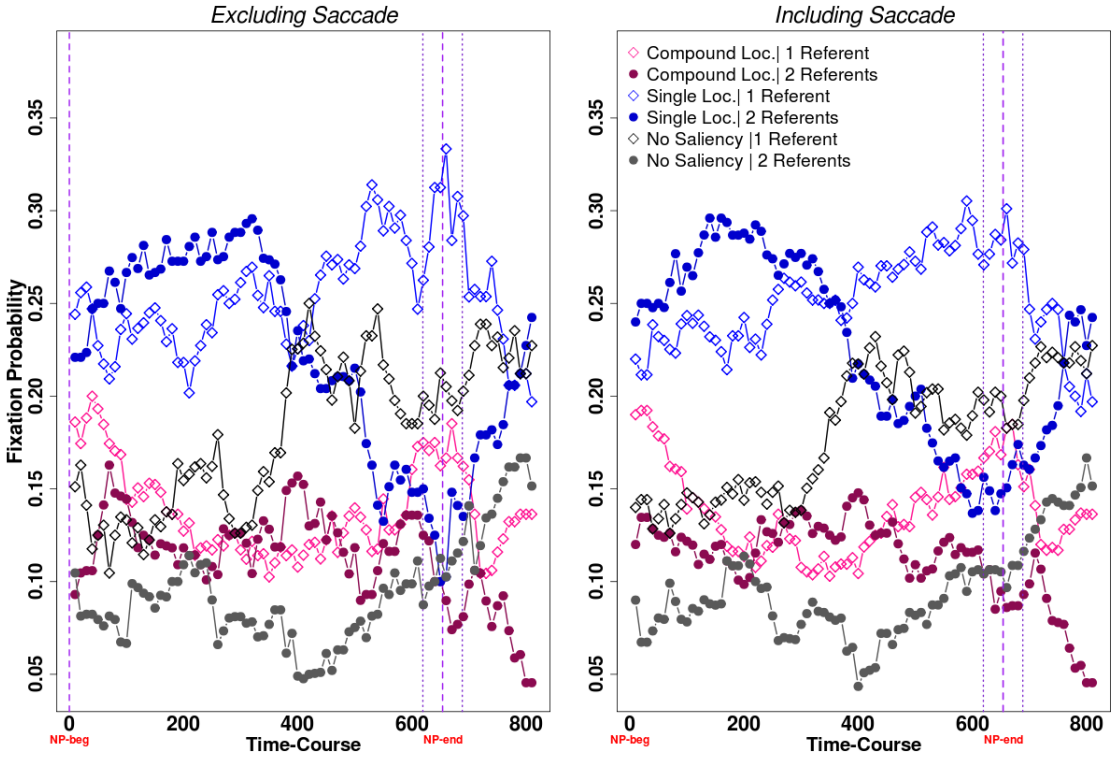
Including saccades and empirical logit comparison In Figure 2.6, we plot proportions of fixation on the object BOWL during mention of the direct object *the orange*. Fixations are aligned at the beginning of the critical word [1] for a window of 800 ms in 80 slices, 10 ms each. We compare two ways of counting fixation duration by *Including* or *Excluding* the time spent during the saccade. In the Including case, the saccade is added on the fixation duration at the landing object, i.e, the receiver object. In practice, we anticipate and extend the fixation duration on the object. In the Excluding case instead a fixation is counted from landing. Comparing the two plots in Figure 2.6, it is evident that there is almost complete equivalence between the two way of counting. When a saccade is included, fixation proportion appears smoother then when it is not. However, the two trends display a very similar distribution.

Turning onto the difference between probability and empirical logit, we compare the equivalence of trends when either proportion of fixation, or empirical logit is the dependent measure used. Empirical logit is a variation of logit, which includes a constant 0.5 on both numerator and denominator to avoid undefined logarithms: $emplog(y) = \log(\frac{y+0.5}{N-y+0.5})$; where y are fixations, either proportions or frequencies, and N is the normalization term. For proportions, y is the probability of gazing at the target object (e.g. BOWL) across conditions at each time-point. N can be either 1 to scale it between $\approx -1 and \approx 1$, or a normalization constant of the design. Both methods give equivalent trends for different ranges of the dependent measure. We set a normalization constant at 6, which is the number of possible objects[2], assuming that objects compete for fixations, i.e., we can look only one object at each time-point. For frequencies, y is the number of fixations on the object along the time-course, over each time-frame, aggregated by subjects and trials per condition; and N is the total number of time-frames across conditions.

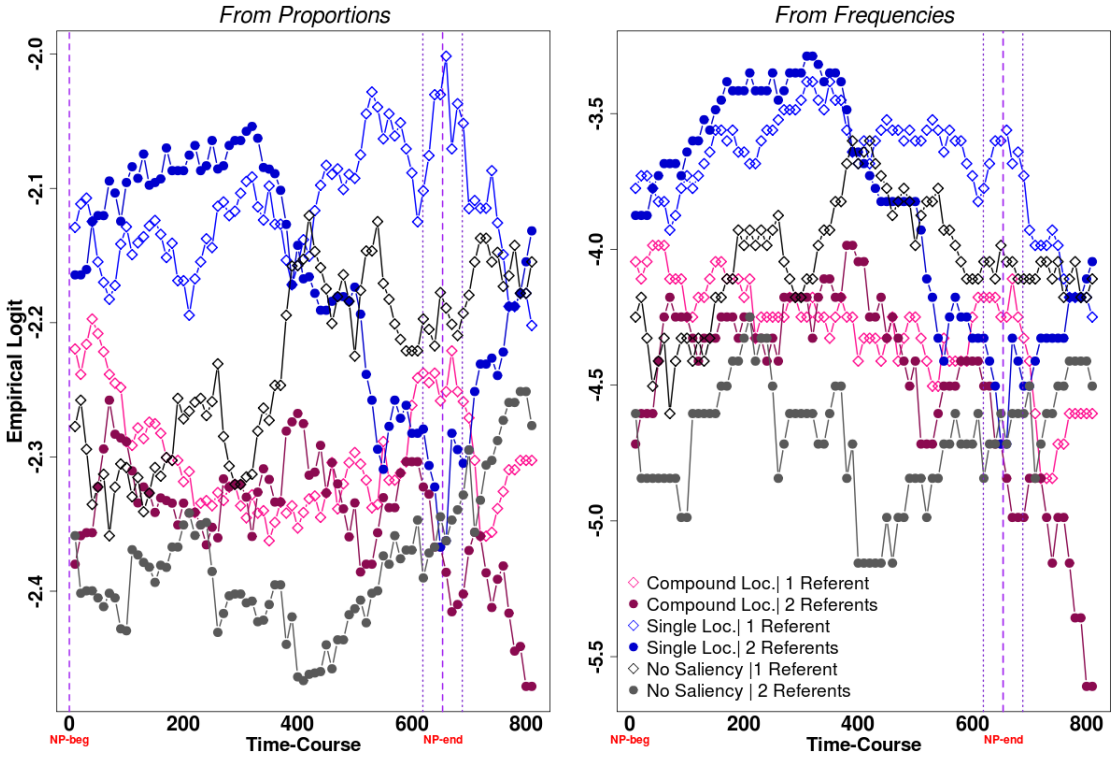
In Figure 2.6, we plot the two types of empirical logit transformation over 800 ms from the onset of NP direct object *the orange*, and compare it with the above plot showing it in proportions. We can immediately observe that, despite the difference in

[1] We mark its mean offset in the plot.

[2] Including the background object.



(a) Including vs Excluding saccades.



(b) Empirical Logit transformations: Proportions vs Frequencies

Figure 2.6: Comparison of methodologies in the time-course analysis of data discussed in section 3.5.2.1 of Chapter 3

the range of the dependent measure, the trend is perfectly equivalent across the different transformations. When comparing empirical logits calculated from proportions and frequencies, the only noticeable different is in the range of the measure, but not between conditions.

Since empirical-logit is more suitable to linear mixed effect modeling (Barr, 2008), we will mainly use this measure for plots and models. For linear mixed effects modeling, we calculate empirical logit such that we avoid data aggregation. The reason is that if we aggregate, we loose our random variables (e.g., participants), which are a necessary component of multi-level modeling. Thus, at the observation level, i.e., for each time-frame, our y is whether a fixation occurred or not (0,1), and N is the length of the time-course, e.g., 80 frames, 10 *ms* each. Throughout the thesis, we mainly show time-course plots of empirical logits calculated from frequencies, as the range observed more faithfully matches the one estimated by the linear-mixed effect models[1].

Interpreting LME coefficients In an eye-tracking experiment, we test the influence of low-level visual features, i.e saliency, during situated language understanding. Participants are asked to listen to syntactically ambiguous sentences, e.g. *the girl will put the orange on the tray in the bowl* while concurrently viewing a visual context containing 4 objects: WOMAN, ORANGE, ORANGE ON TRAY and BOWL; see Figure 2.7 for an example trial. The explanatory variable manipulated is *Saliency* with 2 factors *No-Saliency* (saliency has not been manipulated), and *Single* (saliency is manipulated on the single orange). Our hypothesis is that at the beginning of direct object *the orange*, we expect more looks to single ORANGE when saliency on it is manipulated.

We consider eye-movements on SINGLE ORANGE aligned at the onset of *the orange* for 800ms. The fixed effects of our LME model are the factor variable *Saliency*, with 2 levels (No Saliency/Single) and *Time*, as continuous variable (8 windows 100ms each). The predictors are centered around the mean to avoid collinearity. In a balanced design, this means that our 2 levels of factor variable *Saliency* will take the values -0.5 for No-Saliency and 0.5 for Single. Our random effects are Subjects and Trials, 24 groups each.

[1] By using frequency of inspection, rather than proportions, we maintain the normalization range closer to the observation-level.

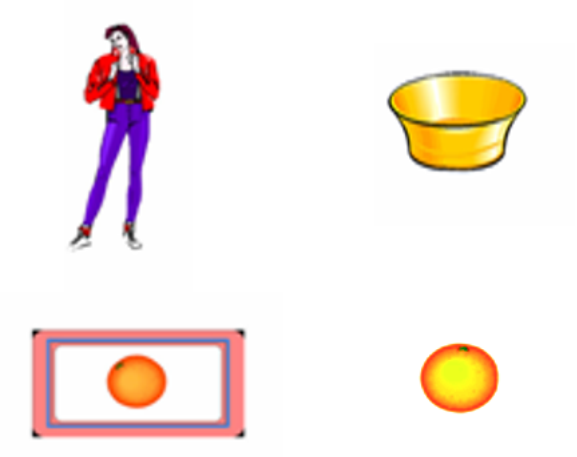


Figure 2.7: Example of made-up image trial based on experiments presented in Chapter 3.

Table 2.1: Example of mixed effects models table of coefficients: *Saliency*: No-Saliency (0.5), Single (-0.5)

	ROI: *the orange*	
Predictor	Coefficient	p
Intercept	-3.4967	0.0002
Saliency	-0.3333	0.0001
Saliency:Time	-0.2252	0.01

In Table 2.1 we show an example of LME coefficients table. In order to calculate the regression coefficient relative to the level for the categorical variable considered, we need to multiply the estimate returned by the model of the explanatory variable *Saliency*, $\beta = -0.3333$ for the level we are interested *Single*, -0.5. We obtain that *Single* has a regression coefficient of: $\beta_{Single} = 0.1666; p < 0.05$; which means that there are more looks on ORANGE when saliency is manipulated, compared to when it is not manipulated: $\beta_{No-Saliency} = -0.1666; p < 0.05$. The interpretation we give is that saliency has triggered more looks to ORANGE in prediction for upcoming post-verbal argument of the sentence. Moreover, we find that *Single* has a positive interaction with *Time* ($\beta_{Single:Time} = 0.1126; p < 0.05$). Looks increase over time. Participants prefer the single ORANGE as direct object *the orange*, compared to ORANGE ON TRAY[1]. The results for random effects are not reported along the thesis. However, for completeness of explanation, linear mixed models returns for each group of a random effect, i.e. *Subject* has 24 groups (the number of participants), the corresponding

[1]If we want to test directly this comparison, we would need to include as explanatory variable *Object* with 2 levels (*orange* and *orange on tray*).

random intercept. We can also have random slopes, which allow each group of the random effect to have a different slope. Then, we can observe the different effect of an explanatory variable, e.g.*Saliency* for each group of the random effect.

A final remark on interpreting linear mixed effect models is the linearity assumption, i.e., each predictor has a unique coefficient in the model indicating its difference from the intercept. Fixations develop over a time-course, and their trend changes as time unfolds. This might result in the measure being non-linear over time. A workaround to this problem is to adopt a polynomial definition of time, where the higher its degree is, the more terms we have to fit non-linearities in the time-course (see Mirman *et al.* (2008) for an application). However, a major drawback of this approach is a loss on the interpretation. In fact, by having a polynomial of 3 or more degrees, the interpretation of underlying cognitive processes becomes harder, and prone to idiosyncrasy. In this thesis, we discuss only linear effects and we don't add polynomial terms to our time variable. For results where the trend of the dependent measure is highly non-linear, we will remind the reader about our linearity assumption.

Chapter 3

The Interaction of Visual Saliency and Intonational Breaks during Syntactic Ambiguity Resolution

3.1 Introduction

During tasks demanding synchronous exchange of multimodal information, e.g. watching a movie, sentence processing has to interact with the other cognitive modalities that are also actively involved, e.g. vision. At the state of art, very little is known about the mechanisms underlying cross-modal interaction during synchronous processing.

Imagine being in a museum listening to an audio-guide describing a painting that we are simultaneously watching: the processing of linguistic descriptions e.g. *the woman depicted ...*, interact with visual information of the canvas, e.g. brightness, that our visual system is concurrently attending.

During cross-modal interaction, however, not all visual and linguistic information available is accessed and integrated at once; it is rather more plausible, instead, to assume that visual or linguistic information is selectively utilized to guide the allocation of visual attention; both depending on the task we are engaged with, e.g. sentence understanding, and the phase we are currently in, e.g. before vs after the description begins.

Suppose that we are approaching the painting and the audio-guide hasn't started yet. In this starting phase, we are freely viewing the scene, and visual attention is mainly directed by image-based mechanisms, e.g. color, intensity, orientation; however, as soon as the linguistic stream begins, visual information is utilized contextually with the linguistic information processed. So, the more linguistic information comes in, the more linguistically structured will the guidance of visual attention be; hence overriding the more 'primitive', i.e. image-based features, acting upon visual attention. Obviously, there are intermediate phases of cross-modal interaction, where visual and linguistic information might compete for visual attentional resources.

In this chapter, we investigate how image-based visual information interacts with prosodic information during a visually situated sentence comprehension task. We find that low-level visual information is utilized in the prediction of linguistic referential information of the sentence, especially when linguistic information is not sufficient, e.g. beginning of direct object, to generate a full prediction about the upcoming material.

Moreover, we investigate the pattern of interaction emerging when visual and linguistic information compete for visual attentional resources. We observe high independence in the way visual and linguistic information are accessed and utilized, which strongly relates to the phase of the task under processing. Furthermore, we find additive effects when both types of information cooperate. When linguistic and visual information point visual attention to the same target object, we observe more looks compared to when cues are tested independently.

3.2 Background

In psycholinguistic research, there is a growing body of eye-tracking research investigating sentence processing situated in visual contexts (**V**isual **W**orld **P**aradigm Tanenhaus *et al.* 1995). A large span of linguistic phenomena, across different levels of sentence processing, have been re-investigated in the light of a visual context, by looking at how referential contextual information is visually accessed under different linguistic manipulations. For example, prosodic cues (Snedeker & Trueswell, 2003; Snedeker & Yuan, 2008) and disfluences (Bailey & Ferreira, 2007) are observed disambiguating, referentially ambiguous visual contexts. Verb semantics (e.g. Altmann & Kamide 1999; Scheepers *et al.* 2008), and thematic role information, (e.g. Knoeferle & Crocker

2006, 2007) are incrementally used to make visual predictions, i.e. anticipation, about upcoming referents of the sentence. On the syntactic level, instead, it has been found that priming of di-transitive structures (Arai *et al.*, 2007), e.g. DO vs PO structure[1], is reflected on anticipatory eye-movements, at verb-site, launched to the visual object expected by the priming condition.
At all levels of sentence processing, it has clearly emerged that linguistic information is utilized in integration with the visual information present in the context, and this has led to the conclusion that visual attention is mediated by the interaction between utterance information (Crocker *et al.*, 2010) and visual context (Altmann & Mirkovic, 2009).

VWP has mainly focused on linguistic phenomena; thus making the simplifying assumption that visual attention during sentence processing would be mostly driven by linguistic stimulation. This assumption is supported by the experimental use of simple visual material, e.g. object arrays or clip-art scenes, which impoverishes visual responses, thereby making them more likely to be guided by linguistic information only.

A visual context, however, carries information which is actively implicated in the attentional mechanisms of visual cognition; especially, when naturalistic scenes are in the place of object arrays[2]. To the best of our knoweldge, beside the study of Huettig & Altmann 2007, which has shown how objects with similar shapes, e.g. a ROPE and a SNAKE, compete on visual attentional resources [3], and a few studies on visual cognition showing how linguistic information can boost search performance (e.g. Schmidt & Zelinsky 2009), not much work has been done to understand the relation between mechanisms of visual attention and sentence processing. During tasks activating synchronous processing, viz. situated sentence comprehension, different cognitive modalities, e.g. vision and language, have to exchange and integrate multi-modal, e.g. visual and linguistic, information in order to achieve the goals of the task, e.g. understanding a sentence in the context of a scene. Thus, in order to fully understand the mechanisms

[1] A DO structure is: *The man is reading the boy a book*; whereas a PO is: *The man is reading a book to the boy*.

[2] We explore explore this issue in Chapter 4.

[3] If SNAKE is mentioned, ROPE will get more looks compare to an object, e.g. TEDDY, which has a different visual shape.

of situated language processing it becomes crucial to establish which visual and linguistic factors are involved, when during a task, and, more generally, what the pattern of their interaction is.

In this chapter, we investigate the impact of low-level (image-based) visual information during processing of syntactically ambiguous PP-attachment sentences. Low-level features, e.g. *color*, *intensity* and *orientation*, are primitive visual information processed by the primary visual cortex. The unification of these features has been theoretically conceptualized and statistically quantified in the notion of *saliency* (Itti & Koch, 2000b); which is a measure of visual prominence computed by aggregating values for the different features at different spatial scales[1]. Saliency is expected to guide visual attention during free-viewing tasks. Fixations are expected to follow saliency information, from highest to lowest saliency with the first fixation launched on the region highest in saliency and the subsequents ordered by decreasing saliency[2]. In the absence of a specific goal, our attention is captured by the locations of the scene, which are richest in saliency (e.g. Parkhursta *et al.* 2002). When our visual system is instead used actively to achieve a certain goal (e.g. Findlay & Gilchrist 2001), e.g. search, the effect of saliency is overriden by high-level object-based cognitive control (Henderson *et al.*, 2007), i.e. we look at objects that are contextually relevant to our task[3].

We believe that situated language processing tasks are between the two extremes of having or not having a goal. A situated sentence understanding task is, in fact, constituted by two main phases: a free viewing phase (before speech onset) where participants inspect the visual material, and the sentence phase (during speech) where the linguistic information listened to is incrementally (i.e. word by word) mapped against the visual context, hence setting up the 'goals' of visual attention. Our expectation is that, in the absence of linguistic information and any specific goal, i.e. free viewing, low-level visual information is utilized to steer visual attention. Then, however, the more linguistic information is processed, the less need there is, for the visual system,

[1]For a Matlab implementation of the measure and a computational model of visual attention refer to Walther & Koch 2006

[2]The **I**nhibition **o**f **R**eturn mechanism makes sure that fixations do not return on previously inspected regions.

[3]For a computational model of visual attention integrating, partially, low and high level visual features refer to Torralba *et al.* 2006

to rely on low-level information. Obviously, there are intermediate stages during situated sentence understanding, where the linguistic information is not yet sufficient to generate a full prediction about which visual objects will be arguments of the sentence, e.g. at the beginning of direct object. This assumption is especially valid, when the visual context is an object array (no contextual dependency among the objects), and the linguistic material is not predictive of any specific object, as indeed is the case in Altmann & Kamide 1999. Thus, we would expect in these intermediate stages, e.g at verb and direct object sites, visual attention to be driven by low-level visual information, which is utilized as a predictive proxy to sentence information.

In three eye-tracking experiments, we investigate the cross-modal interaction between visual and linguistic low-level information during a situated language understanding task, involving referentially ambiguous sentences. In Experiment 1, we investigate the impact of visual saliency, which is expected to trigger anticipatory eye-movements on the salient object when the linguistic information parsed is not yet sufficient, i.e. between verb and direct object, to make a full prediction about upcoming post-verbal arguments. In Experiment 2, we test the effect of intonational breaks, already explored in the VWP literature (e.g Snedeker & Trueswell 2003) on the resolution of PP-attachment ambiguity. We choose intonational breaks as linguistic cues because they do not carry any explicit semantic information, and can be considered, in some respect, low-level, therefore more directly comparable to saliency. Our expectation is to replicate results from the literature, where intonational breaks give focus to the visual object which is referenced within the prosodic phrase enclosed by the breaks. Finally, in Experiment 3, we directly investigate the cross-modal interaction between saliency and intonational breaks by looking at their competition and cooperation. In the competitive scenario, visual saliency and intonational break cue visual attention to different visual objects to resolve referential ambiguity[1]. In the cooperative scenario, they both cue visual attention to the same target object. We assume an integrated cross-modal architecture of cognition, where the different modalities exchange information in a highly interactive manner to optimally achieve the goals of a specific task. In such an architecture, each modality gives a highly specific contribution to the multi-modal integration. Thus, we expect visual and linguistic information

[1]The different strategies of ambiguity resolution are interpreted in terms of eye-movement patterns on visual ROI, refer to Chapter 2 for more details about the approach.

to be selectively utilized at independent points of the task. For this reason, in case of conflicting cues, i.e. competition, we should observe the same pattern as if the cue was tested in isolation. However, in line with research on cooperative integration of cross-modal information (e.g. Evans & Treisman 2010), the joint contribution of visual and linguistic information should strengthen guidance of visual attention to the cued target object. Cooperation, however, occurs only during the phases of the trial where the joint contribution can optimally reinforce the information currently integrated.

3.3 Experiment 1: Visual saliency in syntactic ambiguity resolution.

In Experiment 1, we investigate the impact of low-level visual information, i.e. saliency, during comprehension of syntactically ambiguous sentences. In contrast to standard VWP research, which has focused on sentence processing phenomena, we test the hypothesis that mechanisms specific to the visual system, i.e. visual saliency, can also play an active role during situated sentence understanding. Before going into the details of the experiment, we first discuss syntactic ambiguity resolution in VWP. Here, we highlight the importance of visual referential competition, i.e. to *the apple on the towel* there are two visual apples depicted, in explaining the patterns of visual attention observed in the literature. Then, we describe the notion of visual saliency in the visual cognition literature and its implications for our experimental design.

3.3.1 Syntactic ambiguity resolution in the VWP

Most of the early work in VWP focuses on how a visual context can influence the resolution of syntactic ambiguity (e.g Snedeker & Trueswell 2003; Spivey-Knowlton *et al.* 2002; Tanenhaus *et al.* 1995). A syntactic ambiguity arises when phrases can be combined in different configurations, hence licensing different semantic interpretations of the sentence. A classic syntactic ambiguity is generated by prepositional phrase (PP) attachment, e.g. *put the apple on the towel in the box*: where the PPs can be interpreted either as modifiers, i.e. the apple that is on the towel, or goal locations, i.e. take the apple and put it on the towel. Obviously, the presence of a visual context

acts on the resolution of syntactic ambiguity by constraining the number of possible interpretations. In fact, to each context corresponds certain resolutions of ambiguity.

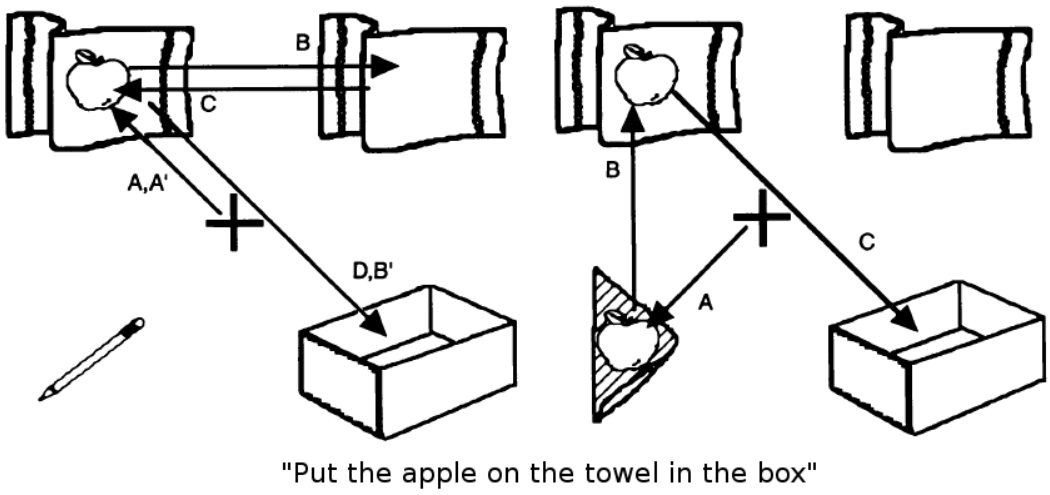


Figure 3.1: Example of visual contexts used in Tanenhaus *et al.* 1995. The arrows indicate how eye-movements are 'distributed' in the different visual contexts.

In Figure 3.1 we show the two visual contexts used by Tanenhaus *et al.* 1995 to test the effect of visual referential information on the resolution of syntactically ambiguous PP-attachment sentences, e.g. *put the apple on the towel in the box*. The main difference between the two visual contexts relies on the number of visual referents (1 or 2) available for the direct object *the apple*. The hypothesis tested is that in one-referent context at linguistic ROI *on the towel*, the visual referent EMPTY TOWEL is ambiguously interpreted as goal location, i.e. put the apple that is on the towel on the other towel; whereas in a two-referent context, the presence of a visual competitor for the linguistic referent apple, i.e. APPLE ON A NAPKIN, would trigger a modifier interpretation for the ambiguous PP *on the towel*, thus neutralizing the goal location effect driven by the visually depicted EMPTY TOWEL. The results show that the presence of a visual competitor helps the resolution of syntactic ambiguity, while more generally showing a clear effect of integration between visual and linguistic referential information. From this study it is evident that the visual context directly constrains the interpretation of a sentence. In particular, the visual referential competition in two-referent context seems to be responsible for the difference in looks. In the two-referent context at linguistic

ROI *on the towel*, APPLE ON TOWEL, APPLE ON NAPKIN and EMPTY TOWEL, are all possible visual candidates for the phrase *the apple on the towel*, however by the time the competition between APPLE ON TOWEL and APPLE ON NAPKIN is concluded and attention is shifted to the EMPTY TOWEL, the intervention of the second PP *in the bowl* shifts attention to the visual object bowl, de-facto neutralizing the goal interpretation of visual object EMPTY TOWEL. In other words, for the two-referent context, sentence processing doesn't have time to evaluate the candidate interpretation where EMPTY TOWEL is a goal location.

The interplay between visual context and utterance mediates the guidance of visual attention (Crocker *et al.*, 2010; Knoeferle & Crocker, 2006). A visual context provides the domain of referents upon which the linguistic information has to be mapped. Thus, the more referential ambiguity (both on the visual context and on the sentence) there is, the more visual referents compete on visual attentional resources. Referential ambiguity relates to the number of possible matchings available at the different point of parsing, and can be found at two levels of situated sentence processing: local and global. A local ambiguity is when by the end of the sentence a unique relation between linguistic and visual referents can be read. (e.g. Tanenhaus *et al.* 1995), i.e. the APPLE will finish in the EMPTY BOX. A global ambiguity is when interpretation remains ambiguous even after having processed the whole sentence, i.e. linguistic and visual referents can still be related in several ways (e.g. Bailey & Ferreira 2007, i.e. the APPLE can either finish in the EMPTY BOWL or on the TOWEL IN THE BOWL, refer to section 3.4 for more details). In this experiment, as in Bailey & Ferreira 2007, we set referential ambiguity both at the local and global level. We want to test how much visual referential competition is responsible for the patterns of visual attention observed during syntactic ambiguity resolution. Moreover, since we assume that visual information is actively utilized to make predictions of upcoming linguistic material, we investigate the impact of visual saliency during situated language understanding.

3.3.2 Visual saliency

Saliency is an image-based, low-level measure of visual prominence, based on the information of three visual features: color, intensity and orientation. For each of these

features, a conspicuity map is computed by taking the changes in values of the feature, at each point of the image, relative to surrounding points, across different spatial scales of comparison (see Figure 3.2 to visualize a saliency map computed using the SaliencyToolBox (Walther & Koch, 2006)). The three different maps are then aggregated into a unique map, the saliency map (Walther & Koch, 2006). Saliency doesn't generally match entire objects.

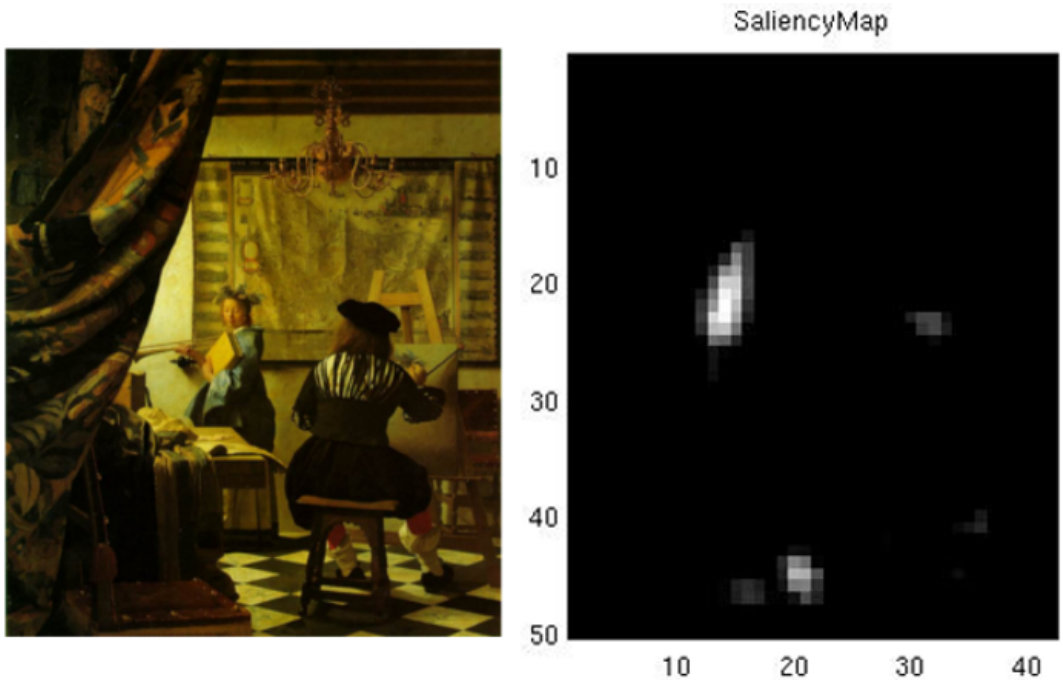


Figure 3.2: Example of a saliency map applied on the painting 'The Art of Painting' by Jan Vermeer (1666-72).

Edges or angles formed by two contiguous objects may have an high saliency, even if they are not whole objects. There is evidence, however, showing that locations rich in objects are positively correlated with saliency (Elazary & Itti, 2008). In our experimental design, since we use object arrays, to make our material comparable with standard VWP studies, saliency is mostly reflected by the prominence of individual objects, rather than image regions. Thus, we manipulate saliency at the level of individual objects. Saliency is expected to have effects only during free-viewing tasks (Henderson *et al.*, 2007). During a situated language processing task, the goals are

set by the linguistic information. At the beginning of the trial, before speech begins, participants are free-viewing the scene, thus image-based information is expected to guide visual attention. During speech, visual attention is expected to be driven by linguistic information. However, there are intermediate phases while speech is unfolding, e.g between verb and direct object, where linguistic information is not yet sufficient to generate a full prediction about the upcoming arguments of the sentence. During these phases, we expect saliency to be utilized by sentence processing as a visual predictive proxy for those upcoming arguments.

3.3.3 Method

In a sentence understanding eye-tracking experiment, we asked participants to listen to syntactically ambiguous PP-attachment sentences, e.g. *The girl will put the orange on the tray in the bowl*, while concurrently presented with a visual context, where *Number of referents* and *Saliency* of objects are manipulated (see Figure 3.3). In contrast with previous studies which are similarly designed, where participants are engaged in a behavioral task requiring an action response (see Novick *et al.* 2008; Snedeker & Trueswell 2003; Spivey-Knowlton *et al.* 2002; Tanenhaus *et al.* 1995), i.e. the participant had to complete the action told in the sentence; our experiment is a look-and-listen comprehension experiment. Thus, the use of imperative ambiguous sentence, *Put the orange on the tray in the bowl* has been substituted with *The girl will put the orange on the tray in the bowl* [1].

Moreover, similar to Bailey & Ferreira 2007, we use fully ambiguous visual contexts, i.e. both locally and globally ambiguous. Local ambiguity arises at the ambiguous PP-attachment phrases, i.e. *the orange on the tray*, and it is resolved after competition between the visual candidates SINGLE ORANGE, ORANGE ON TRAY and TRAY IN BOWL that can be referred to by the linguistic information processed. Global ambiguity arises at the end of the sentence, when there is still referential ambiguity on the possible mappings between linguistic and visual information, i.e. at *in bowl*, the ORANGE can finish either in the EMPTY BOWL or in the TRAY IN BOWL. A fully ambiguous visual context allows us to better explore the mechanism of visual competition.

[1] To account for this variation the subject of the action was introduced in the picture stimuli.

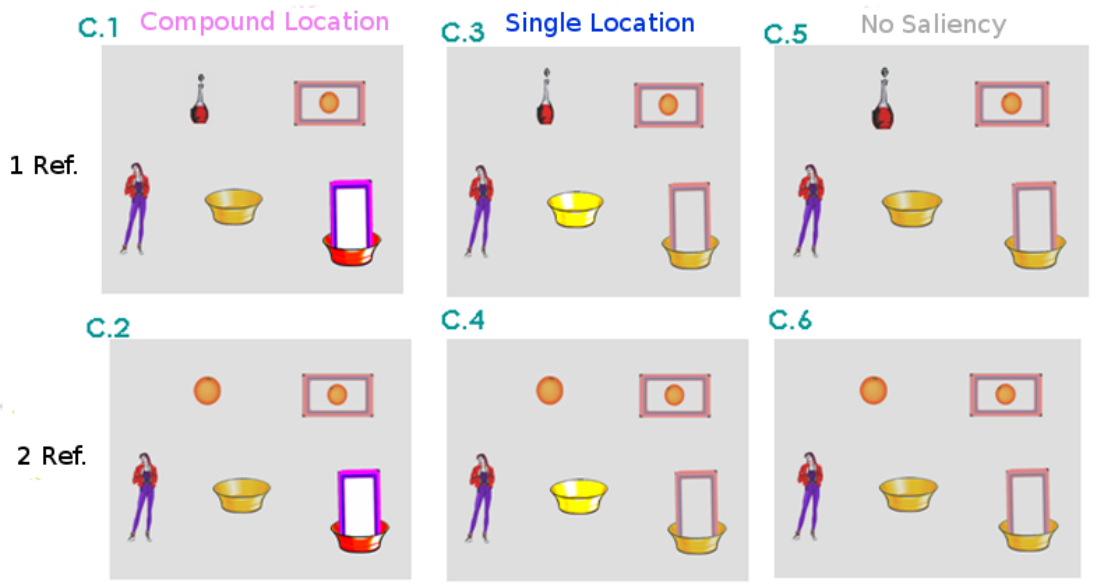


Figure 3.3: Conditions, 2 x 3: Number of referents (One, Two) crossed with Saliency (Single-Location, Compound Location, No Saliency).

We have a 2 by 3 design, crossing *Number of referents* (*1 Referent*/*2 Referent*) and *Saliency* (*Single-Location*/ *No-Saliency*/*Compound-Location*). The *Number of Referent* variable refers to the number of visual objects corresponding to the linguistic referent, direct object, *the orange*. In *1 Referent*, there is only one ORANGE depicted ON A TRAY, whereas in *2 Referents*, together with the ORANGE ON A TRAY, there is also a SINGLE ORANGE. The *Saliency* variable refers to the visual object carrying the saliency manipulation, on which we hypothesize an interpretation at the sentence level. *Single-Location* is the visual object BOWL, which in the sentence reflects the goal location, *in the bowl*, of the action *put*, on the direct object *the orange*. *Compound-Location* is the visual compound object, TRAY IN BOWL, which in the sentence corresponds to the prepositional modifier reading *on the tray in the bowl*, i.e. a tray which is located in a bowl. *No Saliency*, saliency was not manipulated.
We have, moreover, introduced an external condition of scene preview. We have split participants in two groups: **Long preview**, the participants had 1000ms of visual preview before the onset of speech stimuli; **Short preview**, speech and visual stimuli started simultaneously at beginning of the trial.

In naturalistic scenes, change of preview-time has a direct impact on search performance, with longer preview boosting identification of target objects (Vo & Henderson, 2010). We test the impact of preview during sentence understanding situated in an object array context.

3.3.3.1 Participants

Thirty participants from the University of Edinburgh were each paid 5 pounds for taking part in the experiment.

3.3.3.2 Materials

For each condition, a set of 36 experimental items was constructed. During the selection of visual objects we made sure that a depicted experimental item wasn't repeated more than twice; when repeated we used a different visual token. Visual objects were displayed without any imposed grid. The saliency of the target object has been modified using Photoshop CS2. The following manipulations were applied on target object: **Luminosity** +50 percent; **Contrast** +50 percent; **Colour balancing** RGB curves have been modified reinforcing the colors' dominance of target object; **Black and White** input and output curves have been changed only in the cases in which the edges' orientation of target object was too prominent. In order to highlight the saliency of the target object, we have manipulated also the background: **Luminosity** -30 percent; **Contrast** -30 percent. To validate the saliency manipulation, we used the Saliency Matlab ToolBox (Walther & Koch, 2006), by checking that the visual object of interest had the highest saliency relative to the other objects. Regarding the linguistic stimuli, some linguistic variability was introduced in the set of sentences by using, beside the verb *put*, also the synonyms *move*, *place*, and *lay*. Each of the 4 verbs was used for 9 sentences. Moreover in order to avoid prosodic effects, we apply cross-splicing to the spoken sentences utilized in Experiment 2, where intonational breaks are explicitly manipulated. In Experiment 2, we use two types of intonational breaks[1]:

(1) NP modifier: ...[the orange on the tray] BREAK [$_{\text{in}}$ the bowl]

(2) PP modifier: ...[the orange] BREAK [$_{\text{on}}$ the tray in the bowl]

[1] More details on the rationale of the experiment can be found in section 3.4.

which we splice and merge[1] to obtain a neutral prosody, as following:

(3) the orange [$_{PP}$ modifier] + on the tray [$_{NP}$ modifier] + in the bowl [$_{PP}$ modifier]

We take the direct object *the orange* and the second PP *in the bowl* from sentences produced using intonational break PP-modifier, and we merge them with *on the tray* from NP-modifier break. Between each phrase, a 50ms pause was added to yield a more realistic prosody. In addition to the 36 experimental items there were 48 fillers. In one third of the fillers, saliency was manipulated. The manipulation was to balance the total amount of pictures containing salient objects. For counterbalancing reasons, the position of objects was rotated in eight different configurations. Four configurations were created by rotating objects clockwise, the other four scrambling positions along diagonals. The position of the subject visual object was also interchanged, so that on half of the trials it was on the left side of the screen, the other half on the right. In order to keep participants engaged in the task, we asked 24 (yes/no) comprehension questions, 12 about the content of sentences and 12 about the content of the image. Trials were randomized and divided into 4 different lists using the Latin Square rotation. Individual lists were created for each participant making sure that between experimental items there was always at least one filler.

3.3.3.3 Procedure

An EyeLink II head-mounted eye tracker was used to monitor participants' eye-movements with a sampling rate of 500 Hz. Images were presented on a 21" multiscan monitor at a resolution of 1024 x 768 pixels. A test of eye dominance was performed at the beginning of each session and only the dominant eye was tracked. Participants were asked to wear swimming caps in order to avoid the eye-tracker sliding during the experiment. After 0 or 1000 ms of visual presentation (depending on the previewing condition), spoken sentences were concurrently played. The experiment was explained to participants using written instructions. At the beginning of every session, participants were given 4 practice trials to familiarize them with the experiment. Calibration was done at the beginning of the session and repeated again at approximately halfway

[1] We use Adobe Audition to manipulate the sound files.

through the session. Some subjects required more than two calibrations. Between trials a fixation cross appeared within which drift correction was performed. The entire experiment was approximately 30 minutes long.

3.3.3.4 Pre-processing and Analysis

Fixations are extracted and cumulated per visual object by superimposing templates with interest regions (visual objects) marked by colours and unfolded on a time course of 6000ms by slices of 10ms, using the open source software 'Filter' (Saarbrucken) and SBtrans (Saarbrucken). In our experiment we have 5 different visual objects WOMAN, SINGLE ORANGE OR DISTRACTOR, ORANGE/TRAY, TRAY/BOWL, BOWL over which fixations were counted. Blinks or out of range fixations were excluded. The onset of time course is aligned to the beginning of visual presentation. Furthermore, fixations are associated with linguistic regions counting from the onset up to 800ms. In this experiment, we have focused on three linguistic regions where ambiguity occurs: the direct object (ROI:NP *the orange*), the first spatial preposition (ROI:1PP *on the tray*), the second spatial preposition (ROI:2PP *in the bowl*). On the eye-movement data, we perform descriptive and inferential analyses, using the statistical programming language R. The descriptive analysis compares the distribution of fixations on target objects, e.g. BOWL, for the different 6 conditions, e.g. *Single-Location/1 Referent*, over time (800ms) using log-odds as dependent measure (Barr, 2008). The inferential analysis examines the effects of experimental predictors on the trend of fixations. We did our inferential analysis using linear-mixed effects model (LME) with empirical logit as dependent measure on a specific target, e.g. BOWL, over time. The predictors of our LME model, centered to avoid collinearity, are: *Saliency*, *Number of Referent* and *Time*. The random effects are: *Subject* and *Trial*. The final model is selected following an iterative stepwise forward selection procedure (for more details about the analysis, refer to Chapter 2), which allows us to have maximum statistical fit with a minimal number of parameters. We discuss our results in the context of the coefficient estimates of those factors significant after model selection.

3.3.4 Results

Before examining the main effects of our **within** subjects conditions (e.g. *Number of Referent*), we briefly discuss the **between** subjects manipulation *Preview* (*Long/Short*). We found that with a Short preview, participants are slightly slower in responding to linguistic stimulation, i.e. looks to target object after its mention come later compared to Long-preview.
The longer exposure to the visual array allows participants to identify and possibly recall the linguistic labels of all the objects, whereas in a zero preview, the participants are directly faced with the task of matching linguistic and visual information. Thus, compared to Short-preview, in a Long-preview, the visual referent is immediately identified after its linguistic mention. This difference does not however reach significance. Differently from a naturalistic scene, our visual array contains only few objects, which can be quickly identified during the first few fixations, and this probably weakens the impact of preview. For this reason, in the rest of our analysis, we analyze Long and Short Preview aggregated.

We divide the discussion by linguistic ROI, analyzing the results obtained for the different target objects. For this experiment we focus only on the linguistic ROI direct object *the orange*. On the the object ORANGE, in line with previous literature (e.g. Spivey-Knowlton *et al.* 2002), we expect more looks to single ORANGE compared to DISTRACTOR, when visually depicted, i.e. 2 Referent. The referential ambiguity between two visual objects sharing the same linguistic referent, i.e. *orange*, triggers attentional competition. On the single object BOWL and compound object TRAY IN BOWL, we expect anticipatory looks at the beginning of direct object, in prediction to upcoming arguments of sentence, when saliency on object is manipulated, i.e. Single-location (BOWL) Compound-location (TRAY IN BOWL). When linguistic information alone is not sufficient to generate a full prediction of the sentence, i.e. at the auxiliary/verb region, the saliency on objects is visually used to infer it.

ROI NP: direct object *the orange* Figure 3.4 compares empirical logit fixation on target object ORANGE (2 Referent) or distractor (1 Referent) across conditions. At the beginning of the NP, there is no clear effect of *Number of Referents*.

Table 3.1: Experiment 1. Linguistic **ROI** *NP direct object: the orange*; on the three Visual ROI: ORANGE, SINGLE-LOCATION, COMPOUND-LOCATION. Predicted LME coefficient estimates of predictors. *Saliency* is contrast coded (treatment). The level *No-Saliency* is used as reference level, contrasted with *Single-Location* and *Compound-Location*. *Number of Referents* has been centered, and since the experiment is balanced: 1 Referent was recoded as (-0.5) and 2 Referents as (0.5).

	ORANGE	
Predictor	Coefficient	*p*
Intercept	-4.8014	0.0001
Time	0.0096	0.001
Referent	0.0935	0.08
Single-Location	-0.0122	0.5
Compound-Location	0.0075	0.7
Referent:Time	0.0117	0.0001
Single-Location:Time	0.0083	0.0001
Compound-Location:Time	-0.0008	0.5
Referent:Single-Location:Time	0.0063	0.02
Referent:Compound-Location:Time	0.0057	0.01
	SINGLE-LOCATION	
Predictor	Coefficient	*p*
Intercept	-4.8570	0.0001
Single-Location	0.0978	0.0002
Compound-Location	-0.0514	0.01
Referent	-0.0539	0.1
Time	-0.0009	0.7
Single-Location:Referent	0.0353	0.01
Compound-Location:Referent	0.0199	0.1
Referent:Time	-0.0039	0.01
Single-Location:Time	-0.0035	0.001
Compound-Location:Time	-0.0001	0.9
	COMPOUND-LOCATION	
Predictor	Coefficient	*p*
Intercept	-4.7494	0.0001
Single-Location	-0.0695	0.01
Compound-Location	0.08	0.007
Time	0.0066	0.03
Referent	-0.0034	0.9
Single-Location:Referent	0.0617	0.001
Compound-Location:Referent	0.0353	0.001
Compound-Location:Time	-0.009	0.001
Referent:Time	-0.0046	0.001
Single-Location:Time	-0.0012	0.3
Compound-Location:Time	0.0031	0.01

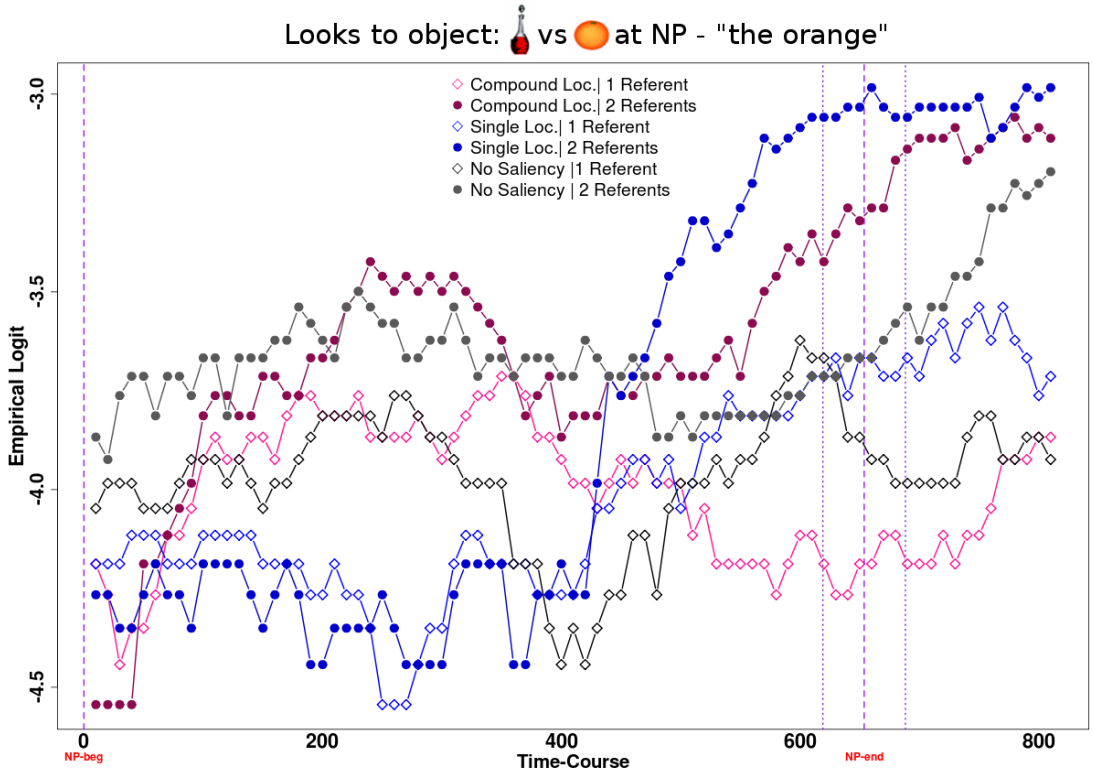


Figure 3.4: Experiment 1. Empirical logit fixation plot on ORANGE or DISTRACTOR at ROI:NP *the orange* for Long and Short preview collapsed.

However after the first 300ms, we observe increasing looks to ROI ORANGE/DISTRACTOR when 2 Referent are depicted, effects which strengthen over time (refer to Table 3.1 for list of coefficients). Moreover, we observe interactions with Time also for both saliency manipulations. When saliency is on Single-Location, looks increase over time, especially when 2 Referents are depicted. For saliency on Compound-location, instead, we observe a significant interaction with time only when associated with Number of Referents. Before the linguistic referent *the orange* is completely spelled out, salient objects are attracting visual attention, thus competing on looks with the mentioned object.

In Figure 3.5, we show empirical logit fixation on BOWL at ROI NP direct object *the orange*. At the beginning of the region, we observe a main effect of Saliency, in that for Single Location, there are significantly higher anticipatory looks, compared to both No-Saliency and Compound-Location. Saliency information generates antic-

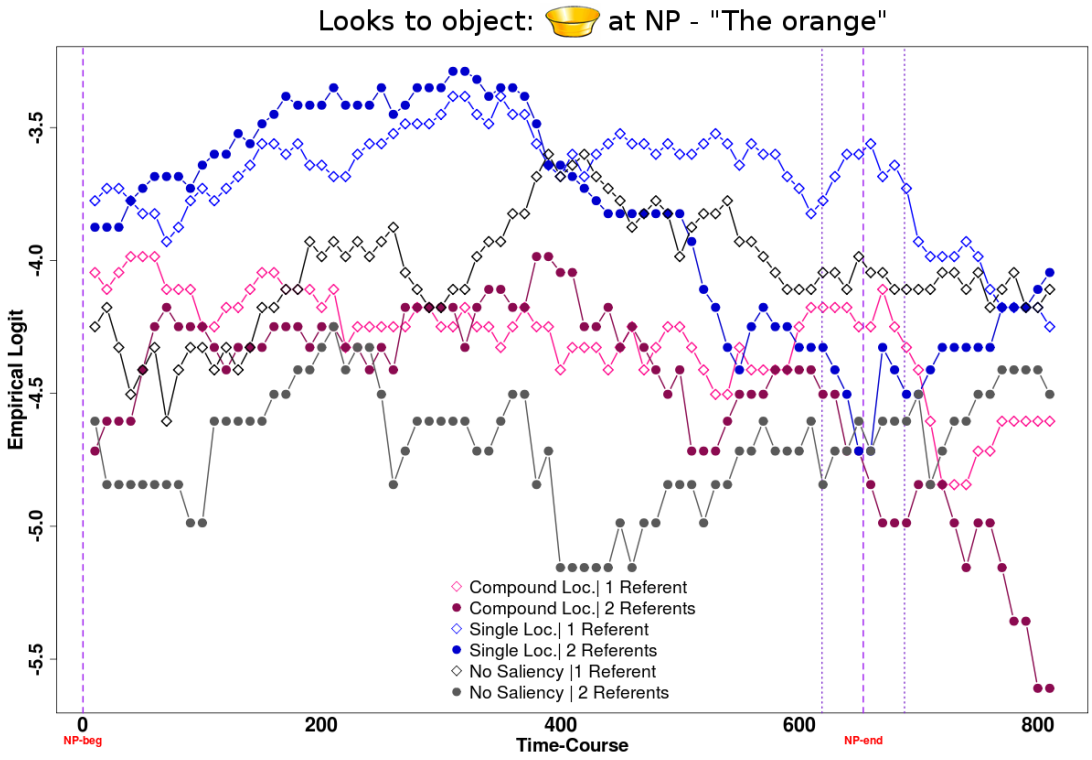


Figure 3.5: Experiment 1. Empirical logit fixation plot on BOWL at ROI:NP *the orange* across conditions

ipatory eye-movements at the beginning of direct object *the orange* used to predict upcoming linguistic material. However, as soon as the linguistic referent *the orange* is spelled out, looks on the salient object decrease over time; especially when 2 Referents are depicted, as more visual objects, sharing the same reference, are competing for attention.

Similar anticipatory effects are found also on the other salient object *Compound-Location*. In Figure 3.6 we show empirical-logit on TRAY IN BOWL at direct object *the orange*. We observe a main effect of saliency on Compound-Location, i.e. more looks at TRAY IN BOWL when visually salient, compared to Single-Location; which decreases over time and, similarly to what observed on BOWL, is negatively influenced by the number of referents. Also on the compound object we confirm that saliency is utilized to predict upcoming linguistic information. When linguistic information is not sufficient to generate a prediction about upcoming arguments, sentence understanding

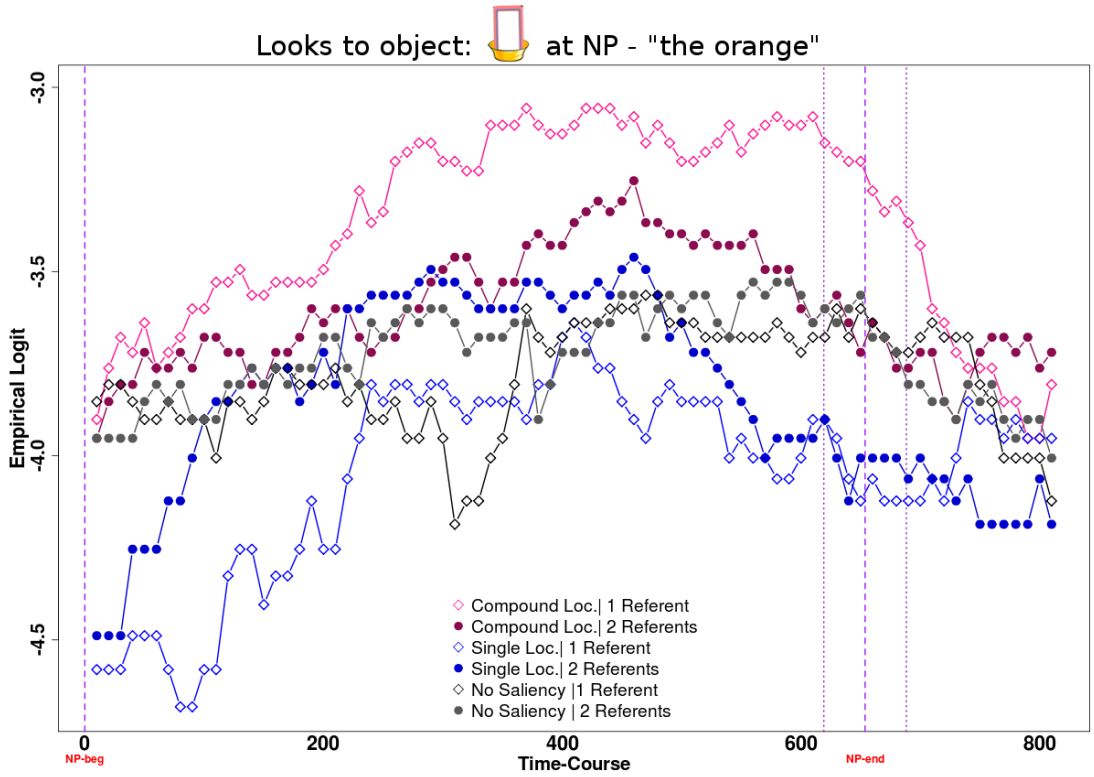


Figure 3.6: Experiment 1. Empirical logit fixation plot on TRAY IN BOWL at ROI:NP *the orange* across conditions.

relies on image-based visual information to provide such information.

3.3.5 Discussion

In this experiment, we have tested the hypothesis that purely visual information, e.g. saliency (Itti & Koch, 2000b), plays a role during situated sentence processing. We have found anticipatory looks on salient objects in both conditions, at the beginning of the direct object region, e.g. *the orange*. The effect was modulated by the number of visual referents sharing the same linguistic referent, i.e. ORANGE and ORANGE ON TRAY. The more referential ambiguity, the more visual competition. In line with previous work on syntactic resolution in VWP (e.g. Tanenhaus *et al.* 1995), we find effects of visual competition in two-referent context between visual objects sharing the same linguistic referent: at *the orange*, SINGLE ORANGE and ORANGE ON TRAY compete

for attention; especially, when saliency is not manipulated *No-Saliency*. When saliency is manipulated instead, we find that visual competition between the two-referent is delayed, as visual attention is evaluating whether the salient object is going to be the direct object of verb *put*. Since, in our visual context we do not have a visual object EMPTY TOWEL that can be interpreted as goal location of 1PP *on the towel*, we cannot directly assess whether, during this region, we find similar results to Tanenhaus *et al.* 1995.

In contrast with visual cognition studies showing that saliency has effect only during free-viewing tasks (Henderson *et al.*, 2009a), we observe an interaction of saliency during a sentence understanding task. The effect of saliency is, however, restricted to the phase (beginning of direct object) of linguistic processing where the information is not sufficient to generate a full prediction of upcoming linguistic material; considering also that the verbs utilized, e.g. *put*, do not favor any particular object of the array (unlike Altmann & Kamide 1999). Within this setting, situated sentence understanding can be configured as a free-viewing task, in that there is no precise goal, e.g. searching for a cued object, and participants do not know which visual objects to look at until they listen to the sentence. Moreover, even if the sentence has started, they need to process at least until the verb, before being able to make linguistically based predictions about sentence continuations. In this first phase of situated language processing, where the linguistic material available is not sufficient, saliency functions as a visual proxy to generate predictions.

The finding of interaction between low-level mechanisms of visual attention and sentence processing suggests a cross-modal architecture of cognition, where the different modalities interact during tasks requiring synchronous processing, e.g. sentence understanding. However, from these results we cannot tell whether during cross-modal interaction, the information coming from a certain cognitive process, e.g. sentence processing, takes precedence over the information gathered by a different cognitive process, e.g. visual attention. The reason is that in the experiment just described, visual information does not compete with linguistic information, but rather complements it. In the next two experiments, we are going to test whether linguistic or visual information takes precedence during situated sentence processing; or if they are used rather independently according to the different phases of synchronous processing. In particular, in experiment 2 we test the effect of 'low-level' linguistic information of

intonational breaks on the same experimental material of experiment 1. We use intonational breaks as linguistic manipulation because they do not carry any explicit semantic information, thus making their effect more comparable to saliency, which also, to some degree, doesn't carry any semantic information. In experiment 3, we put visual and linguistic information either in competition, i.e. saliency and intonational breaks point to a different resolution of ambiguity, or in cooperation, i.e. saliency and intonational breaks point to the same resolution of ambiguity. With experiment 2, we want to test the effect of intonational breaks independently from saliency; thus once we look at their interaction in experiment 3, we can fully compare the results obtained independently in experiment 1, saliency, and in experiment 2, intonational break, with those obtained during their interaction (experiment 3).

3.4 Experiment 2: Intonational breaks in syntactic ambiguity resolution

Based on the experimental design of experiment 1, in experiment 2 we investigate the effect of intonational breaks during situated understanding of syntactic ambiguous PP-attachment structures. We have chosen to manipulate intonational information because it doesn't carry any explicit high-level semantic information; hence making it more suitable to be tested against visual saliency (experiment 3), which also acts on low-level processing of visual information. Before going to the details of the current experiment, we contextualize our work with previous studies on the topic.

3.4.1 The effect of prosodic information during situated ambiguity resolution.

The importance of prosodic information on the resolution of referentially ambiguous visual context has emerged during communicative tasks, where especially intonational break information was used by speakers to contrast the intended visual object from its referential competitor, both by adults and children (Snedeker & Trueswell, 2003; Snedeker & Yuan, 2008).

In the study conducted by Snedeker & Trueswell 2003, pairs of participants, a Speaker and a Listener, were engaged in a dialogue/action task situated in a referentially ambiguous context (see Figure 3.7) to visualize an example trial). Speakers were asked to give instructions of actions to be performed by the Listeners[1], which had a PP-attachment syntactic ambiguity e.g. *Tap the frog with the flower*: where the 1PP *with the flower* can be interpreted either as instrumental, i.e. take the FLOWER and tap the FROG, or as modifier of direct object *the frog*, i.e. tap the frog that has a flower. The results show that speakers make use of prosodic information to resolve PP-attachment ambiguity; and especially prominent was the use of intonational break information (see Figure 3.7 to visualize). When a speaker intended a modifier interpretation, the ambiguous prepositional phrase *with the flower* was emitted together with the direct object in a single prosodic phrase, *the frog with the flower* (i.e. no intervening intonational break). Whereas an instrumental interpretation was obtained by introducing an intonational break after the direct object *the frog*. Further studies have observed that the disambiguating effect of intonational breaks, is reflected also by the eye-movement responses (Snedeker & Yuan, 2008): where an instrumental break, i.e. silent pause after the direct object *the frog*, triggered more looks to FLOWER, during the ambiguous region *with the flower*, compared to the modifier break.

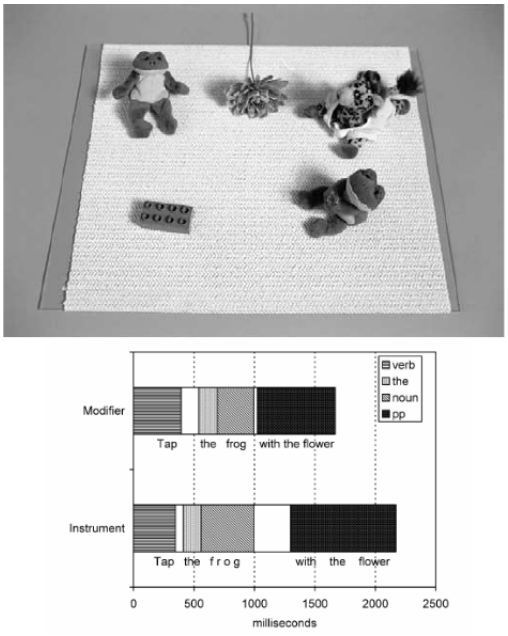


Figure 3.7: Example of trial, and results on prosodic information used by speakers in a dialogue study conducted by Snedeker & Trueswell 2003.

A similar effect, though weaker, was found by Bailey & Ferreira 2007 in a situated language comprehension study, where in place of intonational breaks there were filled pause disfluencies, e.g. *put the uh uh apple on the towel in the box.* In a similar vein,

[1]The speakers had to memorize for 10 sec the instruction, which was given in written form, and asked to repeat it verbatim.

as previously discussed VWP studies, participants were instructed to perform actions while listening to syntactically ambiguous PP-attachment sentences situated in a fully ambiguous visual context (unlike Snedeker & Yuan 2008; Tanenhaus *et al.* 1995); see Figure 3.8 to visualize an example trial. Filled pause disfluencies were introduced to allow different resolution of ambiguity.

For example, a filled pause introduced after the first prepositional phrase, e.g. *put the apple on the uh uh towel in the box*, should lead to the reading where '*the apple that's on the towel should go on the other towel that's in the box'*. Thus, more looks to the TOWEL IN BOX during and after processing of the disfluency. The results show a mild effect in this direction, which is however strongly connected to the type of context (i.e. one or two referents). Similar to Tanenhaus *et al.* 1995 a two-referent context triggers visual competition between the two APPLES, de facto reducing the time to evaluate the other possible reading where *the apple that's on the towel is put on the towel that's in the box*. In a one-referent context, instead, where there is no referential ambiguity between the two APPLES, the competition emerges at 1PP *towel* between the two compound objects sharing TOWEL, i.e. APPLE ON TOWEL and TOWEL IN BOX.

In Figure 3.8, we can observe, in fact, that while *the towel* unfolds, for the one-referent context both APPLE ON TOWEL and TOWEL IN THE BOX have a higher probability of looks; whereas in two-referent context a similar pattern is found between APPLE ON TOWEL and APPLE ON NAPKIN. Beside the effects driven by the different syntactic readings of the ambiguous sentence, we believe that a simple mechanism of visual competition might be also responsible for the patterns of eye-movement observed in both the Tanenhaus *et al.* 1995 and Bailey & Ferreira 2007 studies. Obviously, as the sentence unfolds over time, the linguistic referents that visually compete change. Thus the more visual competition there is at each phrase, the less time there is to evaluate other possible readings before a new phrase is parsed and another visual competition emerges. Together with visual competition, we expect eye-movement to be also influenced by the syntactic/semantic plausibility of the event integrated; e.g. a two-referent context makes more plausible the reading where a single ORANGE is moved in the TRAY IN THE BOWL; rather than in one-referent context, where only the ORANGE ON TRAY can be moved into TRAY IN BOWL, see Figure 3.9 to visualize example.

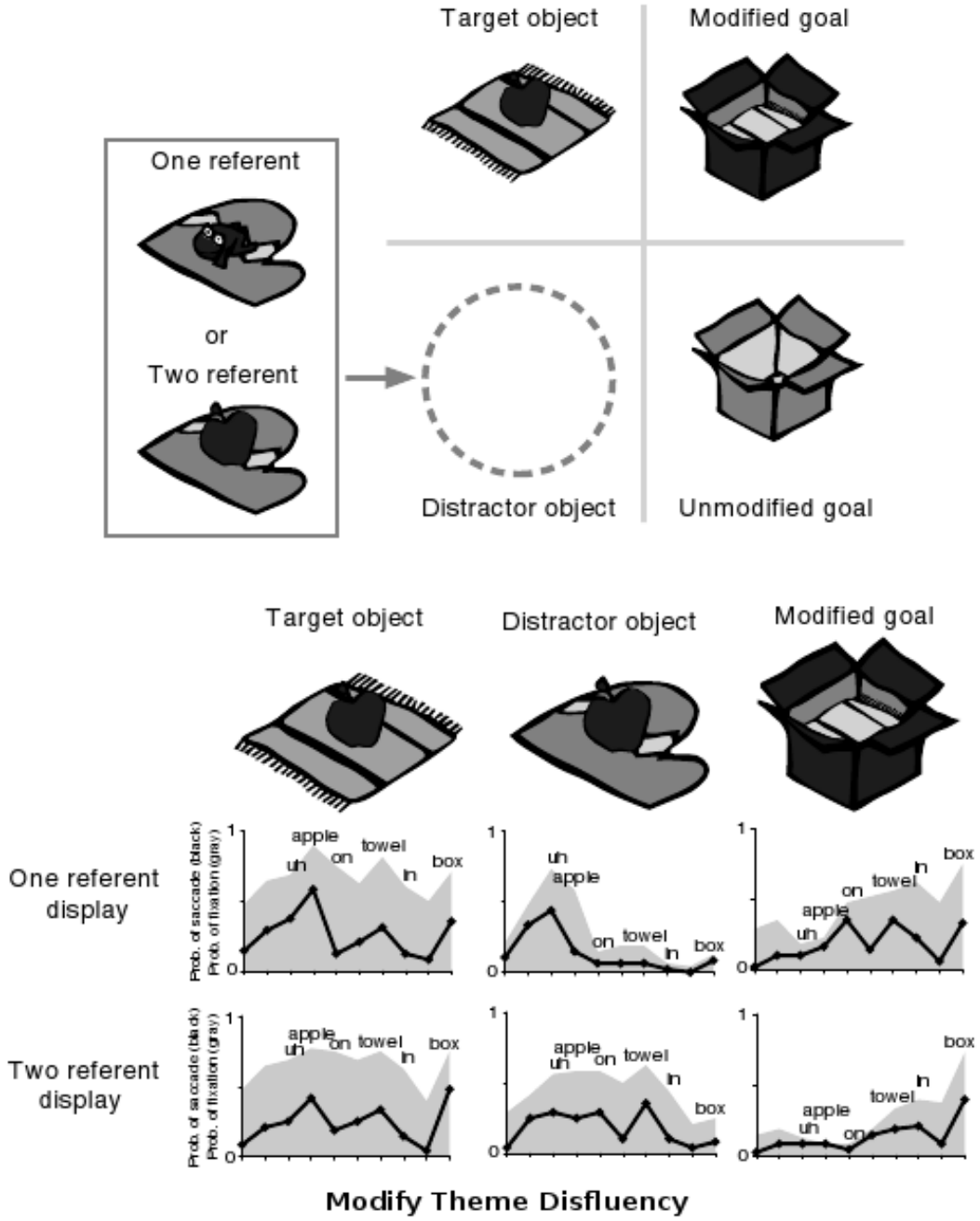


Figure 3.8: Top row: Example of fully ambiguous visual context. Bottom row: Probability of looks to APPLE ON TOWEL, APPLE/DISTRACTOR ON NAPKIN, TOWEL IN BOX, in one and two-referent context for modifier disfluency condition, e.g. *put the uh uh apple on the towel in the box*. The gray polygon indicates probability of fixations, whereas the line refers to saccade. Extracted from a study by Bailey & Ferreira 2007

In this experiment, we test the effect of intonational breaks during disambiguation of ambiguous syntactic structures, which have shown a clearer effect of disambiguation than filled paused disfluencies. In line with experiment 1, we have chosen to situate it in a fully ambiguous visual context to have more visual competition between objects. We expect the visual objects corresponding to phrases bounded by certain intonational breaks receive more looks compared to the different intonational condition (Snedeker & Yuan, 2008). Moreover, visual competition is expected to change according to the phrase observed and, the plausibility of the reading integrated. Finally, we test the independent effect of intonational breaks before investigating it in interaction with visual saliency (Experiment 3).

3.4.2 Method

In a 2x2 eye-tracking experiment crossing *Number of Referent* (1 Referent/ 2 Referents) and *Intonational-Break* (NP-modifier/PP-modifier), similarly to experiment 1, participants listened to PP-attachment ambiguous sentences, such as *the woman will put the orange on the tray in the bowl* while concurrently being presented with a fully ambiguous visual context. We manipulate the intonational breaks similarly to Snedeker & Yuan 2008. Thus, we consider two cases:

(4) *NP-modifier*: Intonational break after the first prepositional phrase (1PP)
   a. *[The girl will put] [the orange on the tray] [in the bowl]*

(5) *PP-modifier*: Intonational break after the second prepositional phrase (2PP)
   a. *[The girl will put the orange] [on the tray in the bowl]*

For the NP-modifier reading, the intonational break is placed after the 1PP modifier, and the phrases *the orange* and *on the tray* are together into a single prosodic phrase to trigger the unambiguous reading where: the orange that's on a tray is put in the empty bowl. In terms of eye-movement, for NP-modifier we expect more looks to ORANGE ON TRAY at 1PP *on the tray*, compared to the other condition. For the PP-modifier reading, the intonational break is placed after the NP direct object, and the phrases *the orange* and *on the tray* are separated by an intonational break; whereas the second and third PP are now aggregated *on the tray in the bowl*. The separation between the

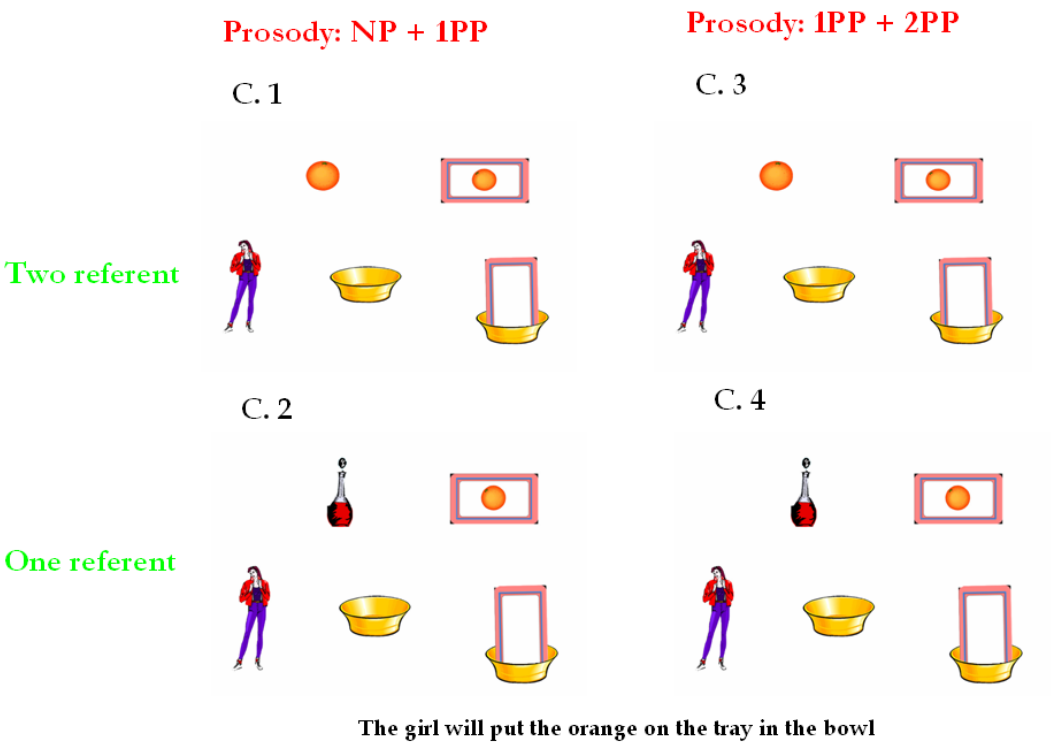


Figure 3.9: Experiment 2: example of visual and linguistic material across the different 4 Conditions.

NP and its 1PP modifier puts the single object ORANGE in referential focus[1] whereas the aggregation 1PP and 2PP does it for the compound object TRAY IN BOWL, thus triggering the reading where the single orange is put on the tray that's in the bowl. In general, we expect visual attention to focus on the objects following the order in which they are mentioned; thus, the intervention of a break on sentence processing signals a visual check on the objects which correspond to the linguistic information processed up to the interruption.

As for experiment 1, *Number of Referents* refers to the number of visual objects corresponding to the direct object of the sentence, e.g. *the orange*. In a 2 Referents visual context, there is a SINGLE ORANGE and AN ORANGE ON A TRAY; whereas in 1 Referent condition, only an ORANGE ON A TRAY is depicted, see Figure 3.9 to

[1] At least during the break.

visualize the four conditions.

3.4.2.1 Participants

Thirty-two participants, native speakers of English from the University of Edinburgh, with normal or corrected to normal vision, were each paid 5 pounds for taking part in the experiment.

3.4.2.2 Materials, Procedure and Analysis

We reuse the same images of experiment 1 but this time saliency is not manipulated (refer to Figure 3.9 to visualize an example trial). For the spoken stimuli, we use the same list of sentences of experiment 1, but repeated for two conditions of intonational break. A female speaker was instructed to read aloud the sentences[1] placing intonational breaks according to the two conditions (NP-modifier, PP-modifier). For the NP-modifier condition a mean break of 413.25 *ms* was placed between the end of 1PP *on the tray* and the beginning of 2PP *in the bowl*; whereas for the PP-modifier, we had a mean break of 637.54 *ms* between the end of direct object *the orange* and the beginning of 1PP *on the tray*.

The experimental procedure is similar to experiment 1; the main difference is that now we have only one preview condition of 1000ms, instead of two (refer to section 3.3.3 for more details). Participants preview the image for 1000ms before the spoken sentence is concurrently played. Calibration is done at the beginning of each session and manual drift correction is done between trials. As for experiment 1, we analyze looks on a target object, across conditions, aligned to the different linguistic ROI; the predictors of the LME models are *Number of Referents*, *Intonation Breaks* and *Time* and the random effects are *Subject* and *Trials*. For this experiment, we consider three linguistic ROI: the NP direct object *the orange*, the 1PP *on the tray* and the 2PP *in the bowl*. We show plots of observed fixation data, calculated as empirical logit, and discuss LME coefficients for those factors remaining significant after model selection.

[1] Recorded using a standard microphone.

3.4.3 Results

For the ROI:NP direct object, e.g. *the orange*, we look at the pattern of fixations on the object ORANGE. Here, similarly to experiment 1 and in line with previous work on ambiguity resolution (e.g. Tanenhaus *et al.* 1995), we expect in the two referent-context visual competition, i.e. when two objects depicted share the same referring expression, they will compete for visual attention while the referent is mentioned. Moreover, visual competition is expected to persist also during the other phrases due to the full ambiguity of the visual context. On ROI:1PP, e.g. *on the tray*, the different ambiguous mappings of the sentence on the visual context arise, and the disambiguating effects of intonational breaks are expected to emerge (Snedeker & Yuan, 2008). We analyze fixations on three different objects, ORANGE ON TRAY, TRAY IN BOWL and BOWL. On object ORANGE ON TRAY, we expect more looks for intonational break NP-modifier, where there is no break present between *the orange* and *on the tray*, compared to PP-modifier, especially when only 1 Referent is depicted, i.e. less referential competition. On the contrary, on object TRAY IN BOWL, we expect more looks for intonational break PP-modifier, where there is a break between the NP direct object and 1PP, and instead 1PP and 2PP are aggregated within the same prosodic phrase *the tray in the bowl*. Moreover, on object BOWL, we expect anticipatory effects for intonational break NP-modifier, where participants imagine a final reading of the sentence where BOWL is a goal-location *in the bowl*, i.e. *the orange* that's *on the tray* is put *in the bowl*. Finally, on ROI:2PP, we look at objects BOWL and TRAY IN BOWL, to check whether the two intonational breaks NP and PP-modifier converge on the two final readings discussed in section 3.4. However, since we find no significant effects on the compound object TRAY IN BOWL[1], we only report results on object BOWL.

3.4.3.1 ROI:NP direct object *the orange*

In Figure 3.10, we show a plot of empirical logit on object ORANGE OR DISTRACTOR across conditions during the linguistic region *the orange*.

At the beginning of the region, we observe a preference of looks to the object, for PP-modifier break, which is not, nevertheless, statistically significant. Also Number

[1]Probably the visual complexity of the compound object, and often its semantic implausibility, has made it less likely to be given the final reading of goal location.

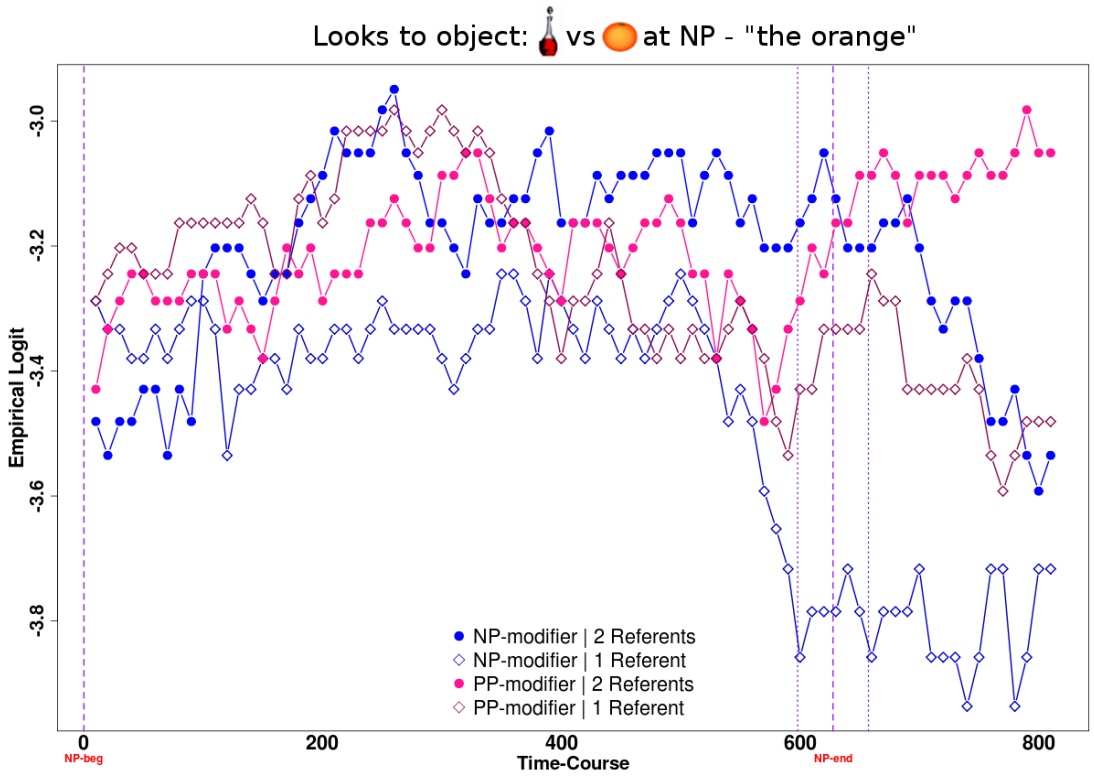


Figure 3.10: Experiment 2. Empirical logit of fixations on target object ORANGE/DISTRACTOR at ROI:NP *the orange* across conditions.

of Referent does not emerge as a main effect. However, as time develops, we observe increasing looks for *2 Referent* compared to *1 Referent*, especially in interaction with intonational break *PP-modifier* (refer to Table 3.2 for coefficients).

Similar to experiment 1, the effect of referents is found only when time is considered. In fact, the competition between visual objects sharing the same linguistic reference begins when it is completely unfolded. Moreover, a PP-modifier break, compared to the NP-modifier, puts the direct object *the orange* in focus; and before the 1PP begins, i.e. during the break, looks to ORANGE have the time to develop.

3.4.3.2 ROI:1PP (modifier vs location) *on the tray*

In Figure 3.11 we show empirical logit on object TRAY IN THE BOWL at linguistic ROI *on the tray*.

Table 3.2: Experiment 2. Linguistic **ROI** *NP direct object: the orange*; on visual ROI: ORANGE. Predicted LME coefficient estimates of predictors. Explanatory variables are centered around the mean. *Number of Referent*: 1 Referent (-0.5), 2 Referent (0.5) and *Intonational Break*: NP-modifier (-0.5), PP-modifier (0.5).

	ORANGE	
Predictor	Coefficient	*p*
Intercept	-3.5422	0.0001
Referent	0.0489	0.1
Prosody	0.0313	0.2
Time	0.0008	0.2
Referent:Time	0.0066	0.0001
Referent:Prosody:Time	0.0087	0.01

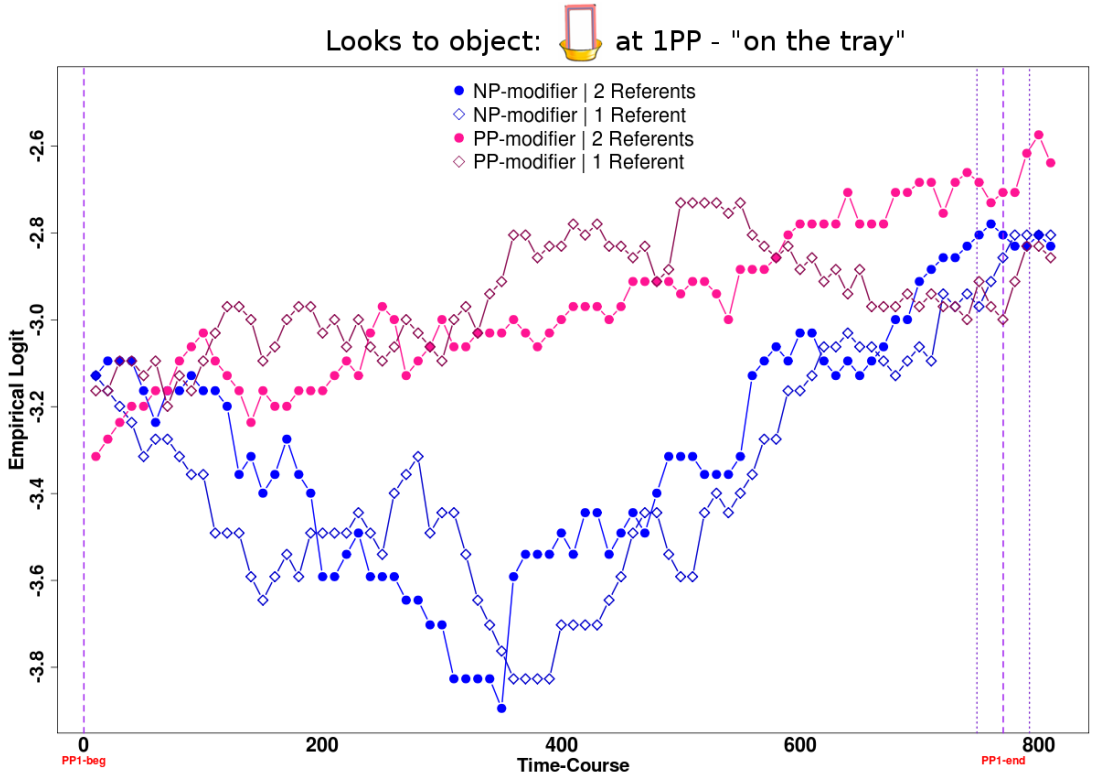


Figure 3.11: Experiment 2. Empirical logit of fixations on target object TRAY IN BOWL at ROI:1PP *on the tray* across conditions.

At the onset of ROI looks do not differ significantly across conditions. However, after the first 200ms, we observe an increase in looks, for PP-modifier, which is statistically significant only in interaction with 2 Referents (refer to Table 3.3 for coefficients). The presence of a single ORANGE in the visual context (2 Referent), opens the reading where the single orange could be moved on the TRAY IN THE BOWL, whereas on the 1 Referent context, the orange is already on a supporting object, thus making this reading less likely.

Table 3.3: Experiment 2. Linguistic **ROI** *1PP: on the tray*; on the three Visual ROI: COMPOUND-LOCATION, ORANGE ON TRAY, SINGLE-LOCATION. Predicted LME coefficient estimates of predictors. Explanatory variables are centered around the mean: *Number of Referent*: 1 Referent (-0.5), 2 Referent (0.5) and *Intonational Break*: NP-modifier (-0.5), PP-modifier (0.5).

	COMPOUND-LOCATION	
Predictor	Coefficient	p
Intercept	-3.4934	0.0001
Time	0.006	0.001
Prosody	0.0444	0.2
Referent	0.0406	0.1
Referent:Prosody:Time	0.0101	0.0001
	ORANGE ON TRAY	
Predictor	Coefficient	p
Intercept	-3.1901	0.0001
Referent	-0.0858	0.07
Prosody	-0.0732	0.09
Time	0.0027	0.09
Prosody:Time	-0.0138	0.0001
Referent:Prosody	0.0479	0.02
Referent:Time	0.0040	0.03
Referent:Prosody:Time	0.0085	0.02
	SINGLE-LOCATION	
Predictor	Coefficient	p
Intercept	-3.6261	0.0001
Time	0.0015	0.3
Referent	-0.0074	0.7
Prosody	0.0115	0.5
Time:Prosody	0.0073	0.0001

This result contrasts with Bailey & Ferreira 2007, who, instead, have found increasing looks to TOWEL IN BOX only for one-referent context. The reason for this difference might be that their visual competitor APPLE was depicted on a supporting object TOWEL, whereas our competitor is depicted alone; and this had a negative influence on the plausibility of the reading where *'the apple that's on a napkin has to be moved on the towel that's in the box'*. Beside the effect of visual competition driven by referential overlap [1], it seems that the compositional plausibility of the event integrated also influences the way eye-movements are distributed across the objects, while modulating the effect of intonational breaks. In one-referent context, where only ORANGE ON A TRAY is depicted, it can be more plausibly imagined to be combined with a goal location BOWL as final destination of the action *put*, rather than with the compound object TRAY IN BOWL. In a two-referent context instead, the presence of a single ORANGE, which is not on any supporting object (unlike Bailey & Ferreira 2007), opens also the possibility to be combined with the compound location TRAY IN BOWL, as final destination of the action *put*. The semantic plausibility of the integrated event is also implicitly connected by the structural plausibility of resulting sentence: a single object is just an NP, whereas a compound-object is an NP modified by a PP; by combining a single object with a compound object, we obtain a simpler and more plausible structure, than combining two compound objects:

(6) a. $[_{NP}$ orange $[_{PP}$ on $[_{N}$ a tray$]]]$ $[_{PP}$ in $[_{N}$ bowl$]]$
b. $[_{NP}$ orange $[_{PP}$ on $[_{N}$ a tray$]]]$ $[_{PP}$ on $[_{N}$ a tray$]$ $[_{PP}$ in $[_{N}$ a bowl$]]]$

We observe similar effects also on the other compound object ORANGE ON TRAY.

In Figure 3.12, again, we do not observe a main effect of Intonational breaks, however, over time, looks have an increasing trend for NP-modifier, especially when only 1 Referent is depicted, see Table 3.3. When 2 Referents are depicted, the referential competition interferes with the effect of intonational break, while making looks more sensible to time. Intonational breaks are used to put in focus the referent bounded by the prosodic phrase. Crucially, however, the effect is dependent on time, over which prosodic information unfolds, and modulated by referential competition: the more visual referents have to be evaluated, the weaker is the effect of the break.

[1] ORANGE ON TRAY and TRAY IN BOWL share object TRAY when the referring expression used is *on the tray*.

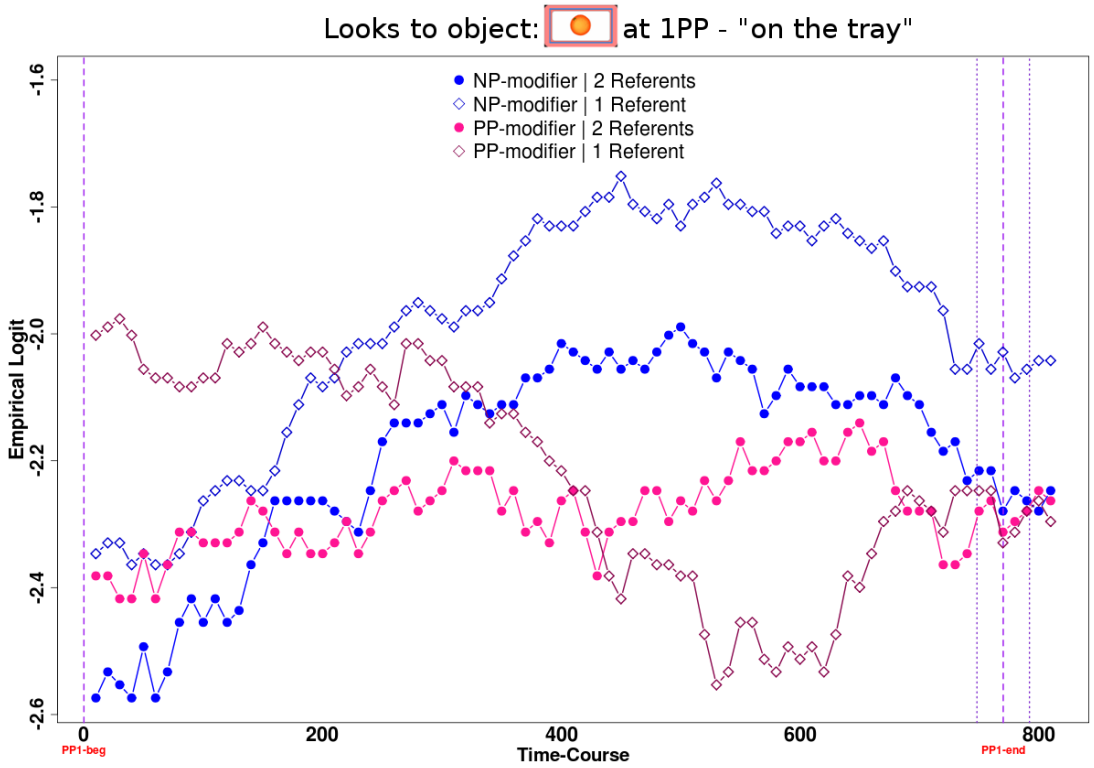


Figure 3.12: Experiment 2. Empirical logit of fixations on target object ORANGE ON TRAY at ROI:1PP *on the tray* across conditions.

Finally, in figure 3.14 we show plot of fixations on target object BOWL.

Confirming previous findings, the effect of Intonational break, NP-modifier, is only found in conjunction with time. Moreover, it is interesting to notice, how the effect relates to NP-modifier, which favors the final reading where the orange that is on the tray is put in the empty bowl. The semantic and structural plausibility of the event resulting by the integration of linguistic and visual information seems to play an important role[1].

3.4.3.3 ROI 2PP: *in the bowl*

In Figure 3.14, we show plots of observed and estimated looks on target object BOWL during linguistic ROI *in the bowl*.

Similar to previous results, we observe no main effects but only interactions (see

[1]More research is needed to disentangle semantic and structural plausibility of the event.

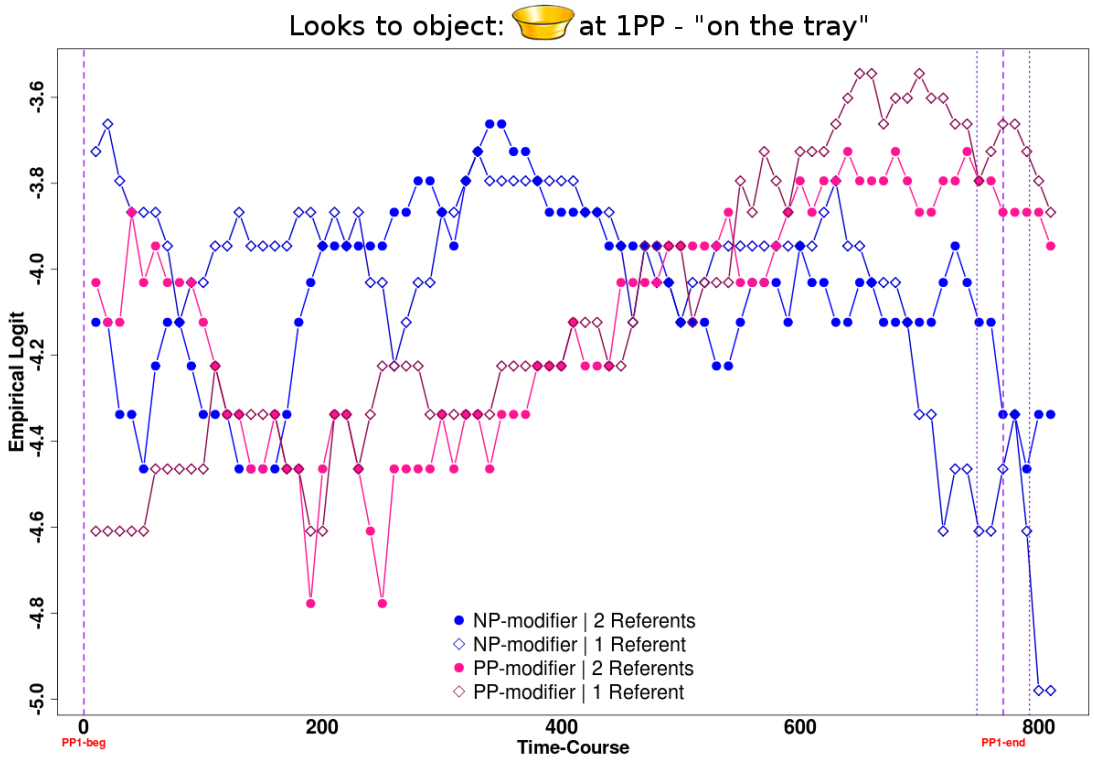


Figure 3.13: Experiment 2. Empirical logit of fixations on target object BOWL at ROI:1PP *on the tray* across conditions.

Table 3.4 for coefficients). Especially prominent is the positive interaction between NP-modifier and 2 Referents; where the effect of 2 Referents weakens over time. In line with previous studies (Snedeker & Trueswell, 2003; Snedeker & Yuan, 2008), the effect of Intonational breaks is to highlight the object in the visual context referenced by the prosodic phrase. Thus for an NP-modifier break, *in the bowl* is mentioned within a single prosodic phrase, which has a direct visual correspondence (BOWL).

This correspondence favors the referential integration between visual, e.g. BOWL and linguistic information, e.g. *the word* over other potential candidates of integration, e.g. TRAY IN BOWL with *in the bowl*. Moreover, the correspondence is modulated by semantic and structural plausibility of the resulting integration. When there are 2 Referents, both allowing competing interpretations of the event, i.e. a single orange in the bowl Vs an orange on the tray in the bowl, the target location BOWL receives more looks than for 1 Referent, where only one continuation is possible, i.e. the orange on

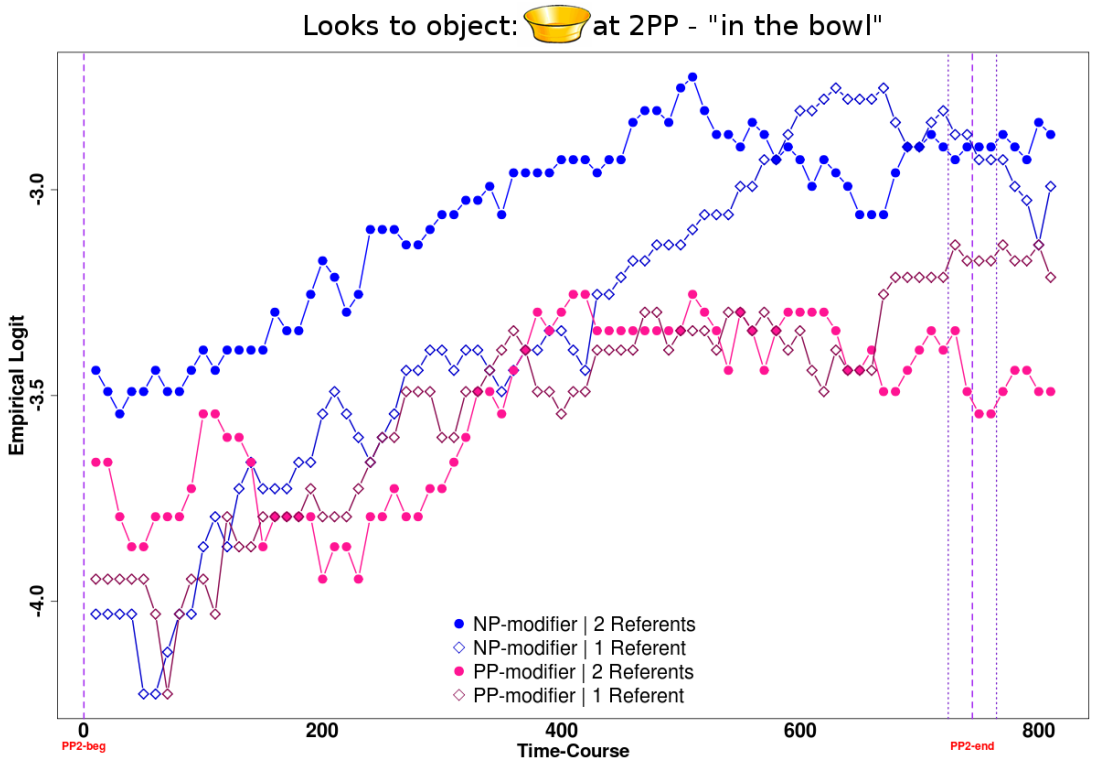


Figure 3.14: Experiment 2. Empirical logit of fixations on target object BOWL at ROI:2PP *in the bowl* across conditions.

the tray in the bowl.

3.4.4 Discussion

The goal of this experiment was to test low-level linguistic information, *Intonational-break*, on the resolution of syntactically ambiguous sentences, before investigating it in interaction with visual saliency (Experiment 3). Intonational breaks are interruptions of the linguistic stream marking the boundaries of a prosodic phrase. In our experiment, a prosodic phrase, e.g. *the orange on the tray* visually corresponded to a certain visual object, e.g. ORANGE ON TRAY. In line with previous studies (Bailey & Ferreira, 2007; Snedeker & Trueswell, 2003; Snedeker & Yuan, 2008); we observe that by changing the position of intonational breaks along the sentence, we change the organization of prosodic phrases hence influencing the integration between linguistic

Table 3.4: Experiment 2. Linguistic **ROI** *2PP: in the bowl*; on Visual ROI: SINGLE-LOCATION. Predicted LME coefficient estimates of predictors. Explanatory variables are centered around the mean: *Number of Referent*: 1 Referent (-0.5), 2 Referent (0.5) and *Intonational Break*: NP-modifier (-0.5), PP-modifier (0.5).

	SINGLE-LOCATION	
Predictor	Coefficient	*p*
Intercept	-3.5334	0.0001
Time	0.007	0.002
Prosody	-0.0444	0.1
Referent	0.0296	0.3
Prosody:Referent	-0.0704	0.0001
Time:Referent	-0.0029	0.03

referents mentioned and visual objects attended. This effect is especially strong during ambiguous regions, e.g. ROI:1PP *on the tray*, which trigger a wider range of readings. The need for disambiguation gives prominence to the intonational break information. However, all Intonational Break effects are found to be significant only in interaction with the other factors, time and number of referents. The number of referents is a fundamental trigger of visual competition: the more visual referents correspond to the referring expression, the more competition; which has also a direct impact on time: the more time is taken at each region by visual competition, the fewer readings of the ambiguous sentence can be evaluated. Time is also related to prosodic information; as prosodic information unfolds over time, eye-responses are conditioned from it. Furthermore, we found that semantic and structural plausibility of the event, resulting by the integration of linguistic and visual information, plays a critical role on visual responses. For example, if the visual context had 2 Referents for the direct object *the orange*, we observed more looks on object BOWL during linguistic ROI *in the bowl* compared to 1 Referent. In a 2 Referents context, both objects, SINGLE ORANGE and ORANGE ON TRAY can be put in the goal location EMPTY BOWL, thus yielding higher looks on target object.

With experiment 1 and 2, we have observed that both visual and linguistic, low-level, information is used during situated language understanding to resolve syntactic ambiguity. However, we do not know yet if the interaction between these two types of information produces results, as independently observed in experiment 1 and 2, or

instead a different pattern emerges. Replicating results of experiment 1 and 2 would suggest an highly interactive architecture of cognition, where visual and linguistic information, depending on the state of integration, is independently accessed and utilized. On the other hand, finding a new pattern, e.g. only linguistic information is used, would imply a more structured organization of cross-modal processing, where certain information, e.g. linguistic information, takes precedence over the other, e.g. visual information.

3.5 Experiment 3: Interaction of visual saliency and intonational breaks

From previous experiments we have seen that the independent manipulation of low-level visual (experiment 1, saliency) and linguistic (experiment 2, intonational breaks) information, during comprehension of syntactically ambiguous sentences situated in a fully ambiguous visual context (for details refer to section 3.3.3) resulted into different patterns of visual resolution. In this last experiment, we test the interaction (competition and cooperation) between visual and linguistic information by bringing together within the same experimental design both types of information. The goal is to discover the relation between visual and linguistic information during synchronous processing. Three possible scenarios of interaction can be imagined: 1) linguistic prominence: linguistic information overrides effects of visual information; 2) visual prominence: visual information overrides effect of linguistic information; 3) full interaction: both types of information are used at different points of synchronous processing, when needed.

3.5.1 Method

In a 2x2 eye-tracking experiment crossing *Intonational-Breaks* (NP-modifier/PP-modifier) and *Saliency* (Single-Location/Compound-Location), similarly to experiment 1 and 2, participants listened to PP-attachment ambiguous sentences, such as *the woman will put the orange on the tray in the bowl* while concurrently presented with a fully ambiguous visual context.

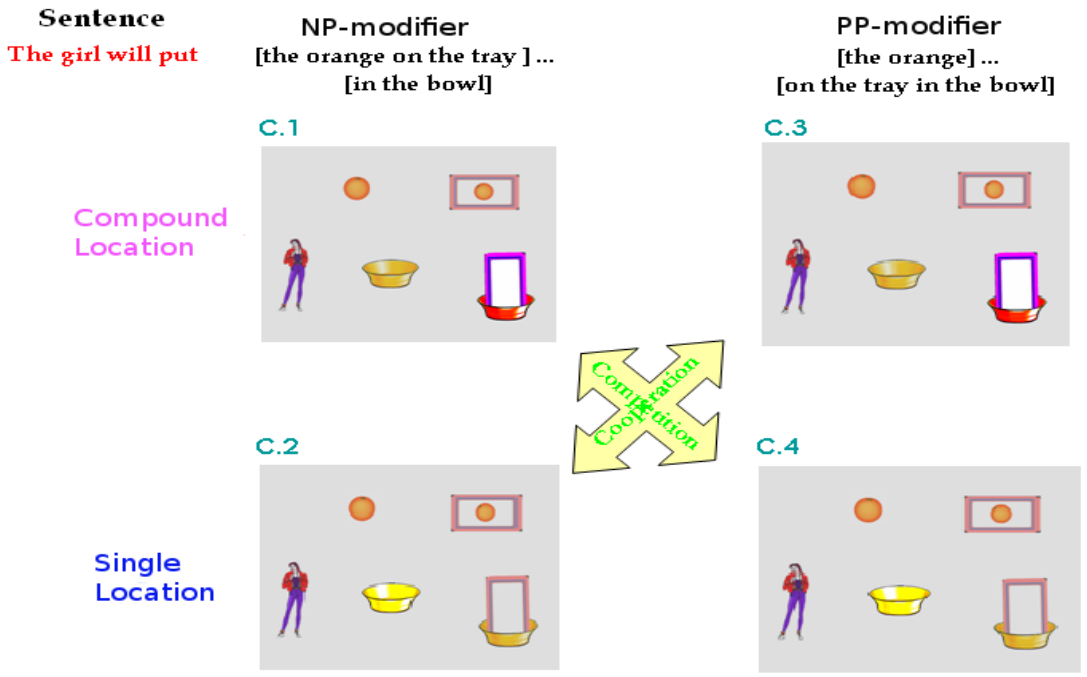


Figure 3.15: Experimental setting, four condition: a) Competition (Single-Location/NP-modifier, Compound-Location/PP-modifier), Cooperation (Single-Location/PP-modifier, Compound-Location/NP-modifier).

In order to have full interaction between saliency and intonational breaks (see Figure 3.15), we create conditions such that, they either point to the same resolving target object (Cooperation), or they point to different ones (Competition). An example of the Cooperative condition is (NP-modifier/Single-Location), where the intonational break NP-modifier put in focus the phrase *in the bowl* by enclosing it within a single prosodic phrase, while the salient object in the visual context is the Single-Location BOWL. An example of the Competitive condition is (PP-modifier/Single-Location), where the phrase in intonational focus is *on the tray in the bowl*, but the salient object is the Single-Location BOWL, rather than TRAY IN BOWL.

3.5.1.1 Participants

Thirty two participants from the same population were each paid five pounds for taking part in the experiment.

3.5.1.2 Materials, design, procedure and analysis

The materials have been imported from previous experiments. From experiment 1, we used the visual stimuli where saliency was manipulated; from experiment 2, instead, we used the linguistic stimuli where the manipulation was on the intonational breaks. From number of referents, we retain the condition where 2 Referents are depicted. Regarding the previewing time condition, that was manipulated in experiment 1 (Long and Short), we use a Long preview of 1000ms, which allows us to compare results with the previous experiments. The analysis is done following the methods previously adopted. Plots of fixation (empirical logit) have been used during the descriptive phase of analysis, whereas for the inferential analysis, we use LME models. The predictors for the models are: *Intonational-Breaks*, *Saliency* and *Time*; as random effects we have *Subjects* and *Trials*. Our best model is selected using a step-wise forward model approach.

3.5.2 Results

We consider the same ROI of previous experiments: NP direct object *the orange*,1PP *on the tray* and 2PP *in the bowl*. As in previous experiments, we show plots of observed fixation data and describe it together in the context of the LME coefficients found significant after model selection.

3.5.2.1 ROI:NP direct object *the orange*

In experiment 1, we have seen anticipatory effects triggered by saliency at the beginning and during ROI:NP direct object.

In Figure 3.16, we show the trend of looks, expressed in empirical logit, on target object BOWL starting from onset of ROI direct object *the orange* along 800 ms temporal window. Confirming experiment 1, we observe a strong anticipatory effect of saliency on Single-Location, which now is nevertheless weakened by the interaction with NP-modifier break (see Table 3.5).

Saliency is inferentially used to predict which visual objects are going to be mentioned. However, during an NP-modifier break, attention is shifted on both SINGLE

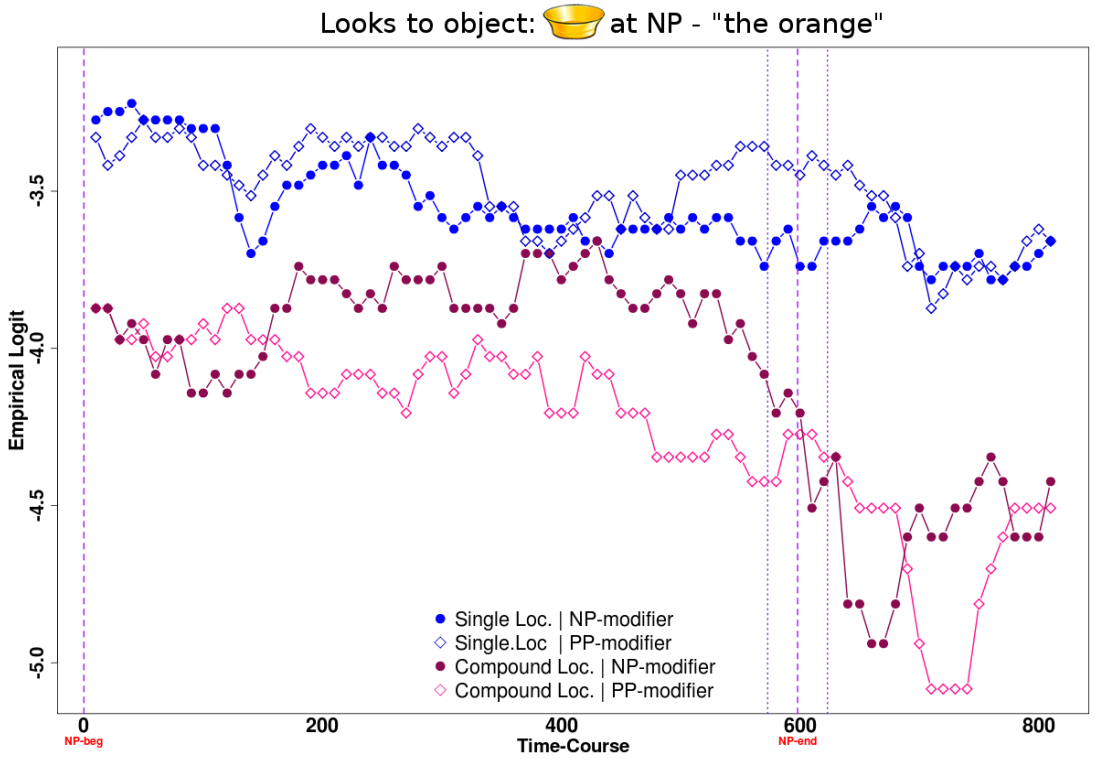


Figure 3.16: Experiment 3. Empirical logit of fixations on target object BOWL at ROI:NP *the orange* across conditions.

ORANGE and ORANGE ON TRAY compared to PP-modifier. The more referential competition reduces the anticipatory impact of saliency.

On the contrary, when we look at Figure 3.17, even if the plot suggests an early effect of saliency on object Compound-Location, this intuition is not confirmed inferentially, where the coefficient doesn't reach significance. Instead, we observe a significant negative interaction of Compound-Location with Time, where the looks on Compound-object decrease while the direct object *the orange* unfolds, shifting visual attention. Moreover, we observe a negative interaction between PP-modifier and time on looks to Compound-Location. The early effect of saliency is weakened by the prosodic information, which at this region, for PP-modifier points toward the SINGLE ORANGE. Furthermore, the semantic and structural plausibility of target object might have also played a role; a single-location can be more easily the goal location for a direct object, compared to a compound-location.

Table 3.5: Experiment 3. Linguistic **ROI** *NP direct object: the orange*; on the two Visual ROI: SINGLE-LOCATION and COMPOUND-LOCATION. Predicted LME coefficient estimates of predictors. Explanatory variables are centered around the mean. Single Location (0.5), Compound Location (-0.5); NP-modifier (-0.5), PP-modifier (0.5).

	SINGLE-LOCATION	
Predictor	Coefficient	*p*
Intercept	-4.8772	0.0001
Saliency	0.1149	0.0001
Time	-0.004	0.0001
Prosody	-0.011	0.5
Saliency:Prosody	0.0399	0.0001
	COMPOUND-LOCATION	
Predictor	Coefficient	*p*
Intercept	-4.7402	0.001
Saliency	-0.0748	0.1
Time	-0.0054	0.1
Prosody	-0.0367	0.4
Saliency:Time	0.011	0.0001
Prosody:Time	-0.0046	0.007

3.5.2.2 ROI 1PP: modifier/location *on the tray*

In this ROI, we replicate the intonational effects seen in Experiment 2.

In Figure 3.18 we show looks on TRAY IN BOWL. We find a main effect of PP-modifier which strengthens over time (refer to Table 3.6 for coefficients). The intonational break highlights the visual object enclosed within the prosodic phrase while it unfolds over time.

In line with results of Experiment 1, we observe effects of Intonational-break also on ORANGE ON TRAY.

In Figure 3.19, we observe initially higher looks to target object for PP-modifier, but the trend changes over time. We find, in fact, a positive interaction between NP-modifier and time. Again, the effect of intonational break is connected to the temporal dimension of prosodic unfolding.

Both visual objects ORANGE ON TRAY and TRAY IN BOWL are referred by the linguistic ROI *on the tray*, thus looks are expected to increase. However, as seen

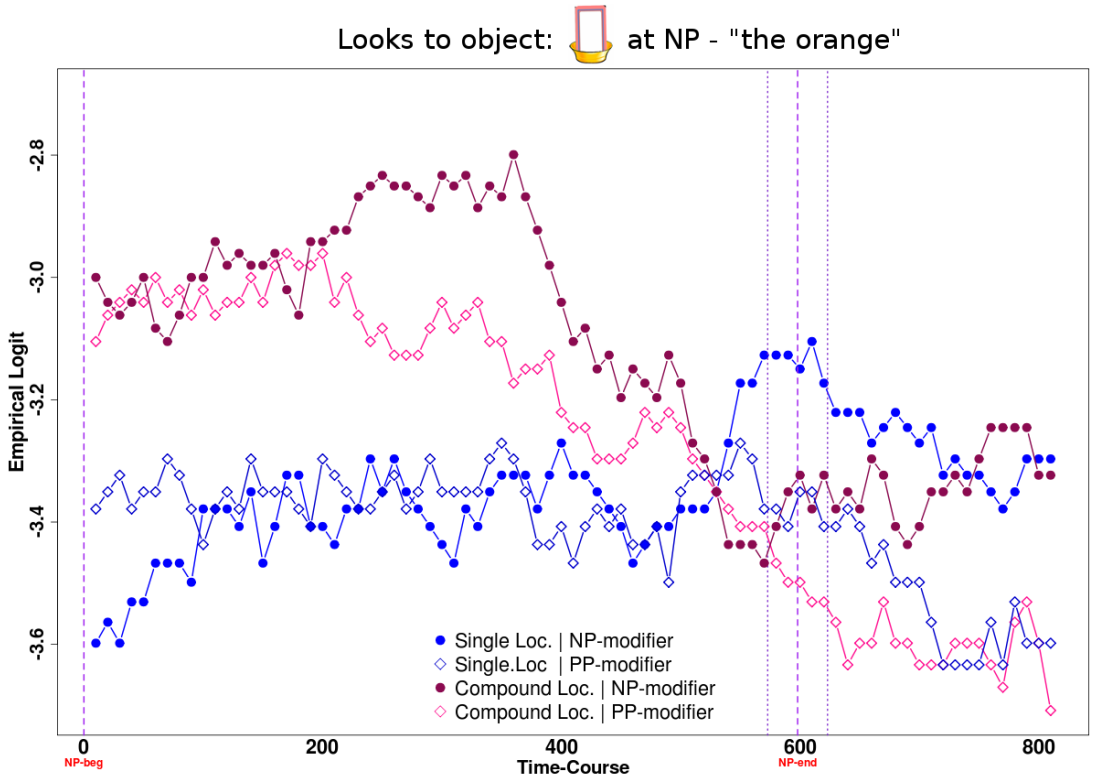


Figure 3.17: Experiment 3. Empirical logit of fixations on target object TRAY IN BOWL at ROI:NP *the orange*

at ROI:NP (*the orange*), we might find anticipatory looks launched to other target objects, e.g. BOWL, which are triggered by predictive processes such as saliency. We test whether anticipatory looks are still found on BOWL.

In Figure 3.20 we observe a main effect of saliency Single-Location over the whole time course, which positively interacts with break NP-modifier (refer to Table 3.6). Saliency on a Single-Location plays still a predictive role. Participants rely on low-level visual information to anticipate arguments of the sentence. Crucially, however, the effect of saliency goes together with semantic and structural plausibility of the event undergoing integration. An EMPTY BOWL could still serve the role of final goal location, whereas for the compound object TRAY IN BOWL, it would be unlikely. This intuition is confirmed by the absence of predictive saliency effects on TRAY IN BOWL at this linguistic ROI. Moreover, we find evidence of cross-modal cooperation. If there is an NP-modifier break, which suggests a final reading where the 2PP *in the bowl* is goal

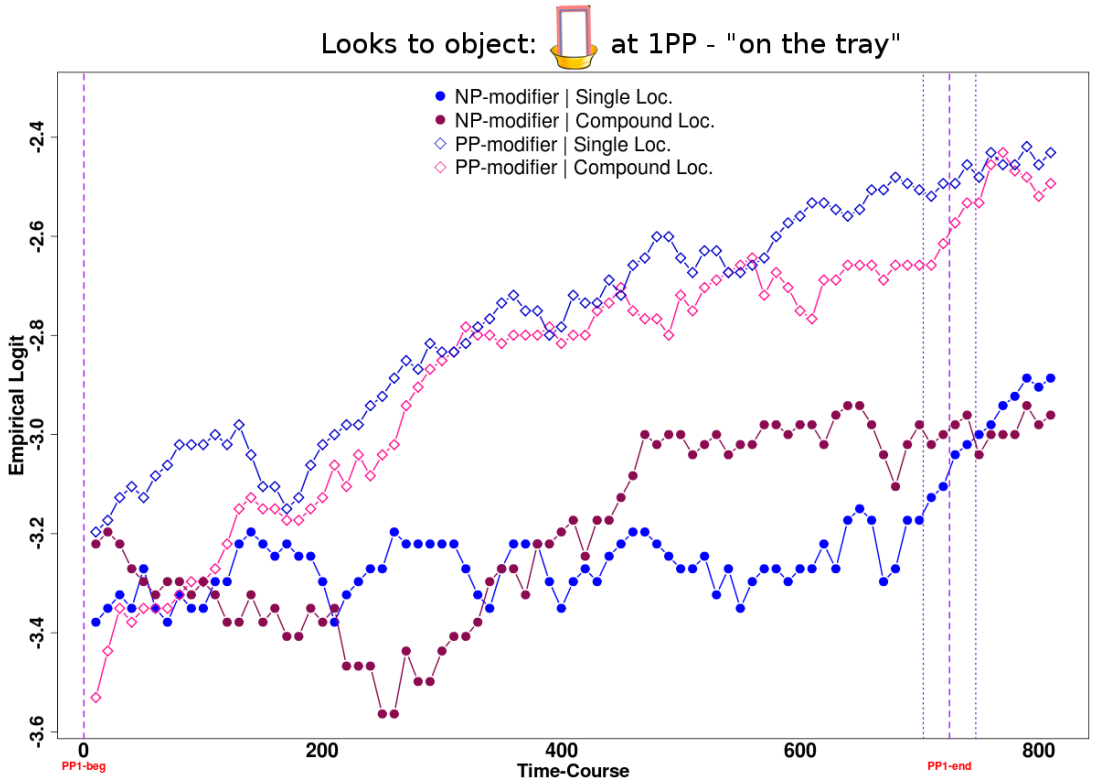


Figure 3.18: Experiment 3. Empirical logit of fixations on target object TRAY IN BOWL at ROI:1PP *on the tray*

location for *orange on the tray*, and the object BOWL is salient (Single-Location), we observe more looks. The interpretation of an event can be strengthen by the cooperative convergence of cross-modal information (Evans & Treisman, 2010).

3.5.2.3 ROI:2PP *in the bowl*

On this last linguistic ROI, we have observed only the effects of intonational break in experiment 2 limited to target object BOWL, with the NP-modifier break yielding more looks. We didn't find an effect of Saliency. However, saliency might show effects when combined with intonational information. Confirming previous experiments, we find effects only on BOWL, thus we report only results for it.

In Figure 3.21 we observe a significantly higher trend of looks when the target object is salient (Single-Location) compared to the other condition of saliency. This

Table 3.6: Experiment 3. Linguistic **ROI** *1PP: on the tray*; on the three Visual ROI: SINGLE-LOCATION, COMPOUND-LOCATION and ORANGE ON TRAY. Predicted LME coefficient estimates of predictors. Explanatory variables are centered around the mean. Single Location (0.5), Compound Location (-0.5); NP-modifier (-0.5), PP-modifier (0.5).

	SINGLE-LOCATION	
Predictor	Coefficient	*p*
Intercept	-3.6261	0.0001
Saliency	0.0758	0.02
Time	0.0035	0.1
Prosody	0.041	0.2
Saliency:Prosody	-0.0384	0.02
	COMPOUND-LOCATION	
Predictor	Coefficient	*p*
Intercept	-4.6331	0.0001
Prosody	0.1668	0.0001
Time	0.0133	0.0001
Saliency	0.004	0.9
Prosody:Time	0.0126	0.0001
	ORANGE ON TRAY	
Predictor	Coefficient	*p*
Intercept	-4.27	0.0001
Prosody	-0.098	0.1
Time	0.005	0.2
Saliency	0.0008	0.9
Prosody:Time	-0.0119	0.0001

positive trend of looks is strengthened by an interaction with NP-modifier break and increase over time. Similar to experiment 2, the intonational break confirms its role of highlighting the visual referent to be looked at, but this time, we also find it cooperating with saliency information. When both visual and linguistic information point to the same resolving object, their integration is strengthened.

3.5.3 Discussion

In experiment 1, we observed that at ROI:NP direct object, *the orange*, saliency is utilized to predict upcoming linguistic information. In experiment 2, as expected from

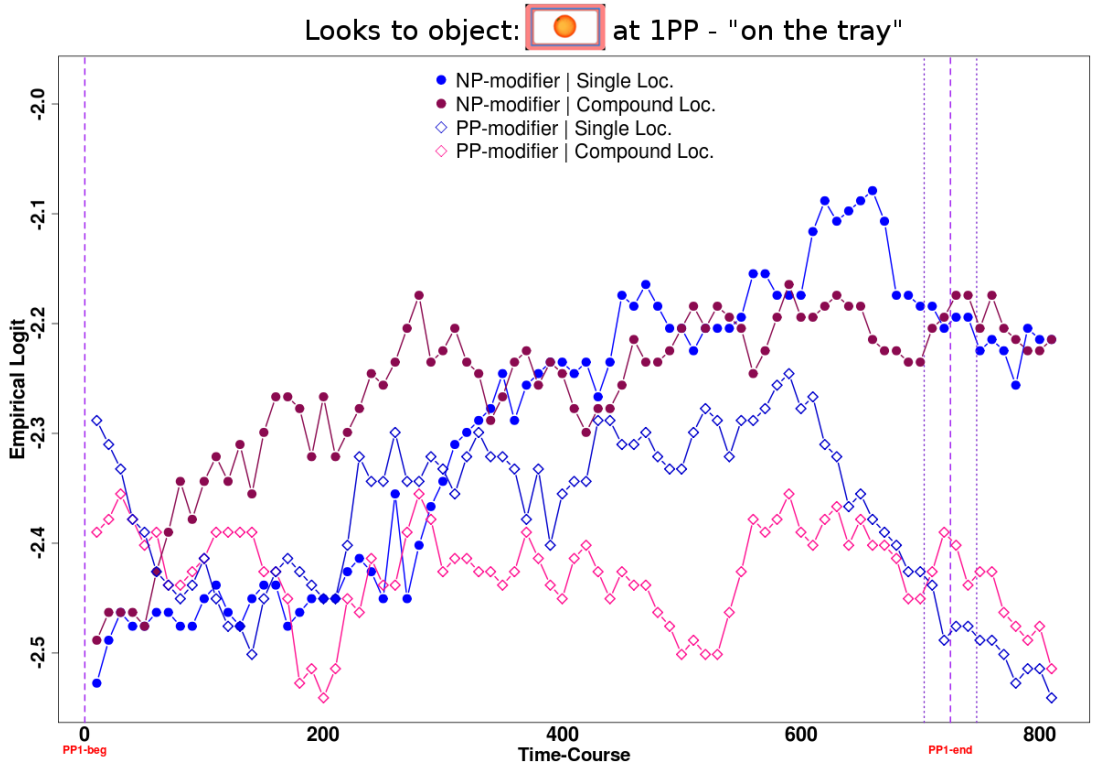


Figure 3.19: Experiment 3. Empirical logit of fixations on target object ORANGE ON TRAY at ROI:1PP *on the tray*

Table 3.7: Experiment 3. Linguistic **ROI** *2PP in the bowl*; on the Visual ROI SINGLE-LOCATION. Predicted LME coefficient estimates of predictors. Explanatory variables are centered around the mean. Single Location (0.5), Compound Location (-0.5); NP-modifier (-0.5), PP-modifier (0.5).

	SINGLE-LOCATION	
Predictor	Coefficient	*p*
Intercept	-4.8772	0.0001
Saliency	0.08	0.05
Time	0.0065	0.008
Prosody	-0.0606	0.1
Saliency:Prosody	-0.0720	0.0009
Prosody:Time	-0.0072	0.0001
Saliency:Time	0.0069	0.0002
Saliency:Prosody:Time	-0.0087	0.01

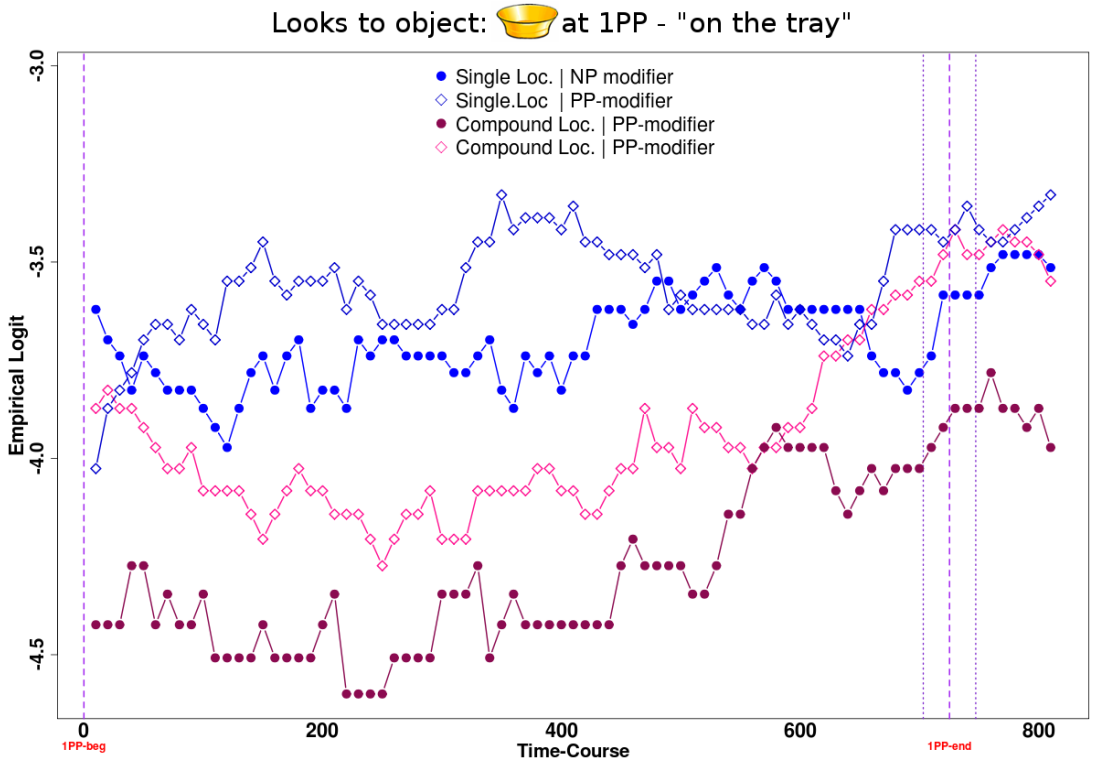


Figure 3.20: Experiment 3. Empirical logit of fixations on target object BOWL at ROI:1PP *on the tray*

previous studies (e.g. Snedeker & Trueswell 2003), we observed that intonational breaks give prominence to the referent enclosed within the prosodic phrase, hence triggering more looks to the corresponding visual object. In experiment 3, we tested the interaction (competition/cooperation) between saliency and intonational breaks. We find similar results as those independently observed in previous experiments. Saliency has anticipatory effects probably generated at verb-site, whereby a salient object is expected to appear as argument of the sentence. Intonational breaks modulate the mapping between visual and linguistic referents by giving prominence to the referent enclosed within the prosodic phrase. However, we also find instances of interaction, mostly cooperation; which were dependent on semantic and structural plausibility of the integrated event. In particular, we observed cooperation when saliency was on Single-Location and intonational break structure was NP-modifier. An NP-modifier break suggests a reading, where the *orange on the tray* is put in the goal location

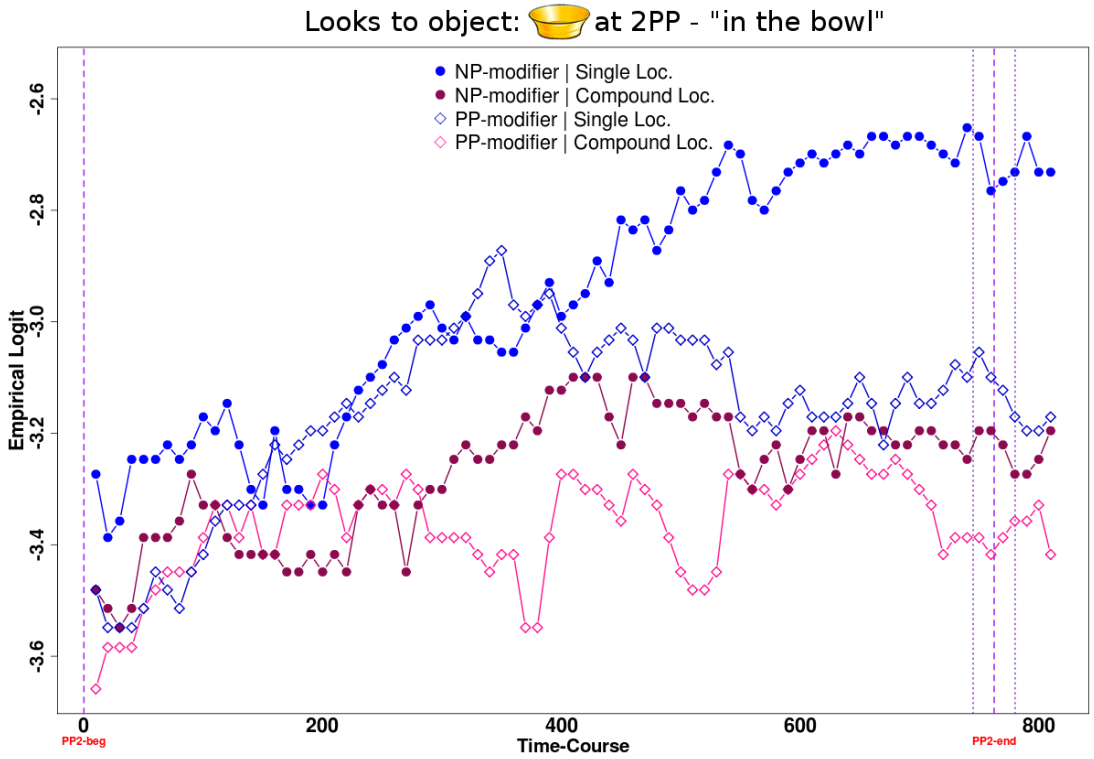


Figure 3.21: Experiment 3. Empirical logit of fixations on target object BOWL at ROI:2PP *in the bowl*

EMPTY BOWL. Thus, when saliency was on the Single-Location (EMPTY BOWL), there was cross-modal cooperation, which resulted into higher looks compared to the other conditions. Crucially, however, a similar effect was not found when cooperation was between saliency on Compound-Location and PP-modifier break. The problem relies on the visual complexity of the compound object, which indirectly also reflects an inherent linguistic complexity. A single object has a larger and more flexible set of events in which it can be imagined, whereas a compound object, beside being in some occasions semantically implausible, has a stricter range of final events. Since comprehension is incremental, for each new word processed, the set of possible continuations shrinks ruling out candidates which aren't plausible. Thus at position 1PP *on the tray*, BOWL can still appear encoded as goal location, whereas the compound object TRAY IN BOWL is a less plausible goal location. This intuition is confirmed by the only effect of competition found; where at ROI:NP (*the orange*) the effect of saliency

on Compound-Location was challenged by PP-break information, which at this ROI is giving focus to object ORANGE (see section 3.5.2.1 for details). The visual complexity of the compound-object taken together with the competing intonational break has neutralized the effect of saliency. Overall, we have found evidence of an highly interactive architecture of cognition, where both visual and linguistic information is utilized on-a-par, when needed during the task. Moreover, when visual and linguistic information is pointing to the same resolving object, their cross-modal cooperation strengthens the ongoing integration process.

3.6 General discussion

Sentence processing often occurs synchronously to other modalities, e.g. vision, and this raises questions about their cross-modal interaction, and the mechanisms underlying this integrated processing.

Previous work in psycholinguistics (e.g. Tanenhaus *et al.* 1995) has shown that visual responses are influenced by the integration between linguistic information processed and the identities of objects, i.e. the visual referents, forming the visual context. Within this approach, visual responses play the marginal role of signaling which contextual object is currently under linguistic processing. However, there are active visual mechanisms, e.g. saliency, which might also influence the allocation of visual attention during situated language processing.

The first goal of this chapter was to explore whether image-based visual information, i.e. saliency, is utilized during situated sentence understanding, and if yes, how. The second was to investigate which pattern of integration arises when visual and linguistic information are investigated in interaction.

In experiment 1, we tested the first issue in an eye-tracking language comprehension experiment, where participants listened to syntactically ambiguous sentences, while concurrently viewing a visual context, where saliency of objects has been manipulated (for details refer to section 3.3). We found that saliency is utilized during sentence processing, especially at beginning of direct object to predict upcoming linguistic information. This finding interestingly contrasts with research in visual cognition showing that saliency is active only during free-viewing tasks (Henderson *et al.*, 2009a); where, differently from a visual search task, there is no goal to be achieved by

the viewer. In a situated sentence comprehension task, we can identify two phases, an initial phase of free-viewing and a second phase of incremental sentence comprehension. During the first phase, the participant has no precise goals, but rather expectations[1] on the type of sentences that might be listened to. In the second phase, the goals are defined through the understanding of the sentence, which is incrementally built during linguistic unfolding. Obviously, at the beginning of the sentence, e.g. at verb site, the linguistic information processed is not sufficient to generate a full prediction of upcoming arguments, thus low-level visual information is utilized to fill in this gap.

This finding of interaction between saliency and sentence processing raises an intriguing question about the relation between visual and linguistic information during synchronous processing. Especially, we asked whether there is any preferential way of accessing visual or linguistic information, or instead they are used rather independently. In order to test this hypothesis, in experiment 2, we manipulate intonational break information, which can be considered low-level linguistic information[2], on the same task and material. We needed to observe the independent effect of intonational breaks before investigating it in interaction with saliency.

In line with previous research (e.g. Snedeker & Trueswell 2003), we find that the visual object, e.g. ORANGE, corresponding to a linguistic referent enclosed by the intonational breaks, e.g. PP-modifier, is looked at more, compared to other conditions of intonation (see section 3.4 for details). The intonational breaks are responsible for temporally organizing the mapping between linguistic and visual referents. Moreover, the plausibility of the event resulting from the integration of visual and linguistic information modulates the effect of intonational break. Thus, a PP-modifier break, which puts in focus the Compound-Object TRAY IN BOWL, has overall less effect than a NP-modifier break, which instead focuses more on Single-Object EMPTY BOWL. A compound-object is, in fact, has less possibility to be combined into a plausible event, than a single-object.

Finally in experiment 3, we test the interaction betwen saliency and intonational breaks by reusing material from the previous two experiments, and designing it such that saliency and intonational breaks either cooperate, i.e., both kinds of information point to the same target object, or compete, i.e. they point at different target objects.

[1] These expectations are probably reinforced during the course of the experiment

[2] It doesn't carry explicit semantic information.

We find similar results to experiment 1 and 2, when visual and linguistic information were tested independently. However, we also observe cases of cooperation, where the integrated contribution of both type of information resulted into higher looks to resolving object (for details refer to section 3.5). Moreover, again, we find that plausibility of the event undergoing integration plays a key role. In fact, effects of cooperation are found only when saliency is on a single-object (Single Location) and the intonational break puts it into focus (NP-modifier).

Overall in this chapter, we have shown that visual information must be taken into account when investigated concurrently with sentence processing; and that the interaction between visual and linguistic information doesn't present a preferential pattern, i.e. linguistic information overrides effects of visual information, but rather both types of information are utilized depending on the state of the task, e.g. at ROI:NP compared to ROI:1PP, during which cross-modal integration is observed.

3.7 Conclusions

An important message of studies on situated language processing is that on referents, visual and linguistic, we can investigate the interaction between sentence processing and visual attention. However, the classic experimental setup utilized presupposes a specific type of manipulated linguistic material, e.g. syntactically ambiguous sentences, which is contextualized in a rather simple visual context, e.g. object arrays, where visual responses are restricted to a finite and small number of disconnected objects. In real world scenarios, instead, sentences are generated or understood on the basis of a task being performed, e.g. scene description, and very rarely have structures of the type investigated in psycholinguistic research, e.g. PP-attachment ambiguity. Moreover, visual objects referred to by a sentence are usually contextualized within a naturalistic setting; in which they have a precise semantic function, e.g. a MUG is used for drinking, and co-occur with other objects of the scene, e.g. a MUG is usually found on a TABLE or a kitchen COUNTER.

In Chapter 4, we explore how sentences are generated from photo-realistic scenes, which visual and linguistic factors are involved, while beginning to unravel mechanisms of cross-modal referentiality.

Chapter 4

Object-Based Factors in Cross-Modal Referentiality during Situated Language Production

4.1 Introduction

In linguistics, a referent is the thing in the world that a word or phrase denotes or stands for (Saeed, 2008). In the larger context of cognition, a referent is a cognitive entity bridging together the multi-modal perception of a real-world counterpart with a linguistic identifier. During tasks demanding the synchronous interaction of different cognitive modalities, e.g. situated language processing, mechanisms of cross-modal referentiality are activated to coordinate such multi-modal processing.

When describing a scene, for example, visual attention retrieves referential information about objects, e.g. MUG, while sentence processing creates linguistic denotations e.g. *the mug* referring to their visual identity. It follows that the visual context constrains the interpretation of the linguistic material, and vice versa. Thus, if we are in a kitchen, and somebody asks us to *take a mug*, visual information about the scene, e.g. a MUG is usually on a COUNTER, is integrated with the linguistic information given, making synchronous processing more efficient and less ambiguous. Obviously, however, the complexity of integration is increased when the scene contains referential ambiguity. If there are many different MUGS on the COUNTER then more visual information, e.g. color, is needed to create an unambiguous referential identity RED

MUG. And this additional information increases complexity of linguistic encoding: *do you want **the red mug**?*.

In this chapter, we focus on the visual and linguistic factors implicated in the formation and maintenance of a shared referential interface, while exploring the pattern of visual responses emerging during referential integration, i.e. before and after a certain visual referent is linguistically mentioned. Our hypothesis is that semantic properties of visual objects, e.g. animacy, together with the density of visual information, e.g. clutter, influence linguistic encoding. Moreover, we expect these factors, together with referential ambiguity, to modulate the latency between fixating a visual entity with respect to its linguistic mention, i.e. *eye-voice span*.

Overall, by studying how referential scene information is visually inspected during linguistic encoding, we lay the foundations for a theoretical understanding of synchronous visual and linguistic processing, while gathering empirical data on which predictions can be tested.

4.2 Background

A realistic theory of the formation and maintenance of reference across modalities has to treat visual information on a par with linguistic information. Such a theory must explain how mechanisms known to operate independently in both the linguistic and the visual modality cooperate in the process of referent assignment. In the previous chapter, we have found that *saliency* influences the resolution of prepositional phrase (PP) attachment ambiguities in language comprehension. Saliency is used to predict which visual objects can be encoded as post-verbal arguments in a given sentence; therefore, specific mechanisms of visual attention actively interact during sentence processing.

However, it is important to note that low-level visual features such as saliency are not referential per-se; they are properties of image regions, not of objects (Henderson *et al.*, 2009a). It is therefore necessary to focus on top-down visual properties, which are object-based and found to be actively implicated during goal oriented tasks, e.g. search[1] (Castelhano *et al.*, 2009; Malcolm & Henderson, 2010; Nuthmann & Hender-

[1] Find cued target object in the scene.

son, 2010; Oliva *et al.*, 2003; Schmidt & Zelinsky, 2009). Categorical information about the cued target is combined with contextual scene information to actively guide visual attention allocation (Torralba *et al.*, 2006). An object-based approach of active visual perception strongly implicates processing of referential information; in fact, the objects of a scene are referents which can be linguistically mentioned. Thus, it seems likely that during sentence processing, situated in naturalistic scenes, visual and linguistic referentiality is established upon object-based information, and influenced by its factors.

In this chapter, we investigate the influence of object-based factors during situated language production in naturalistic scenes with referential ambiguity. We decided to move from situated language comprehension (Chapter 3) to production, in order to investigate more naturally the association between objects fixated and descriptions produced. During situated language comprehension, the sentences that participants have to listen to are chosen by the experimenter. In practice, this reduces the active contribution of visual attention during sentence processing; as it is mainly expected to respond to the linguistic material parsed. During situated language production, on the other hand, the referential information of objects in the scene has to be visually retrieved first, before being linguistically encoded. Therefore, by looking at production, we expect the visual factors of the scene and its objects to modulate the linguistic output generated, while making sentence processing more directly dependent on the visual context inspected. Moreover, the aim of investigating more realistically the relation between visual and linguistic processing is further supported by the use of a photo-realistic referentially ambiguous visual context, in place of the commonly adopted object-arrays. The use of naturalistic scenes give us more realistic visual responses; while referential ambiguity helps us to explore the different strategies of disambiguation used to resolve it.

To the best of our knowledge, there have been only few attempts to investigate language production concurrently with a visual context (Griffin & Bock, 2000; Qu & Chai, 2008).

In Griffin & Bock 2000, participants were eye tracked while they described pictures containing two actors (depicted alternatively as Agent or Patient of the event; see Figure 4.1 for example trial and results). The goal of the study was to investigate the relation between visual entities fixated and the sequential order in which they are

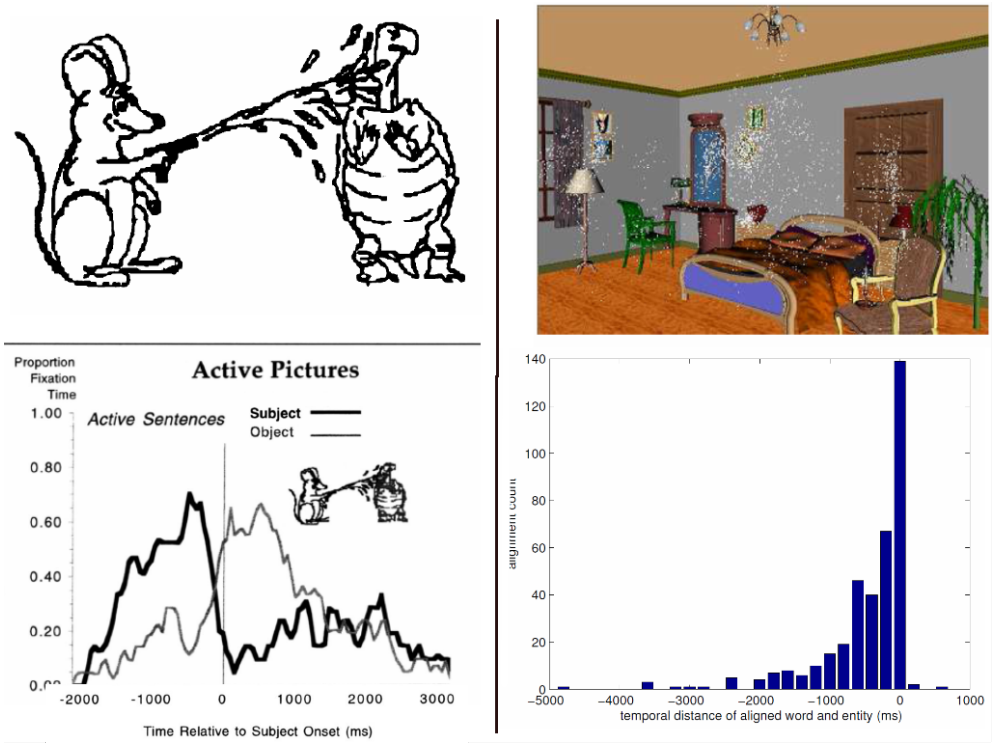


Figure 4.1: In the left panel, we show an example of b/w image (top-left) used in the description task by Griffin & Bock 2000; and proportion of fixation on the two objects (bottom-left), before and after the subject of the transitive action depicted, i.e. MOUSE is mentioned. In the right panel, we show the 3D rendered scene used for the dialogue task by Qu & Chai 2008 with raw fixation over-plotted (top-right); and trend of temporal alignment between gaze on object and linguistic mention (bottom-right).

produced. A key observation is the *eye-voice span*: a visual object is fixated around 900ms before it is named. This observation has been further confirmed in other studies where similar trends of eye-voice spans have been reported (Qu & Chai, 2008, 2010), see Figure 4.1 to visualize their results. However, both studies have shortcomings which might have exaggerated the eye-voice span effects. In Griffin & Bock 2000, the eye-voice span might be due to the simplicity of the visual material utilized, i.e. b/w drawings depicting only two visual referents. Beside the fact that there is a 50% chance to look at either referent, the visual information retrieved in support of sentence production can only be found on the visual referent undergoing mention, as there is no contextual scene information. This limitation is further exacerbated by the absence of referential ambiguity, which simplifies even more which visual referent has to be

looked at. Qu & Chai 2008 use a more complex 3D pseudo-scene [1], containing 28 contextually related objects, some of which are ambiguous, e.g. CHAIR. However, participants had to answer to automatically generated questions regarding the objects in the scene, e.g. *describe the left wall*. We believe that the nature of the task forced attention to be focused on the target of the question; thus triggering serial responses, i.e. look and name.

In this chapter, we investigate the referential effects of top-down (object-based) visual properties, during cued descriptions of photo-realistic, referentially ambiguous, scenes. Our hypothesis is that the selection of referents and the type of structural encoding (e.g. active vs passive) depends on both the visual information of the scene (i.e. *clutter*), quantifiable as density (Rosenholtz *et al.*, 2007), and the semantic properties of the visual object to be described, the most general being: animate vs inanimate[2] (Branigan *et al.*, 2008). Thus, we expect sentence encoding to be dependent on the semantics of the cue and scene information and these effects are also expected to emerge on the corresponding eye-movements records.

In Experiment 4, we get a first glimpse of visually grounded descriptions by designing a web experiment where participants are asked to write descriptions of photo-realistic scenes, which differ by the density of visual information and number of animate actors, after being prompted with a cue word, either referring to an animate or inanimate object depicted. We investigate the reaction times of visual apprehension and sentence encoding, while exploring in detail the structure of the generated sentences.

Contrary to findings in visual search studies, where clutter had a negative impact on search performance, we expect clutter to impact positively on sentence production. In fact, the more visual information there is, the more linguistic encodings are possible, thus boosting both visual retrieval and sentence encoding. In line with language production literature, we expect animate referents to facilitate encoding compared to inanimate referents. Moreover, we assume this effect to be cumulative; thus the more animate referents are depicted, the more conceptual material is available to source sentence encoding.

[1] The same scene is used for all subjects.

[2] More details about the experimental factors can be found in the next section.

In Experiment 5, we move one step forward by investigating how eye-movement patterns are linked to the type of sentences produced. The experimental design is similar, the main difference being that participants are now eye-tracked, and thus descriptions are spoken rather than written. Here, we focus on the phase of referential integration, i.e. the time around the mention of the cued target. In particular, we test the eye-voice span hypothesis by looking at the frequency of fixations on the referenced visual target before its mention. A naturalistic referentially ambiguous setting is expected to modulate the eye-voice span: in such a scene there are more objects that can be looked at. Furthermore, the ambiguous visual referents are expected to compete on visual attentional resources (see Chapter 3). Both factors are expected to modulate the serial gaze-to-name eye-voice span relation. Moreover, we explore the impact of our experimental factors (animacy and clutter) on visual attention during mention of the referent (before and after). In line with visual cognition research, animate referents are expected to receive more attention than inanimate referents, especially when a scene has a low density of visual information. During sentence production, more visual information implies more referential information to be used during encoding, thus in a low density scene, most of referential information relies upon the animate referent.

The main goal of this chapter is to investigate cross-modal referential information processing, while providing an empirical ground to explore the underlying mechanisms of synchronous processing.

4.3 Experiment 4: Clutter and animacy on scene description

In a web-experiment we investigate the impact of visual referential information during description of photo-realistic scenes. In contrast to previous studies (Griffin & Bock, 2000), the language generation task is situated in photo-realistic scenes with referential ambiguity, i.e., two depicted CLIPBOARDS correspond to the cued word *clipboard*. Moreover, global visual information, i.e., clutter, semantic properties of the cued target, i.e., animacy, and of the scene, i.e., the number of animate actors, are manipulated. Our hypothesis is that both semantics and the density of visual referential information available directly impact timing and strategies of sentence encoding. Especially,

the description of animate targets is expected to be easier than of inanimate ones, in particular when the scene has richer visual information density, i.e. more referential information to be used during encoding.

Before going into the technical details of the experimental design and the results obtained, we briefly discuss the implications of our visual and linguistic manipulations and their role for referentiality and sentence encoding.

Clutter A way to define referentiality in vision is to look for a global measure of visual information. Measures of visual information in vision have always been difficult to derive and have often been connected to the notion of *set size*. The bigger the set of visual objects displayed during a visual search task, the slower RT to respond. Hence, the number of countable visual objects was assumed to give a direct measure of visual information (Wolfe, 1998). However, this notion of visual information has recently been criticized (Rosenholtz *et al.*, 2007) especially when correlated to naturalistic scenes where it becomes extremely difficult, if not impossible, to define and count each single object composing the scene[1]. The alternative notion of visual information proposed is *clutter* (Rosenholtz *et al.*, 2007), see Figure 4.2 to visualize clutter.

Clutter is defined as the state (organization, representation) of visual information in which visual search performances start to degrade, and it has been statistically modeled and quantified using the feature congestion method (for further details refer to Rosenholtz *et al.* (2005)). In our study, we use the Clutter measure to investigate the correlation between amount of visual information and sentence encoding focusing on the retrieval of referential information. Moreover, we adopt clutter in place of the previously used saliency (Chapter 3), because it takes into account the edge of objects, which is an important feature to recognize and relevant to quantify their presence in the scene. Contrary to findings for visual search tasks (Henderson *et al.*, 2009a), we expect clutter to facilitate sentence processing: more referential information can be used during linguistic encoding.

[1]The reasoning holds also when we think at the relation between linguistic referents and visual information. For the same object CUP, we can use different linguistic labels highlighting diverse aspects of the same entity. We can use, for example, *the cup* referring to the whole object, or *the handle* if we refer to a detail of the object.

Figure 4.2: On the left, we show an example of a naturalistic scene used by Henderson *et al.* 2009b. On the right, the same scene is displayed after the feature congestion algorithm is applied to measure its visual information. The red color reflects the density.

Number of actors and animacy A crucial classification between types of world-entities is given by the feature of *Animacy*. The importance of Animacy has been long discussed especially in connection to the assignment of grammatical functions and word order distribution (McDonald *et al.*, 1993). Animate entities are conceptually more accessible than those inanimate (Levelt *et al.*, 1999) and are therefore privileged during syntactic processes of production. This privilege reflects into an Animate entity being encoded with the grammatical function of 'subject' whereas Inanimate occurs mostly with the function of 'object'. Moreover, recent findings claim that animacy influences also the word order of sentence structure (Branigan *et al.*, 2008). The effect of Animacy has also been observed in a preferential-looking task[1] (Fletcher-Watson *et al.*, 2008). A scene containing an animate visual referent is preferred to the one without, with faces attracting the majority of fixations. Animate referents carry conceptual information, which is crucial both during linguistic and visual processing.

In this study we took a broader view of the feature animacy; animacy is not only a linguistic notion, but it is also visually encoded. We therefore manipulated animacy in both the linguistic and the visual modality. Visually, we introduce different degrees

[1]Participants were presented with two identical scenes: one with an animate referent, the other without.

of animacy by changing the number of actors depicted in the scene. Linguistically, we either gave an animate or an inanimate noun as the cue for sentence production.

In line with language production studies (Branigan *et al.*, 2008), animate entities are expected to boost a larger activation of conceptual structures, thus reducing processing cost, i.e. faster reaction time, compared to inanimate entities.

4.3.1 Design

In this experiment, participants had to describe a naturalistic scene, after being prompted by a single word (the description cue). As dependent variables we recorded Looking Time, i.e. the time that elapsed before the onset of the response, Description Time, i.e., the time taken to complete the response, and we also investigated the syntactic structure of the response produced. The design of the experiment manipulated both visual and linguistic referential information. We varied the total amount of visual information present in the scene in the factor Clutter (Minimal vs. Cluttered). We also manipulated the number of animate objects present in the scene in the factor Actors (One vs. Two). On the linguistic side, we varied the prompt given to participants for their description in the factor Cue, which could refer either to an animate or an inanimate object in the scene (Animate vs. Inanimate). The scenes were designed such that they always contained at least one animate object and two identical inanimate objects, so as to introduce systematic visual referential ambiguity. As an example, see Figure 4.3, where the clipboard is the ambiguous inanimate object. Note that the animate objects are referentially unambiguous, even in the Two-Actors condition (man and woman in the example stimulus).

The null hypothesis for this experiment is that visual and linguistic factors do not interact in language processing. This would mean that Clutter and Actors should only influence Looking Time in a way that is compatible with behavior in standard visual search tasks: we expect longer Looking Time in the Cluttered condition, as more objects have to be searched, and longer Looking time also in the Two-Actors condition, which contains an additional object. Our experimental hypothesis is that visual information has an impact on language production, which means that we expect an interaction between the visual factors Clutter and Actor and the linguistic factor Cue (in addition to effects that may be caused by standard visual search processes).

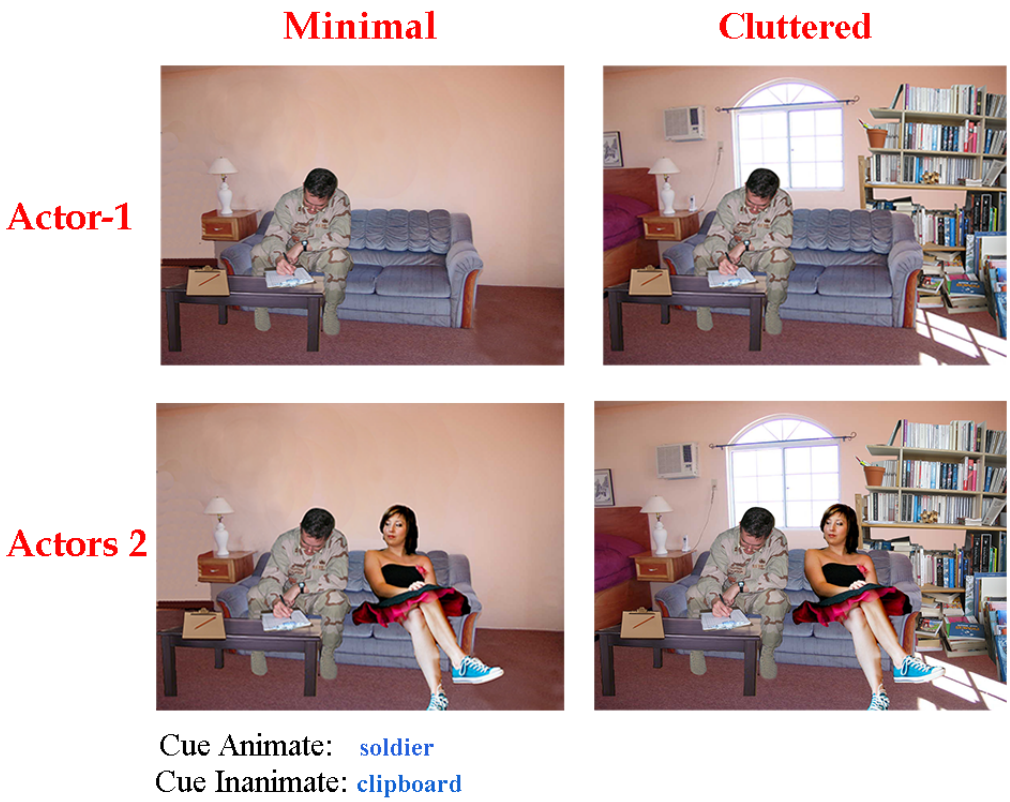


Figure 4.3: Example of the experimental trial. Four visual conditions and linguistic cues.

4.3.2 Method

The experimental design crossed three factors, each with two levels. The two visual factors were number of Actors in the scene (One or Two) and the degree of visual Clutter (Minimal or Cluttered). The linguistic factor was the Cue given to the participants to prompt their sentence production (Animate or Inanimate).

As stimuli, we created a set of 24 photo-realistic scenes using Photoshop by cutting and pasting visual objects from a set of pre-existing photographs. Differences in luminosity, contrast and color balance between the different photographs were adjusted through an accurate use of layers, luminosity masks and color balancing. In order to (1) control for the semantic variability across visual scenes and (2) ground language production in a restricted semantic domain, all pictures were created using six different interior environments: bathroom, bedroom, dining room, entrance, kitchen, and office.

Each interior was represented by four different scenes. For each scene, we created four variants manipulating Clutter and Actors, as illustrated in Figure 4.3. The scenes were designed such that the inanimate object was referentially ambiguous, i.e. there were two instances of it in the picture, while the animate one was unambiguous, even in the Two-Actors condition.

In the experiment, participants were first presented with a set of instructions explaining the task and a set of examples. After a practice phase, they saw one visual stimulus at a time, together with the linguistic cue. They were instructed to provide a written description of the stimulus using the cue. The total of 192 different items were distributed over four lists using a Latin square design. Each subject saw one of the lists, i.e., 48 stimuli in total (each of the 24 scenes was presented twice: once with an animate cue, and in another occasion with an inanimate cue). The stimuli were randomized for each participant, and presented without fillers. The experiment took about 15 minutes in total.

The experiment was realized using the WebExp software package for conducting psychological experiments over the web. WebExp is able to measure reaction times with accuracy comparable to that of lab-based experiments, as shown by (Keller *et al.*, 2009) for self-paced reading data.

Participation was open to both native and non-native speakers of English (this was included as a factor in the analysis). The sample included 32 participants, including 16 native speakers and 16 non-native speakers.

4.3.3 Results and Discussion

We analyzed two response time measures. The first one is Looking Time, i.e., the time participants spent scanning the image before starting to type. It is calculated from the onset of the trial until participants pressed for the first key on the keyboard. The second response time measure, Description Time, is the time participants took to type their response. It is calculated from the first key press until Enter is hit to move on to the next trial.

We also analyzed the syntactic patterns in the responses produced by participants. For this, we tagged each sentence produced using an automatic part-of-speech tagger, viz. (Ratnaparkhi, 1996) maximum entropy tagger, which performs with an accuracy

of 96.6%. The tagger uses the Penn Treebank tagset to assign syntactic categories to words. We collapsed the various tags for nouns in the tagset (e.g., NNS, NNP) and verbs (e.g., VBD, VBN) to two general categories (NN, VB). For each sentence, we recorded the frequency of these two categories, as well as the occurrence of existential *there* and clause coordinator *and*. We also identified and counted the number of passive constructions (for this the full tag set was used, which marks passive verb morphology).

The statistical analyses were carried out using linear mixed-effect models (Jaeger, 2008) to determine the effect of the categorical predictor variables on both reaction times and syntactic frequency. We chose mixed models for their ability to capture both fixed and random effects (Baayen *et al.*, 2008). We included the following predictors in our analysis: Actors (One or Two), Clutter (Minimal, Cluttered), Cue (Animate, Inanimate) and Language (Native, NonNative). The mixed models were built and evaluated following a forward step-wise procedure, where nested models are evaluated on the basis of the log-likelihood fit improvement (see Chapter 2 for details).

Reaction Times Table 4.1 presents the coefficients and p-values of the mixed model for Looking Time. The model intercept represents the response time in the baseline condition in milliseconds, and coefficients indicate the effect a given predictor has on Looking Time (again in milliseconds). We find significantly shorter Looking Time when Cue is animate compared to when inanimate. This can be explained in terms of visual search behavior, as our stimuli contain more inanimate than animate cues, thus making it easier to discriminate animate objects, leading to reduced search time. In addition, the cued animate object was always unambiguous, while the cued inanimate object was always present twice in the scene, creating referential ambiguity, and thus increasing visual search time. There may also be an explanation in linguistic terms: As mentioned above, animate entities are conceptually more accessible than inanimate ones, with gives them a privileged status during syntactic encoding.

We find that participants are faster to scan the pictures when the scene has a minimal density compared to when is cluttered; nevertheless the effect doesn't reach significance. This finding challenges results observed during visual search tasks, where the more clutter, the more difficult was target identification (Henderson *et al.*, 2009b). In a description task, once the object is identified, visual information has to be retrieved according to the demands of sentence encoding. The looking time of a description task

Table 4.1: LME coefficient estimates of Looking Time and Description Time. The centered predictors are *Clutter* (Minimal, 0.5; Cluttered, -0.5), Cue (Animate, 0.5; Inanimate, -0.5), Language (Native, -0.5, NonNative, 0.5), Actors (One, -0.5; Two, 0.5).

	Looking Time	
Predictor	Coefficient	*p*
Intercept	3666.3	0.0001
Cue	-1036.1	0.01
Language	1123.7	0.03
Actors	239.7	0.06
Clutter	-200.5	0.1
Clutter:Cue	591.2	0.01
Actors:Cue	-470.5	0.05
	Description Time	
Predictor	Coefficient	*p*
Intercept	12461	0.0001
Cue	-1777.2	0.0001
Actors	993.1	0.001
Language	1970.9	0.05

does not just indicate how fast participants are to locate the object, but it also informs about the time of visual retrieval prior to sentence encoding.

We also found that in the condition Language-NonNative, participants take longer to scan the picture. This can be explained by the fact that non-native speakers presumably take longer to decode the cue and to plan their utterance.

Turning to the interactions, we found that Clutter significantly interacts with Cue: participants look longer to respond to animate prompts in the minimal clutter condition. Confirming our explanation about the absence of a main effect of Clutter, this interaction suggests that visual attention is not performing a search behavior. A description of an animate object embedded in a minimal scene is a condition with the fewest competing objects to consider; visual search should therefore be particularly fast, and the interaction should be absent or have a negative coefficient. The fact that we find a positive interaction indicates that a linguistic process is at work. In a visual scene with few objects it is more difficult to retrieve enough information about actions that a potential actor can perform. Thus, participants spend more time scanning the

scene and planning their utterance before sentence encoding starts.

There is also a significant negative interaction of Actors and Cue; Looking Time is reduced when two actors are depicted and the description task is cued with an animate referent. Again, this cannot be explained purely in visual terms; the presence of two actors cued by the animate cue should lead to longer search times, as two objects need to be considered instead of one. Instead, we find a negative coefficient for this interaction. Presumably, the conceptual accessibility of animacy is a cumulative property. The more animate entities the scene contains, the more conceptual structures are activated. The step of selecting a conceptual structure to encode is thus facilitated by the larger set of possibilities. Moreover, the unambiguous visual reference of animate objects may boost the selection of those conceptual structures that are related to the actor cued, decreasing looking time. This interpretation is supported also by our syntactic analysis (see next section) in which two actors and animate cue positively correlate with the use of nouns and verbs. Participants produce longer sentence structures, often encoding both Actors.

Table 4.1 also presents the mixed model coefficients for Description Time. The results overlap with those for Looking Time. For condition Cue-Animate, participants were faster to generate a sentence compared to Cue-Inanimate. As for Looking Time, this result can be explained by the fact that animate entities are more accessible in language production, and that visual search is faster, as there is only at most one other animate object in the scene. We also find significantly increased Description Time when two actors are depicted. An inspection of the responses (see below) shows that participants tend to encode both actors in their descriptions of the scene, which explains why encoding takes longer in these conditions, compared to the Actor-One condition, in which only one actor is encoded. Again, non-native participants show a longer response time than native ones, presumably because sentence production is slower in non-native speakers.

Syntactic Categories Table 4.2 present the results for the syntactic analysis of the picture descriptions generated by the participants. We fitted separate mixed models to predict the number of nouns and the number of verbs included in the responses. The intercept represents the noun or verb frequency in the baseline condition, and the

Table 4.2: LME coefficient estimates of Noun and Verb. The centered predictors are *Clutter* (Minimal, 0.5; Cluttered, -0.5), Cue (Animate, 0.5; Inanimate, -0.5), Language (Native, -0.5, NonNative, 0.5), Actors (One, -0.5; Two, 0.5).

	Noun	
Predictor	Coefficient	*p*
Intercept	2.2676	0.0001
Cue	-0.217	0.01
Actors	0.1461	0.01
Actors:Cue	0.2272	0.02
	Verb	
Predictor	Coefficient	*p*
Intercept	1.821	0.0001
Cue	0.2307	0.002
Actors	0.1063	0.03
Clutter	0.0824	0.1
Clutter:Cue	-0.2267	0.005
Actors:Cue	0.1628	0.03

coefficients indicate how this frequency increases or decreases under the influence of the relevant predictor.

The results indicate that significantly fewer nouns are produced when participants are cued with an animate referent. This condition was visually unambiguous, and thus required less elaborate descriptions compared to the Cue-Inanimate condition, for which participants generated longer sentences in order to unambiguously pick out one of the two visual referents available in this condition. Moreover, the competition between the two visual objects for the inanimate cue was often resolved by encoding both visual referents within the same sentence structure. An example of a sentence produced in this condition is *The mug is beside the man, another is on top of the files, both mugs have pencils in them*. Except for the referring expression itself, all nouns are used in combination with spatial prepositions to unambiguously differentiate each visual referent.

When two actors are depicted, and especially if the cue is animate, participants produced significantly more nouns. This correlates with the shorter Looking Times found for the same interaction. Participants often encoded referentially both visual actors

within the same sentence structure. An example is *A man stands behind a counter in a hotel while a customer writes on a piece of paper*. Even though the cue given (here, *the man*) refers only to one actor and is visually unambiguous, the participant encoded also the second actor.

Turning now to the analysis of the number of verbs produced, we again find a significant effect of Cue-Animate, but with a positive coefficient, which means that participants generated more verbs than in the Cue-Inanimate condition. This underlines the connection between the feature Animacy and the semantics of verbs. As verbs encode actions, they are less likely to occur in descriptions of inanimate entities. The latter tend to activate verbs that describe static, mostly spatial, relations like *lie* or *place*, whereas animate entities can be related to a broader range of events, both static and dynamic, resulting in a wider range of verbs generated.

An interaction between Actor-Two and Cue-Animate is also present, which is consistent with the main effect of Cue-Animate. The more animate entities are presented in the visual scene, the more verbs are used to relate them with the event that is being encoded. An example description is *A woman drinks from a cup while a man prepares a chicken to be cooked*.

A significant negative interaction is observed also between Clutter and Cue. The minimal amount of visual information available in the Clutter-Minimal scenes makes it more difficult to select and encode the actions performed by the actor (Cue-Animate), resulting in the generation of fewer verbs. This result is in line with the longer Looking Time for the same interaction. We can assume that the greater number of verbs found in Clutter-Minimal can be attributed to Cue-Inanimate, in which the ambiguous visual reference leads to more elaborate descriptions. An example description that illustrates this interpretation is *An open book is sitting on the counter and there is another one sitting on the table.*

Syntactic Constructions We also selectively analyzed a number of syntactic constructions contained in the responses generated by the participants.

Such constructions provide information about the sentence structures employed to describe the pictures. We counted how often participants employed the existential *there* construction. The results show that this construction occurred less frequently in

the Cue-Animate condition ($\beta_{Animate} = -0.2153; p < 0.05$). This indicates that participants were less likely to give static spatial descriptions of animate visual referents, compared to inanimate ones.

We also find that *and* is used less frequently in the Cue-Animate condition ($\beta_{Animate} = -0.0868; p < 0.05$). This result can be attributed to the ambiguous visual reference of Cue-Inanimate. The use of *and* marks a strategy of ambiguity resolution when both visual referents for Cue-Inanimate are linguistically encoded. The connection between referents is established by combining clauses through coordination.

When we analyzed the number of passive constructions, we again found a significant negative effect of Cue-Animate ($\beta_{Animate} = -0.0436; p < 0.05$). This is in line with standard findings in the sentence production literature: animate entities are more likely to be realized as subjects of active constructions, while inanimate tend to be realized as subjects of passive constructions (assuming that the cued entity is typically realized as a subject). An example of a production that contains the use of both coordination and passive is *A teddy is being hugged by the girl sitting on the bed and another teddy is sitting on the floor at the corner of the bed.*

4.4 Discussion

In Experiment 4 we investigated how visual factors influence sentence encoding during a web-based experiment, where participants are asked to write a description of a photo-realistic scene, after being prompted with a cue word. We assumed visual and linguistic processing to interact on-a-par over a shared referential interface, hence enabling cross-modal synchronous processing. Thus, we hypothesized that by changing the referential information in one modality, i.e. vision, we influence the processing of the other one, i.e. language. We manipulated visual reference such as visual clutter and the number of potential actors, and the animacy of the cue word used for sentence production. Moreover, we systematically introduced visual referential ambiguity for the inanimate cue in order to investigate the strategies of ambiguity resolution adopted.

The analysis of Looking Time showed significant effects of the visual factors such as Clutter and Actors: the more clutter or actors, the longer the time the participants spent before starting to type the sentence. The Animacy of the cue was also significant: an inanimate cue resulted in longer Looking Time, mainly because of visual referential

ambiguity. However, more interesting were the interactions between the visual factors and Animacy. If we assumed independence between visual and linguistic processing, we would expect response latencies typical of standard visual search tasks, based on the referential properties of the cue and influenced only by the visual properties of the stimulus. Instead, we found a clear interaction of visual information and Animacy. A visual scene with minimal clutter means that the set of actions that can be used to relate animate actors is impoverished. Thus, longer visual search is required to integrate the animate cues with information of the scene, the opposite of what is predicted under an explanation in terms of visual search alone. On the other hand, two actors in a scene mean a larger set of conceptual structures is available to relate to the animate cue. This interpretation meshes with the results we obtained for the syntactic analysis of the responses produced by participants. For the Actors-Two and Cue-Animate conditions, we found that longer sentences were produced (containing more nouns and verbs), often encoding both actors. Such results can only be explained by an account where linguistic and visual processing interact closely.

We also analyzed Description Time and the syntactic structure of the responses and found that these are strongly influenced by the Animacy of the cue and the presence of visual referential ambiguity. When the cue was inanimate, participants spent more time resolving the visual ambiguity. The sentences produced in this condition contained more nouns, which were used to spatially disambiguate between the two competing visual objects. Moreover, the disambiguation often occurred together with the use of the conjunction *and*. In line with previous research on language production, the use of passives and existential *there* was also correlated with the inanimacy of the cue. An inanimate cue is more likely to be a subject of a passive clause and in fact, correlated with static spatial descriptions by our participants.

With experiment 4, we gathered convincing evidence for the active impact of visual factors during sentence production. However, a web-experiment can only give us a coarse and indirect representation of how visual attention is actively retrieving information in support of sentence production. In experiment 5, we design a similar experiment, but this time participants are eye-tracked. The goal is to investigate how visual attention is influenced by the referential information of scene, i.e. clutter, and target, i.e., animacy, during sentence production. Moreover, we explore the different

ambiguity resolution strategies emerging when visual and linguistic referential information is integrated, i.e., during mention of the linguistic referent.

4.5 Experiment 5: Object-based information on situated language production

In experiment 5, we investigate how scene clutter and the animacy of cued targets influence visual attention during descriptions of photo-realistic, referentially ambiguous scenes. The density of visual information (clutter, Rosenholtz *et al.* 2007) is expected to negatively correlate with visual search performance: the more cluttered the scene is, the less efficient the identification of target object (Henderson *et al.*, 2009b). In experiment 4, on the contrary, we observed clutter to facilitate sentence production, especially for animate referents, whereas an opposite effect was found with minimal scenes.

Visual density can be seen as a coarse measure of referential information: the more visual information there is, the more linguistic material can be retrieved; thus, in line with experiment 4, we expect a positive correlation between scene information and sentence encoding. On the visual responses, this expectation should be reflected by more inspections on the visual referent before its mention when the scene is cluttered. While the action of mentioning is taking place, visual attention is contextualizing the mentioned referent within the scene while helping the resolution of referential ambiguity.

Moreover, in experiment 4 we have observed that object-specific information, i.e. animacy, triggers different types of sentence encoding. In experiment 5, we expect this effect to emerge also on the visual responses. In particular, animate referents, which are associated with larger conceptual structures and expected to facilitate both linguistic (Branigan *et al.*, 2008) and visual (Fletcher-Watson *et al.*, 2008) processing, should be inspected more than inanimate referents, especially when the scene has minimal clutter. A minimal scene has overall less referential information to be used during sentence encoding; thus animate referents, which carry crucial action information, are more thoroughly inspected.

Finally, this experiment makes it possible to investigate the effect of referential ambiguity on the eye-voice span. In previous work, the relationship between linguistic and visual referents was unambiguous: looks to the visual referent always preceded naming (Griffin & Bock, 2000) and this trend exponentially increases towards its linguistic mention (Qu & Chai, 2008). In our setting, we expect a more complex gaze-to-name relationship caused by a process of visual disambiguation that arises both before and after the intended referent is mentioned.

4.5.1 Method

The experimental design used is similar to experiment 4. The major difference is a simplification on the number of conditions. We removed *Number of Actors* and made both linguistic *Cue* (Animate, e.g. *man*; Inanimate, e.g. *clipboard*), visually ambiguous: two MEN and two CLIPBOARDS are depicted in the scene[1]. The factorial design used crossed the two factors *Clutter* (Minimal/Cluttered) and *Cue* (Animate/Inanimate). Participants' eye-movements were recorded while they described photo-realistic scenes after being prompted with a cue word, which ambiguously corresponded to two visual referents in the scene (see Figure 6.1).

We use the same 24 scenes of experiment 4, but now in each scene together with the two referentially ambiguous inanimate objects, we inserted two ambiguous animate objects using Photoshop; *Clutter* was either added or removed. In addition to the 24 experimental items there were 48 fillers, in which we vary the number of visual referents corresponding to the cue: either 1 or 3.

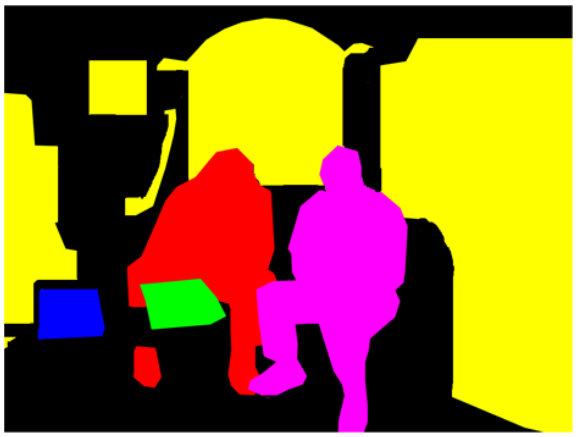
Twenty-four native speakers of English, all students of the University of Edinburgh, were each paid five pounds for taking part in the experiment. They each saw 72 items randomized and distributed in a Latin square design that made sure that each participant only saw one condition per scene.

An EyeLink II head-mounted eye-tracker was used to monitor participants' eye-movements with a sampling rate of 500 Hz. Images were presented on a 21" multiscan monitor at a resolution of 1024 x 768 pixels; participants' speech was recorded with a lapel microphone. Only the dominant eye was tracked. A cue word appeared for 750 ms at the center of the screen, after which the scene followed and sound recording

[1] In experiment 4, only Inanimate Cue, e.g. *clipboard*, was visually ambiguous.

Clutter: *Minimal;* **Cue:** *Man/Clipboard*

Clutter: *Cluttered;* **Cue:** *Man/Clipboard*



Background: Black
Clutter: Yellow
Primary Animate: Red
Secondary Animate: Pink
Primary Inanimate: Green
Secondary Inanimate: Blue

Figure 4.4: Example of an experimental trial, with visual region of interest considered for analysis. PRIMARY indicates that the ANIMATE and INANIMATE visual objects are spatially close and semantically connected (e.g., the MAN is doing an action using the CLIPBOARD). SECONDARY is used to indicate the remaining referent of the ambiguous pair. BACKGROUND and CLUTTER are defined in opposition: BACKGROUND is everything other than CLUTTER.

was activated. Drift correction was performed at the beginning and between each trial. There was no time limit for the trial duration and to pass to the next trail participants pressed a button on the response pad. The experimental task was explained using written instructions and took approximately 30 minutes to complete.

4.5.2 Data Analysis

We defined regions of interest (ROIs) both for the visual and the linguistic data. The visual data was aggregated into six different regions: PRIMARY and SECONDARY ANIMATE, PRIMARY and SECONDARY INANIMATE, BACKGROUND, and CLUTTER (see Figure 6.1).

For the linguistic data, we made a general division between time windows *Before* and *During* production. This allows us to capture the overall trend of the two main phases of a trial. For the analysis of eye-voice span, we consider a window of 2000 ms before the referent was mentioned, similar to Qu & Chai 2008. The resolution of visual ambiguity is analyzed using a window of 1600 ms (divided into 40 time slices of 40 ms each): 800 ms before and after the mention of *Cue*. This makes it possible to explore how the linguistic referent is visually located before being mentioned and just after.

In order to unambiguously analyze fixated and named referents, we aggregate eye-movements responses in four blocks (Primary, Secondary, Ambiguous and Both) by manually checking which referent was mentioned in each sentence. The distinction between Primary and Secondary is based on the spatial proximity and semantic relationship holding the cued objects. PRIMARY means that the Primary Animate or Inanimate is mentioned, and they are spatially and semantically related (e.g., *The man is writing on the clipboard*). SECONDARY is used when the Secondary Animate or Inanimate is mentioned (e.g., *The man is reading a letter*), and they are unrelated, i.e. man and clipboard are spatially and semantically independent[1]. A result emerging from this visual difference between Primary and Secondary is that across the whole set of sentences, Secondary objects are encoded only 14.23% of the time, in contrast to the 41.66% for Primary objects[2]. We introduced referential ambiguity as predictor in the inferential model described below to investigate how looks to the mentioned object differ from those to its competitor. We present analysis for the *Primary* and *Secondary* objects mentioned. The effect of mention on eye-movements' pattern is evaluated by comparing Primary with Secondary objects. Thus, for example, when Primary Animate is mentioned, more looks are expected on the man writing on the clipboard, whereas if Secondary Animate is mentioned, the other man should receive more looks (see Figure 6.1 to visualize).

As an initial exploration of our data, we investigate the overall trend of fixations *Before* and *During* production. Production is a task with large between-participant vari-

[1] AMBIGUOUS is used when is unclear which one is referred to (e.g., *the man is sitting on the couch*). BOTH indicates that both referents are mentioned (e.g., *the man is writing on a clipboard while the other man reads a newspaper*).

[2] In section 4.5.3.3, we see that this result carries important consequences for the visual responses observed.

ability, e.g., one participant will spend 2000 ms *Before* and 1000 ms *During* production, whereas another one will show the opposite pattern. Normalizing the production data is therefore crucial, in particular as we want to interpret eye-movements in relation to phases of linguistic processing. We normalize each sequence S^i_{old} of eye-movements by mapping it onto a normalized time-course of fixed length S^i_{new}. The length of S^i_{new} is set on the basis of the shortest eye-movement sequence $\min_i[\text{length}(S^i_{old})]$ found between Before and During production, across all participants.[1] For each sequence S^i_{old}, we obtain the number of old time-points k^i corresponding to a new time-unit u, as $k^i = \text{length}(S^i_{old})/\text{length}(S^i_{new})$. Proportions are then calculated over k^i old time-points and subsequently mapped into the corresponding unit u of the normalized time-course. In the Results section, we show plots of normalized proportions for *Primary* and *Secondary* (Animate and Inanimate) across conditions over the two regions *Before* and *During* production, 30 bins of normalized time each.

To explore the eye-voice span hypothesis, we compute the number of fixations to the mentioned object compared to the competitor. We also look at latencies, i.e., the onset of the last fixation to the referent or competitor before the mention, and gaze duration as a function of latencies, i.e., the time spent looking at the referent or competitor for the different latencies.

We also report inferential statistics for the referent region (for the time windows previously described). The dependent measure is the empirical logit (Barr, 2008), calculated as $emplog(\phi) = \ln \frac{0.5+\phi}{0.5+(1-\phi)}$, where ϕ is the number of fixations on the region of interest. The analysis is performed using the framework of linear-mixed effect (LME) models as implemented by the R-package lme4 (Baayen *et al.*, 2008). The predictors included were *Animacy*, *Clutter*, *Time* and *Object*. The random factors were *Participant* and *Item*. To reduce co-linearity, factors were centered.

The model selection followed a conservative stepwise forward procedure that tests the model fit based on a log-likelihood test comparing models each time a new parameter is included. If the fit improves the likelihood, we accept the new model, otherwise we keep the old one. We include predictors, random intercepts and slopes ordered by

[1] We remove outliers that are two standard deviations away from the mean, after having log-transformed our data. The data are not normally distributed, due to right skewness. The log-transformation helps us to reduce the skew.

their log-likelihood impact on the model fit. We iterate until there is no more improvement on the fit leaving us with the best model. In the results section, we report and interpret the LME coefficient estimates of the predictors retained after model selection for *Primary* and *Secondary* mentions, fitted in separate models.

4.5.3 Results and Discussion

4.5.3.1 Before and During Production

We first look at how fixations are distributed when we collapse the two main phases of the experiment: *Before* and *During* production. This analysis does not distinguish whether the *Primary* or *Secondary* referent was mentioned. Figure 4.5 shows normalized proportions of looks on the competitor visual objects corresponding to the *Cue* (Animate/Inanimate).

The first thing to note is that for the visual ROI corresponding to the *Primary* referent, the pattern of fixations is more complex than for the ROI of the *Secondary* referent. The spatial proximity and semantic relatedness of the two *Primary* referents result in a more complex pattern of interaction. The clearest effect is found in relation with the animacy of *Cue*; we observe more fixations to the animate referent when the cue is also animate. When looking at the *Primary* ROI, the effect is seen at the beginning of both the Before and the During region. At the beginning of the trial, the visual system retrieves information about the cued objects; when production starts, the referents are fixated again, probably before being mentioned. For the *Secondary* ROIs, the relation with the *Cue* is stronger, probably reinforced by the referential competition. Moreover, the pattern of looks is much clearer than for the *Primary* ROI. This confirms that spatial proximity and semantic relatedness increase the interaction between visual referents. *Clutter* does not have a strong effect, though there is a small increase of looks when the scene is minimal and the animacy of the target matches that of the cue.

4.5.3.2 Eye-Voice Span

We analyzed eye-voice span to investigate the gaze-to-name relation for the mentioned referent and its competitor. Table 4.3 shows percentages of looks to referent or com-

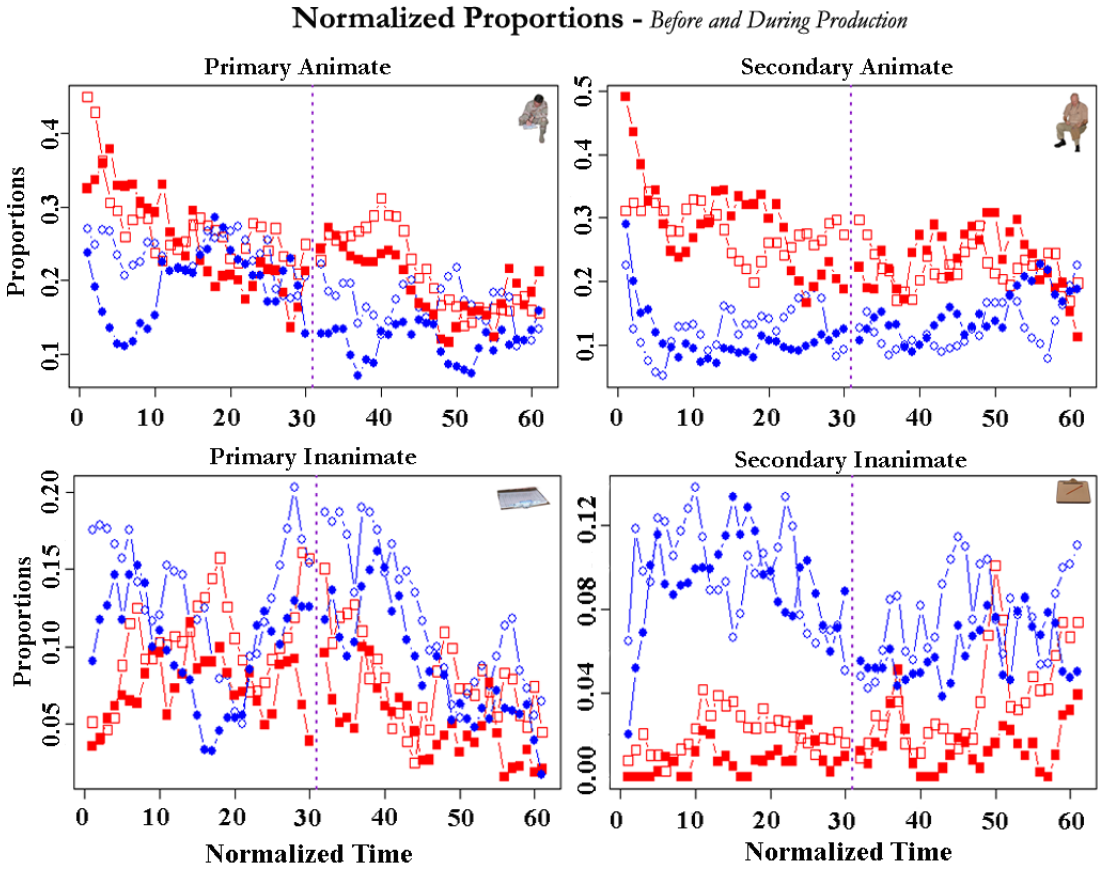


Figure 4.5: Normalized proportions of looks (60 bins) across the four conditions, Before and During production, for the different visual ROIs. The colors are used to indicate the animacy of *Cue* (red - Animate; blue - Inanimate); the line type instead indicates *Clutter* (open - Minimal; closed - Cluttered). The purple dashed vertical line indicates Before (to the left) and During (to the right) production.

petitor with mean latencies and gaze durations.[1]

There is a preference for looks to the referent over looks to the competitor, with a latency of about one second, confirming previous findings (Griffin & Bock, 2000). In a minority of cases, participants only look at the referent (36.44%); competition between the two ambiguous visual referents is the norm (71.65%). Moreover, we notice that the competitor is fixated earlier than the referent and the duration is shorter for the Including condition (which includes trials in which both referents have been fixated).

[1]The measures are calculated only when the Primary and Secondary referent are mentioned; thus, we exclude the Both and Ambiguous cases, for which it was not possible to establish unambiguous eye-voice span relation.

Table 4.3: Eye-voice span statistics. *Excluding* indicates that the percentage is calculated considering only those cases in which either the referent or competitor have been fixated, *Including* takes into account also cases where both have been fixated.

Measure		Referent	Competitor
Percentage of looks	Including	71.65	43.30
	Excluding	36.44	8.09
Mean Latency	Including	1032 ms	1203 ms
	Excluding	1012 ms	1325 ms
Gaze Duration	Including	489 ms	432 ms
	Excluding	568 ms	623 ms

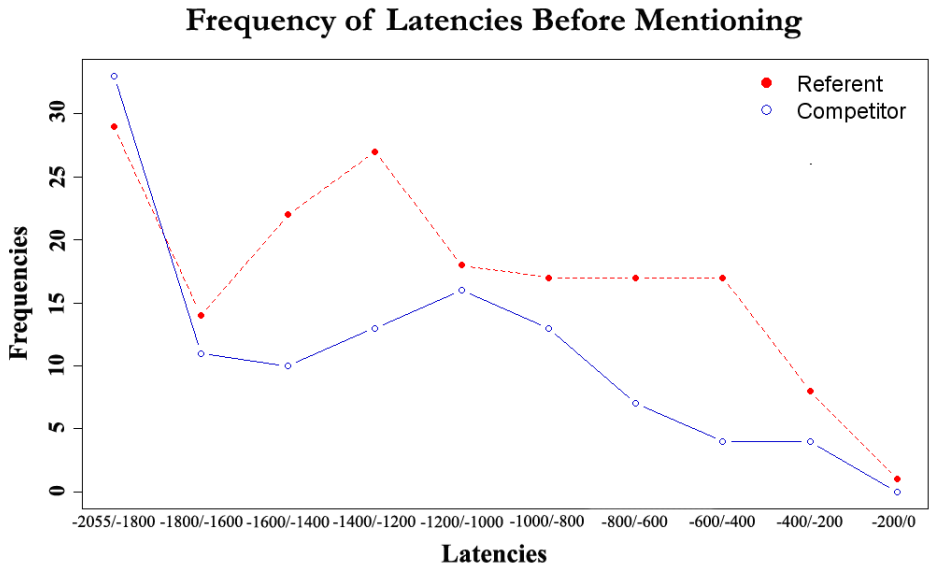
This may indicate that the final decision on which referent is mentioned is made after discarding the competitor.

Figure 4.6(a) shows frequencies of *Latencies* at different temporal blocks (200 ms each) within a total window of two seconds. We find that latency frequency decreases towards the mention for both the referent and the competitor. This finding contrasts with Qu & Chai (2008) who found the opposite trend, i.e., the closer to the mention, the more gazes are associated with the referent object. Note also that this effect cannot only be due to the presence of a competitor, e.g. comparative looks before mention, as these present a similar decreasing trend.

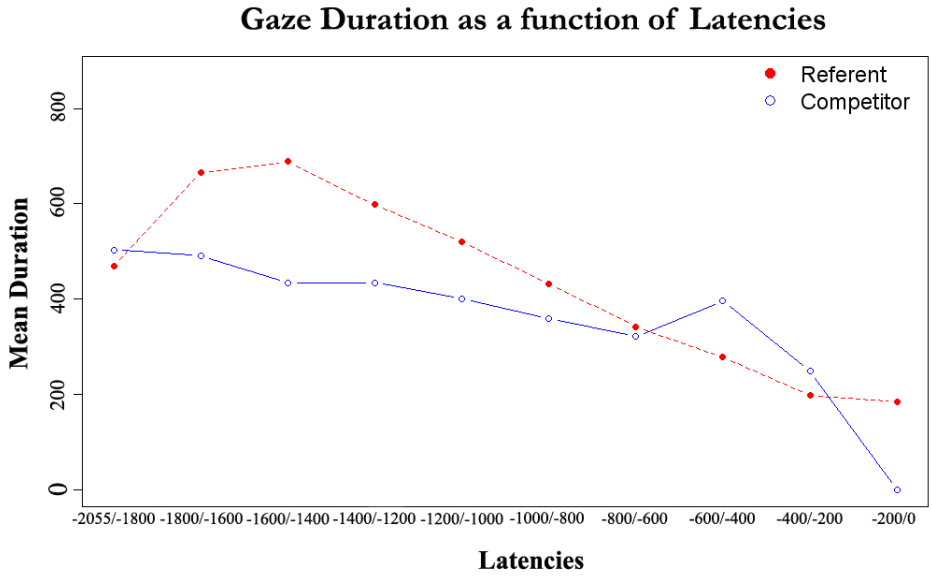
In Figure 4.6(b) we show mean gaze duration as a function of the different latencies. Again, a decreasing trend is clearly visible: the closer the latency to the mention, the shorter the gaze duration. Interestingly there is a peak of gaze duration at 1600/1400 ms. The higher duration found at this latency might be an indicator of referential selection (gaze-to-name binding). We also find evidence of competition at 600/400 ms, where the competitor receives longer gazes compared to referent. A last visual check on the competitor is probably performed before referentiality is encoded linguistically.

4.5.3.3 Inferential Analysis

We now analyze the pattern of eye-movements before and after the mention of the cue word. We focus on the case where the Primary visual object is mentioned, and briefly



(a) Frequencies of latencies at different temporal blocks (from two seconds to mention): red is the referent, blue the competitor. The latency measures the time elapsed from the beginning of the last fixation to the object (referent or competitor) until it is mentioned.



(b) Mean gaze duration as a function of latency. The mean of gaze duration is calculated for the different blocks of latencies. We analyze only cases where gaze duration is shorter than latency, thus avoiding cases where fixations spill over into the region after mention.

Figure 4.6: Eye Voice Span statistics.

discuss results where the Secondary visual object is mentioned. Based on the eye-voice span analysis, we expect to find a decreasing trend of looks before the referent is mentioned, and the presence of competition should weaken the gaze-to-name relationship. Recall that our experiment had two factors (Cue: animate/inanimate; Clutter: minimal/cluttered); we also include the object fixated (Object: primary/secondary) and Time (in 40 ms slices, see Data Analysis above) in the analysis.

Primary Mentioned In Table 4.4 we report LME coefficients estimates for the predictors retained after selection of four separate models: Animate and Inanimate objects, Before and After mention[1].

Beginning with the animate visual objects, we expect the *Primary Animate* to receive more looks than the *Secondary Animate*, and the number of looks should increase. We observe a preference for looks to Primary Animate, but the difference is not statistically significant.

However, we find a main effect of *Cue*: an animate cue facilitates looks to Animate visual objects. When looking at the time course, we find a general decreasing trend, partly compensated by a three-way interaction of *Animacy*, *Object*, and *Time*. Moreover, we observe a two-way interaction of *Clutter* and *Time*: a minimal scene makes it difficult to retrieve disambiguating information for the animate referent, forcing the visual system to look for this information on the referent itself. It is also conceivable that the minimality of the scene makes visual responses similar to those found for line drawings (Griffin & Bock, 2000), thereby explaining the increasing trend. In a cluttered environment, instead, there are more ways to relate the referent to the surrounding context, hence helping language production to disambiguate. This explains the decreasing trend of fixations on the referent in the cluttered condition.

After mention, we observe interactions of *Cue* with *Clutter* and *Object*, confirming both the facilitation of the cued referent and the preference for referent information when scenes are minimal. In contrast with previous findings, we observe increasing looks to the referent after mention. This effect could be due to referential ambiguity: the visual system is connecting disambiguating material retrieved before mention to the referent just uttered. For the *Secondary Animate*, we find an increasing trend of looks

[1]The intercepts for Before and During are different because they are calculated over distinct time intervals.

Table 4.4: LME coefficients of Animate Referents; Before and After Primary object is mentioned; Contrast coding: *Object*: Primary (-0.47), Secondary (0.53); *Cue*: Animate (-0.53), Inanimate (0.47); *Clutter*: Cluttered (-0.52); Minimal (0.48)

	Region-Before	
Predictor	Coefficient	*p*
Intercept	-3.5100	0.0001
Cue	-0.1011	0.02
Time	-0.0046	0.01
Object	-0.0538	0.3
Clutter	0.0241	0.5
Cue:Time	0.061	0.0003
Cue:Clutter	0.0441	0.07
Clutter:Time	0.043	0.007
Object:Time	0.045	0.004
Cue:Object:Time	0.098	0.001
	Region-After	
Predictor	Coefficient	*p*
Intercept	-3.4956	0.0006
Object	-0.0647	0.1
Clutter	0.0271	0.5
Cue	-0.0790	0.2
Time	0.0005	0.7
Clutter:Time	-0.002	0.2
Object:Time	-0.056	0.0002
Object:Clutter	0.0674	0.0005
Clutter:Cue	-0.0632	0.01
Object:Cue	0.0551	0.008
Cue:Time	-0.0016	0.3

when *Cue* is Inanimate and especially for minimal scenes. The minimality of the scene gives prominence to animate referents; probably the spatial and semantic proximity of one of Primary Inanimate and the Primary Animate also trigger comparative looks to Secondary Animate, i.e., participants check whether it can also be contextually related to the cue.

After the referent is mentioned (*Primary* in this case), looks to the *Secondary* Animate decrease over time in all conditions. Competition is triggered by visual am-

biguity, but once the association of the visual with the linguistic referent has been established (i.e., after the mention), participants look back to the referent mentioned, presumably finalizing the choice made.

Looking at inanimate referents, we observe a statistically significant preference for looks to the Primary Inanimate; refer to Table 4.5 for full list of coefficients. This preference could be due to the spatial proximity and the semantic relation with the primary animate, which makes the primary inanimate more likely to be encoded either as a direct object or as subject of the description. As a consequence, we find an interaction with the animacy of the *Cue* but not a main effect. In contrast with standard visual search task, where performance degrades as a function of clutter, here we observe instead a positive interaction of *Clutter* and *Cue* on the target, which increases over time. The visual system is not performing a search task, rather it is sourcing information to ground language processing. In a cluttered scene, an inanimate referent could be spatially related to many other different objects, whereas a minimal scene has fewer points to anchor the referent. The visual system therefore needs to select among the different spatial relations to find one that optimally situates the object within the contextual information.

For the secondary inanimate, there is a negative relationship between the animacy of *Cue* and the minimality of *Clutter*; the proximity and relatedness of the primary inanimate and the primary animate is highlighted when visual information is minimal, which results in the secondary inanimate referent being fixated less. We don't find any significant effect after mention.

Secondary Mentioned As explained in section 4.5.2, Secondary objects have a less prominent role in the scene than Primary objects, which is due to the absence of any semantic or spatial relationship linking the two Secondary objects. This implies less looks and more independent looks on the Secondary objects (see Figure 4.5 to visualize the global pattern), as their visual ROI are distinctly depicted in the scene (refer to example of ROI 6.1).

When looking at the Secondary Animate mentioned, we find a main effect of *Object*, in that the Secondary receives overall more looks compared to the Primary, which increase in time the closer to mention and it is positively influenced by visual density and animacy of cue (see Table 4.6, for list of coefficients). As expected, compared to

Table 4.5: LME coefficients of Inanimate Referents; Before. Primary object is mentioned; Contrast coding: *Object*: Primary (-0.34), Secondary (0.65); *Cue*: Animate (-0.6), Inanimate (0.4); *Clutter*: Cluttered (-0.54); Minimal (0.45)

	Region-Before	
Predictor	Coefficient	*p*
Intercept	-3.4800	0.0001
Object	-0.1786	0.03
Cue	0.0465	0.6
Clutter	-0.0325	0.6
Time	-0.0018	0.5
Object:Cue	0.0724	0.03
Cue:Clutter	-0.0874	0.03
Clutter:Time	-0.05	0.01
Cue:Object:Clutter	0.3955	0.001
Object:Cue:Time	-0.082	0.05
Cue:Clutter:Time	-0.131	0.001
Object:Clutter:Time	0.0048	0.2

primary animate, a secondary animate is more clearly inspected before its mention and this effect is due to the absence of semantic and spatial interaction with other ROI. We observe a positive interaction of Secondary Animate with Clutter, but only when the object is not cued. When the object is cued and situated in a cluttered scene, we observe less looks, which decrease over the time of mentioning. The scene context is a source of referential information visually accessed to support sentence encoding. Similarly to what was observed for the Primary Animate, before a referent is mentioned, sentence processing demands visual attention to retrieve referential information, in order to resolve referential ambiguity while continuing the process of encoding.

After mention, again, we find a main effect of *Object* and *Cue*, especially when the object is mentioned, and the scene is cluttered, which shows an increasing trend. Similar to what observed when Primary Animate is mentioned, scene information retrieved before mentioning is visually linked to the referent just after the act of mention-

Table 4.6: LME coefficient estimates. Animate Referents: Before and After Secondary object is mentioned. Contrast coding: *Object*: Primary (-0.53), Secondary (0.47); *Cue*: Animate (-0.46), Inanimate (0.54); *Clutter*: Cluttered (-0.53); Minimal (0.47)

	Region-Before	
Predictor	Coefficient	*p*
Intercept	-3.4891	0.0001
Object	0.2121	0.001
Cue	-0.0572	0.02
Clutter	0.0533	0.008
Object:Cue	-0.2975	0.001
Object:Clutter	-0.1836	0.001
Clutter:Time	0.107	0.002
Object:Time	0.079	0.006
Object:Cue:Clutter	-0.1666	0.001
Cue:Clutter:Time	-0.1620	0.001
Object:Clutter:Time	0.1050	0.04
	Region-After	
Predictor	Coefficient	*p*
Intercept	-3.4910	0.0001
Object	0.2413	0.001
Cue	-0.1204	0.001
Clutter	0.0465	0.01
Time	-0.031	0.02
Object:Clutter	-0.1881	0.0001
Cue:Time	-0.0135	0.0001
Clutter:Time	-0.133	0.0001
Object:Cue	-0.1149	0.0001
Object:Time	0.079	0.0001
Object:Cue:Time	0.1450	0.01
Object:Clutter:Time	-0.1320	0.01

ing starts; and this is in contrast with previous findings regarding the eye-voice span, where looks to the visual referent are found abruptly decreased after its mention (see Figure 4.1 for a comparison).

Turning to the inanimate referents, before mention, the Secondary object receives more looks than Primary, especially when the scene is cluttered, see Table 4.7 for the

Table 4.7: LME coefficients of Inanimate Referents; Before and After Secondary object is mentioned; Contrast coding: *Object*: Primary (-0.46), Secondary (0.54); *Cue*: Animate (-0.67), Inanimate (0.33); *Clutter*: Cluttered (-0.6); Minimal (0.4)

	Region-Before	
Predictor	Coefficient	*p*
Intercept	-3.5141	0.0001
Object	0.2229	0.001
Time	-0.0070	0.001
Cue	0.0719	0.1
Object:Clutter	-0.3077	0.001
Cue:Clutter	0.2522	0.001
Object:Time	-0.079	0.01
	Region-After	
Predictor	Coefficient	*p*
Intercept	-3.5172	0.001
Object	-0.0433	0.1
Cue	-0.0078	0.9
Object:Cue	-0.3515	0.001
Cue:Clutter	0.2475	0.001
Object:Clutter	-0.1245	0.01
Cue:Time	0.095	0.0001

full list of coefficients. In line with visual search studies, inanimate referents are more easily identified when a scene is minimal. However during description, visual attention focuses on the information of the mentioned object to spatially locate it within the scene. This process is especially important in cluttered scenes, where more locations can be spatially related with the target object. Regarding the eye-voice span, we confirm our findings of decreasing looks, especially when the object is mentioned.

After mention, we find interactions similar to those found before mention. Inanimate referents are more easily identified in minimal scenes. However, during description, a cluttered scene makes visual attention focus more on the mentioned object. Moreover, looks tend to increase on inanimate referents after mention, but significantly less for the mentioned object. The referential ambiguity drives visual comparison between the mentioned referent (Secondary) and its unselected competitor (Primary).

4.6 Discussion

In an eye-tracking experiment, we have investigated the impact of visual information density, i.e. clutter of the scene, and target semantics, i.e. animacy of the target object, on visual attention during a cued language generation task situated in referentially ambiguous, photo-realistic scenes. Our hypothesis, supported by evidence gathered in experiment 4, is that visual factors have a direct influence on sentence encoding. Here, we tested how this influence is reflected on the patterns of visual attention emerging during referential integration, i.e. when the visual referent is linguistically mentioned. Previous accounts investigating the gaze-to-name relation, i.e. eye-voice span, have found that a visual object is looked at just before being mentioned (Griffin & Bock, 2000; Qu & Chai, 2008). Here, we go beyond these results by initially testing the eye-voice span relation, but situated in naturalistic referentially ambiguous scene. Then we analyze the influence of visual factors on visual responses to the mentioned target, before and after its linguistic mention.

We expected clutter to facilitate processes of sentence production, especially during the resolution of referential ambiguity. In fact, the more visual information there is, the more linguistic material is available to solve referential ambiguity. From experiment 4, we also expected the conceptual properties of visual target, i.e. animacy, to influence sentence encoding and interact with visual information density.

The results of our fifth experiment showed that the animacy of the cue facilitates looks to animate objects, especially at the beginning of two main phases of linguistic production: before and during the mention of the referent. The data therefore indicate that a visual search is performed to localize the objects matching the cue word (Malcolm & Henderson, 2009). But our results also contrasted interestingly with findings for visual search, where clutter decreases search performance (Henderson *et al.*, 2009b). In cases in which an animate referent is mentioned, we found that there were fewer fixations to the target object in the cluttered condition compared to the uncluttered one. In other words, clutter makes language production easier, not harder: the visual system is not just searching for the target object, but it is also retrieving visual information that can be used to linguistically anchor it (e.g., for disambiguation). The more clutter there is, the easier this process becomes, explaining the reduced number of fixations in the cluttered condition.

Turning to the relation between fixating and naming an object (the eye-voice span), previous work found that referents are fixated shortly before being mentioned (Griffin & Bock, 2000). It has also been observed that fixation probability increases with decreasing distance to the mention (Qu & Chai, 2008). In our data, we found and quantified a preference for looks to the mentioned referent over looks to the competitor, but this preference was not confirmed in the inferential analysis (see Table 4.4). Only if the primary inanimate was mentioned, it received significantly more fixations than the secondary inanimate. This preference is likely due to the proximity, spatial and semantic, between the primary animate and inanimate. Moreover, we found that fixation probability decreased with decreasing distance to the mention, contrary to previous results, in particular when the scene was cluttered and the mentioned referent was animate. The competition between visual referents seems to override the standard eye-voice span effect. Interestingly, we also observed an increasing trend of fixation to the referent object *after* its mention. Once production has started, the visual system needs to retrieve contextual information to produce disambiguating linguistic material, resulting in an increase in the number of looks after mention.

4.7 General discussion

When sentence processing and scene understanding interact in situated language processing tasks, cross-modal mechanisms are activated. In Chapter 3, we have seen how image-based, i.e. saliency, visual information is exploited by the sentence processor to make predictions of upcoming arguments during situated understanding. Image-based information, however, is not referential, whilst both scene understanding and sentence processing are known to highly rely on reference and its factors. Thus, in this chapter we investigated how the object-based factors of *clutter*, a measure of scene information density (Rosenholtz *et al.*, 2007), and *animacy*, a conceptual property of individual objects (Branigan *et al.*, 2008), modulate the cross-modal interaction of visual and linguistic processing in a situated language production task. We decided for a production task, instead of comprehension, in order to observe how scene information, actively retrieved by visual attention, is naturally associated with different types of sentence encoding.

In experiment 4 (web scene description), we tested how clutter and animacy influence reactions times for the different phases of sentence production, while exploring the descriptions generated by looking at their syntactic information. The results show that descriptions of animate referents are positively correlated with scene information: when description is situated in a minimal scene, we find slower reaction times and shorter descriptions. Moreover, in line with language production studies (Branigan *et al.*, 2008), animate referents are more quickly encoded and this effect cumulates over the number of actors depicted: the more actors, the larger the number of conceptual structures available for description.

In experiment 5 (eye-tracking), we looked at the impact of clutter and animacy on eye-movement responses during the phase of referential integration, i.e. before and after the cued linguistic referent is mentioned. In contrast with previous studies (Griffin & Bock, 2000), we show that looks decrease before a referent is mentioned, especially when the scene is cluttered and the visual referent encoded is animate. Visual attention is retrieving disambiguating visual material to support sentence production. For inanimate referents, looks are positively correlated with clutter both before and after mention. This interaction is unexpected from visual search studies, where clutter is found negatively correlated with search performance (Henderson *et al.*, 2009b). We find that an inanimate referent is usually described in spatial relation with another object. A cluttered scene has overall more spatial locations in relation to which the inanimate referent can be described. Thus, visual attention focuses more on the visual referent during referential integration to evaluate which spatial relation best contextualizes it within the scene.

Clutter (Rosenholtz *et al.*, 2007) is a measure of visual information density, which differently from saliency (Itti & Koch, 2000b) considers objects' edges, thus making it a better indicator of referential scene information. In the visual cognition literature, clutter has been found negatively correlated with search performance: the more clutter, the more difficult is the identification of target objects (Henderson *et al.*, 2009b). However, in a language generation task, the density of visual information has an opposite and rather beneficial effect. In fact, the more referential information is visually available, the easier the process of sentence encoding becomes.

Animacy is a conceptual property of referents, which has important implications for both linguistic and visual processing. In line with previous literature, we find ani-

mate objects to boost sentence encoding (e.g. Levelt *et al.* 1999) and facilitate visual retrieval (Fletcher-Watson *et al.*, 2008). However, the animacy of the object described shows important interactions with the surrounding scene information. In particular, during mention of an animate object, participants inspect the scene to retrieve referential information to support encoding. Thus, when there is more clutter, more visual material can be used to source the ongoing description. During mention of inanimate objects, instead, when the clutter is large, a higher specificity is required in the selection of spatial information to anchor the described object.

Regarding the mechanics underlying the association between visual and linguistic referents, our study demonstrates a more complex pattern of gaze-to-name relation than that previously shown (Griffin & Bock, 2000; Qu & Chai, 2008). In particular, an object is gazed at either before or after its mention, and this relation is modulated by its referential ambiguity, animacy, and the density of the surrounding scene.

Taken together, our results indicate that visual factors such as clutter interact with conceptual factors such as animacy in language production. The simple view according to which referents are fixated in the order in which they are mentioned, with a fixed eye-voice span between fixation and mention, does not seem to generalize to more realistic settings in which speakers describe naturalistic scenes that involve referential ambiguity.

4.8 Conclusions

During description of naturalistic scenes, visual referential information is actively sourced by visual attention to drive sentence encoding. During this cross-modal interaction of referential information, different visual factors have a direct impact on the type of sentences produced and on the patterns of corresponding visual responses.

This finding suggests a close relationship between linguistic descriptions and visual responses, triggered by the cross-modal interaction of scene and object properties, which implies a general mechanism of cross-modal referential coordination.

In Chapter 5, we explore the association between visual and linguistic referential information by quantifying their cross-modal coordination. The hypothesis we test is that the more similar two sentences are, the more similar the associated visual responses (scan patterns) will be. Moreover, if the cross-modal coordination is based on

the referential identity shared between objects and words, rather than on visual properties of the scene per-se, we predict this coordination to hold across different scenes.

Chapter 5

Cross-Modal Coordination between Scan Patterns and Sentence Production

5.1 Introduction

The majority of everyday tasks demands a range of cognitive modalities to be actively involved; and this raises the question to which extent these modalities are coordinated. During cross-modal synchronous interaction of vision and language, e.g. a scene description task, visual attention and sentence processing have to be coordinated in their referential choices. If, for example, we are in a kitchen about to describe a MUG on a COUNTER, visual attention retrieves referential scene information, e.g. looks to MUG or COUNTER, according to the sentence processing output, e.g. *the mug is on the counter next to the toaster*. In a nutshell, what we are looking at, has to be associated with the description synchronously generated.

Coordination is a fundamental mechanism guiding cross-modal interaction during synchronous processing, which is expected to emerge similarly across different people over different contexts. In fact, if two different people are similarly performing the same cross-modal task, the output of modalities involved has also to be similar. Thus, in a description task, if two participants give a similar description of a scene, e.g. *the mug is on the counter* vs *the yellow mug on the counter*, we expect also the corresponding scan patterns to be similar. Furthermore, if the referential information is

similar across scenes, e.g. different kitchens, we should still be able to observe cross-modal coordination: i.e. similar sentences generated on different scenes, should still be associated with similar scan patterns.

In this chapter, we investigate the cross-modal coordination between linguistic structures (sentences) and visual attention (scan patterns) during descriptions of naturalistic scenes[1]. Our main hypothesis is that similar scan patterns are associated with similar sentences. Our results show evidence of cross-modal coordination, which is consistently found across different phases of the generation task (visual planning, visuo-linguistic encoding and linguistic production); and for within-scene and between-scene analyses.

Moreover, when we include temporal information on scan-patterns, we find that coordination is strengthened when linguistic processing is actively involved. When linguistic processing acts directly on visual attention (encoding and production), the cross-modal referential integration demands modalities to be temporally coordinated. Finally, in line with findings in Chapter 4, an analysis of the factors, animacy and clutter, involved during cross-modal coordination, reveals that animate targets, e.g. man, trigger higher cross-modal similarity than inanimate targets; especially in low density scenes, where coordination is overall higher. However, when time is included, the target is inanimate, and linguistic processing is directly implicated, i.e. during production, we observe higher coordination for cluttered scenes. During linguistic production, the inanimate referents, e.g. CLIPBOARD, have to be spatially situated in the scene, thus once the ground referent, e.g. TABLE, is selected, the two modalities coordinate to unambiguously point at the intended referent. Thus, the more referential information is available in the context, the more specific coordination has to be, in order to precisely associate the visual information retrieved with the associated linguistic description.

5.2 Background

Cognition arises through the integrated contribution of several modalities, which constantly interact on a wide variety of everyday tasks. A simple task like making a tea requires, for example, the joint interaction between visual attention, e.g. informing

[1] We use the same dataset obtained in experiment 5.

about the position of the tea-pot, and motor-actions, e.g. grasping the handle. This multi-modal flow of information generated during the task, however, has to be coordinated, i.e. looking at the tea-pot before moving the arm, in order to correctly perform the action. Understanding the coordinative mechanisms allowing cross-modal interaction is crucial to formulating a unified theory of cognition.

Research on eye-movements during daily actions (Land, 2006) has shown that scan-patterns, i.e. eye fixations across spatial locations in temporal order (Noton & Stark, 1971) as well as the sequence of motor actions, are similarly coordinated across participants. Despite the high variability of scan patterns (Henderson, 2003); in tasks requiring active allocation of attention (Findlay & Gilchrist, 2001; Henderson, 2007; Neider & Zelinsky, 2006), top-down, object-based[1], information guides the deployment of visual attentional resources (Brockmole & Henderson, 2006; Torralba *et al.*, 2006). The cognitive control imposed on visual attention is driven by the combined interaction of contextual expectations (Castelhano & Heaven, 2010; Malcolm & Henderson, 2010) and tasks' goals (Castelhano *et al.*, 2009).

Contextual expectations about a scene are generated within the first fixation (gist) (Potter, 1976; Vo & Henderson, 2010). Immediately after, visual attention is driven to those scene's locations relevant to the task. During visual search, for example, contextual expectations are combined with referential information about the cued target to identify it (Malcolm & Henderson, 2009; Yang & Zelinsky, 2009). Moreover, the localization and identification of cued targets is boosted by linguistic referential information, available prior to the task (Schmidt & Zelinsky, 2009). Linguistic information, e.g. *on the left corner*, can narrow the space of visual search to precise regions of the scene. The contextual expectations are built on the referential relationship, both visual and linguistic, occurring among the objects composing the scene. Objects, e.g. MUG and COUNTER, which tend to co-occur within the same scene context, e.g. kitchen, are expected to be contextually related, e.g. a mug is usually on a kitchen counter. The guiding effect of contextual information is observed also in other visual tasks, e.g. memorization (Humphrey & Underwood, 2008; Hwang *et al.*, 2009), where scene layout information and semantic relationship between individual objects is found predictive of similarity across scan patterns of different participants.

[1] In the context of a kitchen, for example, we expect to find a knife on a counter rather than on the floor.

In goal directed tasks, the co-occurrence of scene referential information modulates endogenous allocation of visual attention, while providing a referential interface where modalities can coordinate. Each task, however, follows specific goals which directly influence the allocation of attention (Castelhano *et al.*, 2009; Henderson, 2007). Therefore, the referential information in the scene is accessed and utilized according to the type of task performed. Moreover, tasks differ on how modalities are involved, and on the type of interaction occurring among them. In a search task, for example, only the visual system is actively involved; whereas in a scene description, beside vision there is the intervention of language. Here, we want to make clear what we define as similarity within the same modality, and what instead is coordination across modalities (cross-modal similarity).

Similarity within a unique modality has been observed, for example, across participants in a memorization task, where scan patterns were found to be more similar on the same scene than on different scenes (Humphrey & Underwood, 2008). Here, similarity, driven by referential scene information, is limited to the visual modality. In tasks demanding synchronous cross-modal interaction, e.g. making a tea, instead, coordination emerges as a result of cross-modal similarity. In this case, top-down control is activated to integrate referential information across modalities: where the output of one modality, e.g. looks to handle of the tea-pot, modulates the processing of another modality, e.g. arm movements.

The coordination of cross-modal information, e.g. auditory and visual, has been observed to strengthen the detection of objects and events, as reflected by the integration time (Evans & Treisman, 2010; Iordanescu *et al.*, 2010). During a recognition task, Zelinsky & Murphy (2000) observed that objects with longer names, e.g., *helicopter*, are fixated longer than objects with shorter names, e.g., *man*. The linguistic sub-vocalization of visual referents directly controlled visual attention on the object. Visual and linguistic entities are temporally coordinated on the referential object identity.

Evidence of the referential relation between visual and linguistic entities comes also from a psycholinguistic eye-tracking paradigm of studies (VWP, Tanenhaus *et al.* 1995), which has investigated language processing concurrently with a visual context. Research in this field has demonstrated clear links between the processing of certain linguistic constructions and access to visual contextual information (Knoeferle

& Crocker, 2006). The main finding regards the time-locked relation between linguistic material processed (understood or produced) and visual responses observed. Such a temporal relation is established as an interaction between different aspects of linguistic information processed, e.g. the prosodic phrasing of a sentence (Snedeker & Trueswell, 2003), and the referential information expressed in the visual context, e.g. edible vs non-edible objects depicted in association with the verb *eat* (Altmann & Kamide, 1999). Overall, these results suggest the existence of a general cognitive mechanism, underlying the cross-modal organization of visual and linguistic processing, which allows responses to be coordinated.

In this chapter, we bring together findings from the visual cognition and situated language processing literature by investigating the coordination of vision and language. Active visual attention has to utilize referential scene information to perform several different tasks. Likewise, during situated language processing, visual attention is mediated by the interaction between referential information of scene and sentence. It follows that during tasks demanding synchronous cross-modal processing visual attention and sentence processing are expected to be coordinated on the referential interface: their referential outputs, sentences and scan patterns, should be mutually associated. For example, if two different people say *the mug is on the counter*, their scan pattern should also be similar, e.g. looks to MUG and COUNTER, and this coordination should hold across different scenes, as long as the referential information contained overlaps, e.g. two different scenes containing at least a MUG and a TABLE. By quantifying the coordination of vision and language, we deepen our understanding of cross-modal processing, and this allows us to make predictions across different modalities, e.g. given a sentence, we can predict which scan pattern could be possibly associated.

5.3 Experiment 6: Cross-modal coordination of vision and language

In Experiment 6, we test the hypothesis that similarity of scan patterns is associated with the similarity of corresponding sentences. We assume similarity to be driven by a shared **referential interface** on which visual and linguistic processing are cross-modally **coordinated**. The referential interface modulates the top-down, object-based,

allocation of visual attention during goal directed tasks, i.e. only regions contextually and semantically related to the goals are observed (Henderson, 2007). For tasks demanding synchronous processing, viz., scene description in a visual context, the endogenous access to referential information has to be coordinated, viz. if two trials involve similar scan patterns, then the sentences produced in these two trials will also be similar. As seen in Chapter 4, referential information has general visual, e.g *clutter*, and linguistic, e.g. *animacy*, properties directly influencing the behavioral responses observed, e.g., visual attention, during descriptions of naturalistic scenes. In line with previous literature, animate targets are, for example, found facilitating linguistic production compared to inanimate targets (e.g., Branigan *et al.* 2008) whereas low clutter favors identification of targets (e.g., Henderson *et al.* 2009b), refer to Chapter 4 for details. However, beside these expected effects, we find several interactions between visual density and animacy of targets. Especially, during mention of an animate referent, scene context is widely inspected to source referential information in support of sentence encoding. Thus, the denser the scene is, the more referential information can be used to describe the target. For inanimate referents, which are often described using spatial relations, e.g. *the clipboard is on the table*, a higher visual density means more locations to spatially ground the object. Thus, once a spatial ground is selected, e.g. TABLE, visual attention focuses on the target object during its mention to unambiguously establish its spatial location. We expect cross-modal coordination to be modulated by these factors in line with what was observed in Chapter 4. Thus, higher coordination is expected for animate targets, especially when situated in minimal scenes. However, when visual and linguistic processing are more tightly coupled, e.g. during production, we expect coordination to increase in cluttered scenes, especially for inanimate targets. The more referential information there is, the more spatial locations can be used to locate inanimate referents; thus coordination tightens to unambiguously integrate the visual referential information about the target object to its linguistic realization.

In the next sections, we first summarize how the data was collected and processed, and explain how we computed the measures of scan pattern and linguistic similarity. Then, we investigate the coordination of vision and language by looking at the correlation between linguistic and visual similarity measures, across different phases of the language generation task, computed over the whole data set (global), and selecting a

subset of it where similarities are aggregated by scenes[1] (local). Moreover, in order to examine the role played by visual and linguistic factors, we explore the impact of target *animacy* and scene *clutter*, already examined in Chapter 4, on the cross-modal similarity, i.e., an aggregated measure of coordination.

5.3.1 Data Collection and Pre-processing

In an eye-tracking language production experiment (Coco & Keller, 2010b), discussed in Chapter 4, we asked participants to describe photo-realistic indoor scenes after being prompted with cue words which referred to visual objects in the scenes (refer to Chapter 4 for details).

We collected a total of 576 sentences produced for 24 scenes which were drawn from six different scenarios (e.g., bedroom, entrance). The sentences were manually transcribed and paired with the scan patterns that participants followed while generating them. We removed two pairs because the sentences were missing. Each scene has been fully annotated using the LabelMe toolbox (Russell *et al.*, 2008) which allows the drawing of bounding boxes around the regions of interest in the scene (see Figure 6.1), and the labelling of them using words. Notice that objects can be embedded into other objects, e.g. HEAD is part of the BODY (HEAD $<$ BODY). Setting the granularity of embedding influences the mapping of eye-movements. A coarse level assigns eye-movements to the object with larger area, e.g. BODY, whereas a fine level assigns eye-movements to the object with smaller area, e.g. HEAD. A finer level has more objects, thus there is a higher variability for each scan pattern. In our eye-movement analysis, we consider the highest level of embedding favoring always the mapping on objects with smaller area. The polygons are used to map the x-y fixation coordinates into the corresponding labels.

Across all set of images, we have a mean number of objects labeled of 28.65 with a standard deviation of 11.30. More than one object can be labeled with the same linguistic referent, e.g. *man*. We make a distinction for the ambiguous visual referents corresponding to the cue given in production [2] (MAN-L vs MAN-R). Then, a scan pattern

[1] Only similarities between sentences and scan pattern generated within the same scene are considered.

[2] By distinguishing ambiguous referents, we introduce more variability on the sequence alignment. The more labels the less likely it is to find matching during alignment.

Figure 5.1: Example of scene and cues used as stimuli for the description task. Each scene has been fully segmented into polygons, drawn around visual objects, using the LabelMe toolbox (Russell *et al.*, 2008). Then, each polygon has been annotated with the corresponding linguistic label.

is represented as a discrete sequence of temporally ordered fixated labeled objects (see Figure 5.2). The data varies across participants and scenes both in terms of the complexity of the sentences (i.e., *one man waits for another man to fill out the registration form for a hotel* vs. *the man is checking in* for Figure 6.1) and in the length of the scan patterns produced both in preparation for production (min = 800 ms; max = 10205 ms) and during production (min = 2052 ms; max = 18361 ms). This variability is taken into account by using metrics for sentence and scan pattern similarity which do not need the sequences to be normalized on their length. Moreover, we explicitly test the effect of time for the scan pattern data by including it in one of the measures used to compute similarity.

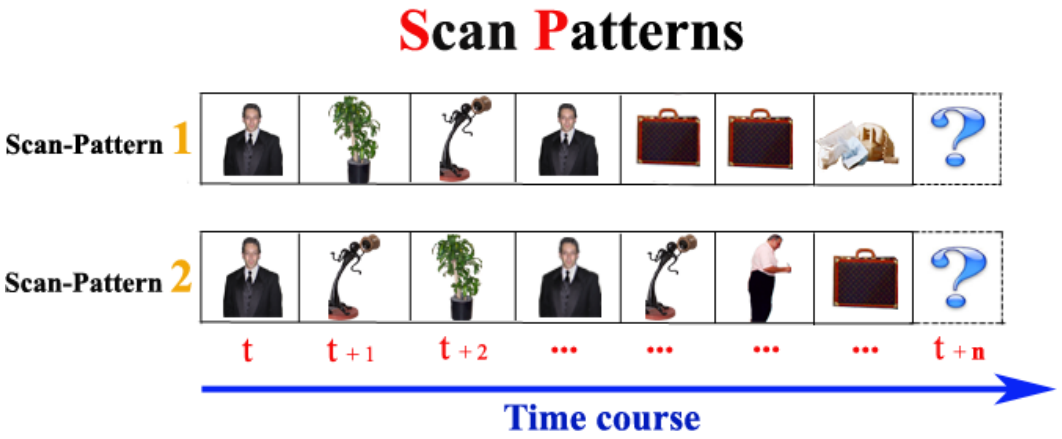


Figure 5.2: Each scan pattern is represented as a sequence of temporally ordered fixated objects. The fixation coordinates are mapped into the corresponding objects by using the labeled polygons.

5.3.2 Similarity Measures

Before quantifying the association between scan patterns and sentence productions, we measure similarity within each modality. We defined two similarities for both modalities. Applying more than one measure makes it less likely that our results will be an artifact of the type of measure used. The similarity measures are calculated using sequence analysis, i.e., **L**ongest **C**ommon **S**ubsequence (LCS, Gusfield 1997). Moreover, for sentences only we use **L**atent **S**emantic **A**nalysis (LSA, Landauer *et al.* 1998) a measure of semantic distance based on vector representation. We begin summarizing the measures of sequence analysis applied to calculate similarity both in sentences and scan-patterns, which are extensively explained in Chapter 2; and finish the section discussing how LSA is calculated on sentences.

5.3.2.1 Sequence Analysis

Both sentences (words) and scan patterns (fixated objects) are sequential data. Similarity between sequences can be found by looking at the information shared when aligned (Durbin *et al.*, 2003). We implement two simple measures of sequence analysis, Longest Common Subsequence (LCS, Gusfield 1997) and Ordered Sequence Similarity (OSS, Gomez & Valls 2009), which allow us to compute similarity between

sequences of different lengths while taking into account relative distance between common elements. A recent application of sequence analysis (Needleman-Wunsch) to eye-movement data, which similarly tackles the issues described above, has been proposed by Cristino *et al.* (2010) and implemented as a Matlab toolbox (ScanMatch); see Chapter 2 for a more detailed discussion. Since we found high correlation between results obtained using LCS compared to those found with ScanMatch ($\rho \approx 0.98; p < 0.001$), we restrict the discussion of our results to the simpler method LCS.

In LCS, we search the longest subsequence common to two sequences by iteratively exploring the space of all possible subsequences. Once we find the longest subsequence, we calculate the similarity score as the ratio between the length of the LCS and the geometric mean of the two sequences.

The second method used is Ordered Sequence Similarity OSS which is based on two aspects of sequential data: the elements the sequence is composed of and their positions (Gomez & Valls, 2009). We calculate the relative distance on the elements shared by the two sequences compared, and integrate this result with the number of uncommon elements. Finally, we normalize the obtained measure on the basis of sequence lengths (refer to Chapter 2 for more details).

5.3.2.2 Compositional model of semantics: LSA

A computational approach widely used to calculate lexical meaning of individual words is LSA (Landauer *et al.*, 1998). LSA measures the similarity between words based on the co-occurrence of content words within a collection of documents (in our case the British National Corpus). It indicates how likely two words are to occur in the same document. Different from Hwang *et al.* (2009) where LSA is calculated between individual words, we implemented a version of LSA generalized to compute the similarity of sentences (Mitchell & Lapata, 2009). In this approach, the meaning of a sentence is represented as the composition of the individual words forming it. We compute an LSA vector for each content word in the sentence (context window of size five; low frequency words are removed) and then combine these vectors using addition to obtain a sentence vector (an alternative discussed by Mitchell & Lapata 2009 would be vector multiplication). Similarity between sentence vectors is measured using cosine distance. It is important to stress that we use an implementation of LSA that does not

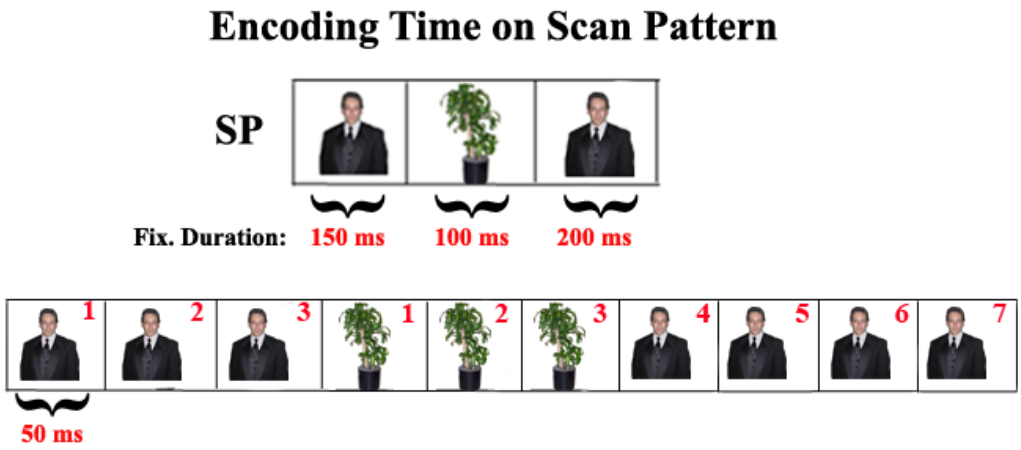


Figure 5.3: Encoding information about fixation duration into scan pattern.

take into account the order of words in the sentence[1], which makes it a non-sequential similarity measure.

5.3.2.3 Measures

Sequence analysis (LCS and OSS) has been applied to both scan pattern and sentence data. Especially, LCS was used on sentences (LCS.L) where low frequency words are removed, and on scan patterns without time information (LCS.V). OSS has been used only on scan-patterns including time information (OSS-TIME). Time was included, in slices of 50ms, by re-labeling the objects along the duration of fixation with an increasing numerical index. For example, if MAN was fixated for 150ms, we spread the object in three different slots, 50ms each, re-labeling it with an increasing index (man-1,man-2,man-3; see Figure 5.3 to visualize it). OSS has also been tested without temporal information on scan patterns and words, giving similar results of LCS. In order to simplify the discussion, however, we only report results with OSS-Time[2]. LSA was computed only on sentences.

[1] In Mitchell & Lapata 2008 is suggested how to include sequentiality of words in the LSA measure.

[2] LCS has also been tested on scan patterns enriched with temporal information. Nevertheless, we preferred to report OSS, because it includes relative distance between common elements, hence accounting more precisely for temporal difference in the scan patterns.

5.4 Analysis

To analyze the coordination between sentences and scan patterns, we divide the data into three regions, which correspond to the different phases of the language generation task: *Planning*, *Encoding* and *Production*. The *Planning* region considers from the onset of the image until the target is found, and it captures the endogenous control taking place during visual search. The *Encoding* region considers from the first fixation to the target object until description begins, and it describes the visual information retrieved in support to sentence encoding. Finally, the *Production* region is from beginning to end of the description, and it refers to cross-modal integration, when linguistic and visual information are combined. For each region of analysis, all measures of similarity are computed pairwise, i.e., every trial (sentence and scan pattern) is paired with every other trial. This resulted in a total of 382,973 pairs.

We perform two types of analysis: descriptive and inferential. In the descriptive analysis, we investigate the data at two levels: (1) globally, by performing comparisons between all pairs of trials in the full data set, and (2) locally, by comparing only the trials that pertain to a given scene (24 in total). These two forms of analysis make it possible to test whether the coordination between sentences and scan patterns is scene specific. For comparison, we also report a baseline correlation (Humphrey & Underwood, 2008) that is obtained by pairing sentences and scan patterns randomly (rather than pairing the scan patterns with the sentences they belong to). We quantify the strength of the correspondence between similarity measures by computing Spearman's ρ for all pairs of measures. We do not report coefficients for the baselines, as they are not significant across all combined measures: $\rho \approx 0.002; p > 0.1$. The distinction we made between global and local similarity has implications for the nature of coordination. A correlation found globally (across all scenes) would imply that scan patterns are partially independent from the precise spatial configuration of the scene, e.g. position or size of the objects, as these factors varied across scenes, but rather dependent on the referential structure shared, i.e. the visual referents common across scenes. A correlation found at the local level would be consistent with well-known scene-based effects, both bottom-up and top-down, which guide visual attention (Humphrey & Underwood, 2008; Itti & Koch, 2000b).

Global and local coordination are further explored by looking at the density distribution of cross-modal similarity, which is an aggregated measure obtained summing visual (e.g., LCS.V) and linguistic (e.g., LCS.L) similarities, and normalizing it to range between 0 and 1. We compare the distributions of cross-modal similarity on two groups: *within* scene and *between* scenes. Notice, this is a slightly different way to divide the data. Global contained all comparisons between and within scenes; whereas with local we select only comparisons on the same scene. Here, we separately consider all the between and within comparisons. In the within scene analysis, we only consider similarities of sentences and scan patterns generated in the same scene; whereas for the between scene analysis, only similarities between different scenes. We focus on two cross-modal similarity measures, one based on sequential similarity only (LCS.V + LCS.L; acronym: CROSSSEQ), and the other one combining LSA on sentences and OSS-Time on scan pattern with temporal information (acronym: CROSSTIME). For analysis involving cross-modal similarity we only consider pairs with a similarity score greater than 0, thus excluding those sentence/scan pattern pairs which were completely dissimilar. It is important to notice that we find completely dissimilar pairs only on the sequential cross-modal similarity (43%). The main reason is that LSA, in our set of sentences, gives always similarity scores that are greater than 0 (min = 0.0346). Moreover, the inclusion of time on scan pattern hugely increases the chance of finding alignments compared to when time is not included[1]. Cross-modal similarities are expected to have higher mean within the same scene compared to across scenes. Within the same scene the referential information is identical, i.e. the same objects are present, and there is more chance that similar sentences or scan patterns are generated. Across different scenes, the referential overlaps sensibly vary, e.g. a kitchen and a bathroom have fewer objects in common than two kitchens, thus there is less chance that similar sentences or scan patterns are generated.

In order to unravel the role played by visual and linguistic factors on the coordination, we analyze the impact of *animacy* of target referent and *clutter*, extensively explored in Chapter 4, on the cross-modal similarity measures. Since our similarity scores are calculated pairwise, for each explanatory variable with two-levels (e.g.

[1]OSS-Time gives a similarity score of 0 only 0.0002% across all pairwise comparison, whereas for the LCS.V we observe a striking 49% of zero cases.

Cue: Animate/Inanimate), we can have three different cases of pairwise combination: in two cases both trials have the same level (e.g. Animate/Animate or Inanimate/Inanimate), and a third case where the trials compared have different levels (e.g. Animate/Inanimate or Inanimate/Animate). We use a simple contrast coding to define these three cases both for *Cue* and *Clutter*. We contrast same levels, e.g. Ani/Ani, with different levels [1], e.g. Ani/Ina.

In the inferential analysis, we apply linear mixed effects modeling (LME) (Baayen *et al.*, 2008). In the first set of models, we assess the relation between the different measures of sentence and scan pattern similarity across the three phases of the task. We use sentence similarity as the dependent variable (fitting a separate model for LCS.L and LSA). Scan-pattern similarity (LCS.V and OSS-Time) and task phases [2] (planning, encoding or during) are the predictors. In the second set of models, we evaluate the impact of the explanatory variables, animacy and clutter, on cross-modal similarity across the different phases of the task. We use cross-modal similarity as the dependent variable (fitting a separate model for CrossSeq and CrossTime), *Cue*, *Clutter* and task's phases as predictors. For both sets of models, we included participants and trials as random effects [3].

All fixed factors were centered to reduce collinearity. The models are built following a forward step-wise procedure. We start with an empty model, then we add the random effects. Once all random effects have been evaluated, we proceed by adding the predictors. The parameters are added one at time, and ordered by their log-likelihood improvement of model fit: the best parameter goes in first. Every time we add a new parameter to the model (fixed or random), we compare its log-likelihood against the previous model. We retain the additional predictor if the log-likelihood fit improves significantly ($p < 0.05$). The final model is therefore the one that maximizes the fit with the minimal number of predictors (see Chapter 2 for details).

[1] We don't make any distinction on whether the different levels of the pair are Inanimate/Animate or Animate/Inanimate as this difference does not have theoretical implications.

[2] We use a simple contrast coding for the phases, where encoding and during are contrasted with planning, which is then incorporated at the intercept.

[3] Similarity is calculated pairwise. Thus, we need to include, as random variables, two participants and two trials for each pair.

5.5 Results and Discussion

The first section of results is dedicated to the discussion of observed data. We show plots, means and confidence intervals of visual similarities measures binned [1] as a function of linguistic similarity. In this context, we discuss the correlation coefficients obtained across all combined measures of visual and linguistic similarity. We also analyze how the strength of cross-modal coordination changes whether similarities are calculated between different scenes or within the same one. In the second section, we report inferential LME analysis testing the relation between linguistic and visual similarities across the different phases of analysis. Moreover, we examine the impact of clutter and animacy on cross-modal similarity, across the task's phases.

5.5.1 Descriptive analysis

In this section, we explore the relation between visual and linguistic similarity by showing plots of observed data and reporting coefficients of their correlation at two levels of analysis: global, i.e., across all sets of comparisons, and local, i.e., only comparisons belonging to the same scene.

Global analysis Figure 5.4 plots the linguistic similarity measures LCS.L and LSA against the scan pattern similarity measure LCS.V and OSS-Time, computed globally, i.e. across all scenes. We bin the data on the x-axis and include 95% confidence intervals. The plots also include the random baseline.

For both types of linguistic similarity (LCS.L, LSA) we observe a clear trend between sentence and scan pattern: when LCS.L or LSA similarity increases, scan pattern similarity (LCS.V) also increases. Even including temporal information (fixation duration) on the scan-pattern similarity (OSS-Time), we observe the same trend, i.e. visual similarity increases along with the increase of linguistic similarity. This effect is consistently observed across all different regions (Planning, Encoding, Production), but not in the random baselines. However, the strength of correlation varies across regions and combination of metrics used.

[1] Ten bins in intervals of 0.1.

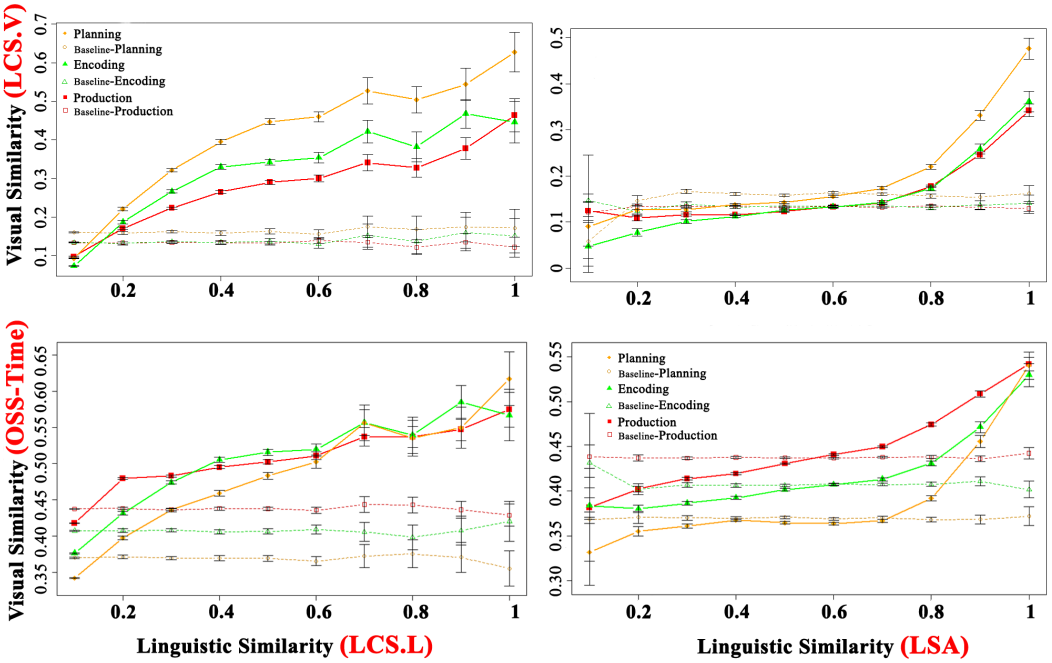


Figure 5.4: Correlation between linguistic (LSA, LCS.L) and visual similarity (LCS.V, OSS-Time)

Similarities computed using sequence analysis, both on visual and linguistic information, give the strongest correlations ($Region_{Planning}$: $\rho_{LCS.V/LCS.L} = 0.48; p < 0.05$) compared to a combination of metrics ($Region_{Planning}$: $\rho_{LCS.V/LSA} = 0.1; p < 0.05$); see Table 5.1 for the complete list of global correlations. When comparing the mean coordination across the different regions, we observe that the more linguistic processing is involved the less coordination occurs, i.e., Planning has overall the highest similarity, and strongest correlation coefficient. Probably, at the beginning of trial, a target search is launched (Planning) and the visual system is uniquely controlled by object-based allocation of visual attention (Henderson, 2007; Malcolm & Henderson, 2010). When, however, linguistic processing begins to select the visual referents forming the sentence (Encoding), the allocation of visual attention narrows around the visual information crucial to build the sentence. When production starts (Production), visual attention integrates the information retrieved to the ongoing linguistic production. When time is included on scan-pattern similarity (OSS-Time) and it is paired to the sequential similarity of linguistic information (LCS.L),

Table 5.1: Correlations (Spearman ρ) between the different similarity measures across regions of analysis: **Plan**ning, **Enc**oding and **Prod**uction

Measures	LCS.V			OSS-Time			LSA		
	Plan	Enc	Prod	Plan	Enc	Prod	Plan	Enc	Prod
OSS-Time	0.82	0.80	0.77						
LSA	0.1	0.11	0.13	0	0.11	0.22			
LCS.L	0.48	0.47	0.38	0.34	0.39	0.35	0.36	0.35	0.38

we still find a positive correlation[1] with a similar coefficient across all different regions ($Region_{Planning} : \rho_{OSS-Time/LCS.L} = 0.34; p < 0.05$). However, when the correlation is obtained by pairing OSS-Time and LSA, the strength of coefficient varies across the different regions of analysis. Especially, it seems to increase along the involvement of linguistic processing: from absence of correlation during Planning ($Region_{Planning} : \rho_{OSS-Time/LSA} = 0; p > 0.1$) to a weak correlation during Production ($Region_{Production} : \rho_{OSS-Time/LSA} = 0.22; p < 0.05$). LSA is based on lexical statistics about the words composing the sentence. LCS instead, more generally, focuses on sequential co-presence, thus disregarding any relational information among words. When scan-pattern includes temporal information, we enforce similarity to be time-locked; this constraint makes LSA be correlated only when the scan-pattern similarity temporally refers to it, that is during Production. The same effect does not emerge when linguistic similarity is sequential (LCS). In such a case, the similarity between two sentences is driven by the simple co-presence of referents (e.g. woman), which can be easily shared in the corresponding scan patterns.

Local analysis Figure 5.5 plots local similarity values, i.e., values computed separately for each scene (LCS.V against LCS.L)[2]. Generally, the trend previously observed at the global level is confirmed, across all regions, though we observe variation in the strength of correlation between scan patterns and linguistic similarity across

[1] The coefficients are slightly weaker.

[2] For conciseness, we show only one pair of combined measures, LCS.V/LCS.L. However, we observe a similar trend for all the other pairs.

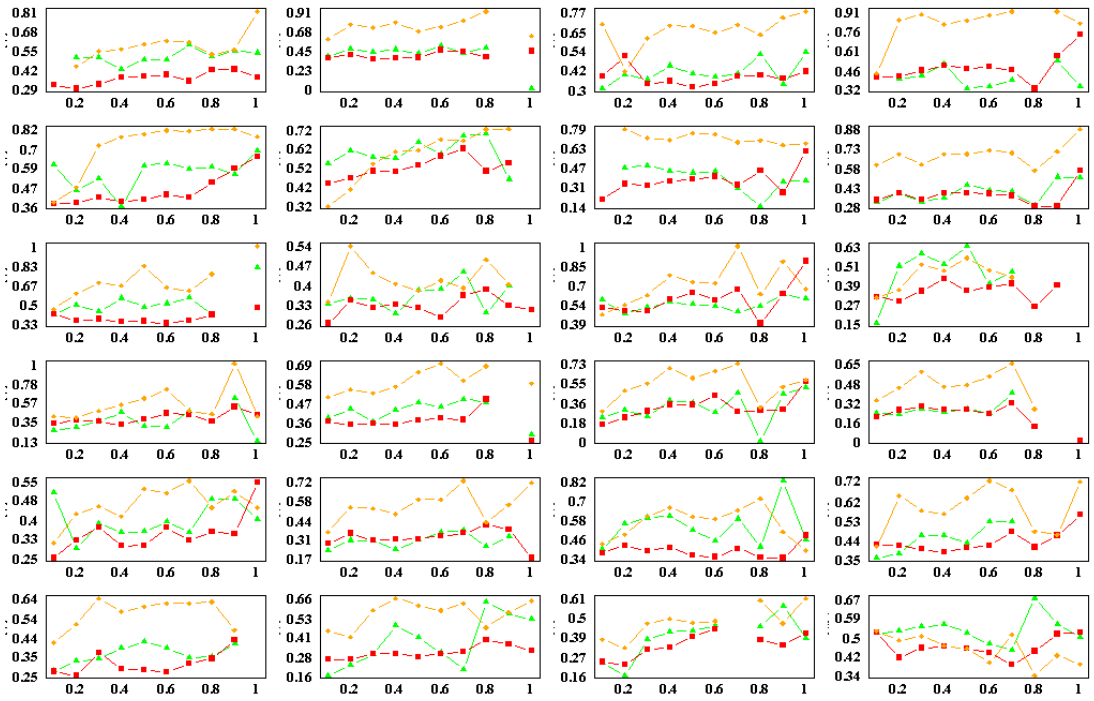


Figure 5.5: Scan pattern similarity (LCS.V) as a function of the Linguistic Similarity (LCS.L) across all 24 scenes

scenes. Moreover, also locally, the more linguistic processing (Production) is involved, the less overall similarity emerges.

Table 5.2: Mean and standard deviation of correlations (Spearman ρ) across scenes between similarity measures for the different regions of analysis

Measures	LCS.V			OSS-Time			LSA		
	Plan	Enc	Prod	Plan	Enc	Prod	Plan	Enc	Prod
OSS-Time	0.81±0.05	0.79±0.07	0.77±0.1						
LSA	0.09±0.1	0.1 ±0.14	0.08±0.1	0.03±0.13	0.07±0.17	0.11±0.13			
LCS.L	0.47±0.1	0.46±0.17	0.34±0.11	0.33±0.1	0.39±0.16	0.31±0.11	0.35±0.11	0.36±0.13	0.35±0.1

In Table 5.2, we show mean ± standard deviation of the correlation coefficients for similarity measures observed locally, i.e. aggregated by scene. As expected from the plots in Figure 5.5, correlation coefficients vary across scenes for all pairs of measures. The context of the individual scenes modulates the coordination between scan patterns and linguistic productions. In line with the results observed at the global level, co-

ordination based on sequential similarity (LCS.V/LCS.L) reaches highest correlation compared to combination of metrics (LCS.V/LSA). Moreover, when time is included (OSS-Time), and it is correlated with the linguistic similarity based on vector semantics (LSA), coordination is observed during linguistic production, where it reaches the maximum correlation.

Cross-modal Similarity: Between Vs Within scenes The coordination between visual and linguistic similarity is found both at a global, i.e. across all comparisons, and a local level, only within scene comparison. However, we don't know yet how important it is for the coordination to be exactly within the same scene compared to being between scenes[1]. In order to have a unique measure of cross-modal similarity, we sum and normalize the similarity scores obtained independently by the visual and linguistic measure. Since, during the descriptive analysis, we have observed that the correlation changes according to the combination of similarity measures considered, and whether temporal information was included or not, we calculate two measures of cross-modal similarity. One purely sequential (CrossSeq), which it is obtained by summing[2] and normalizing LCS.L and LCS.V, and the other one (CrossTime), instead, obtained by combining LSA and OSS-Time. Here we have finer temporal information on the scan pattern, and a more lexicalized measure of sentence similarity.

In Figure 5.6, we plot the density of cross-modal similarities aggregated within, i.e. only trials from the same scene, and between scenes, i.e. only trials from different scenes. In the upper panel (CrossSeq), we observe that within scene cross-modal similarity is likely to be normally distributed with an overall higher mean (≈ 0.3) than between scenes (≈ 0.1), which instead presents a right skewed distribution. It is interesting to notice how the distribution changes, especially for the between case, when the other cross-modal similarity measure (CrossTime) is observed. Now we observe a normal distribution rather than right skewed also for the between case. Probably, the finer temporal and lexical resolution of the measures used in CrossTime allows us to capture more sensibly the differences between sentences and scan patterns. Regardless of the measure used, we notice that within the same scene we have a higher cross-modal

[1] Notice, global contained both between and within scene analysis.

[2] An alternative way to calculate cross-modal similarity would be the harmonic mean, which better account for the relative contribution of each similarity.

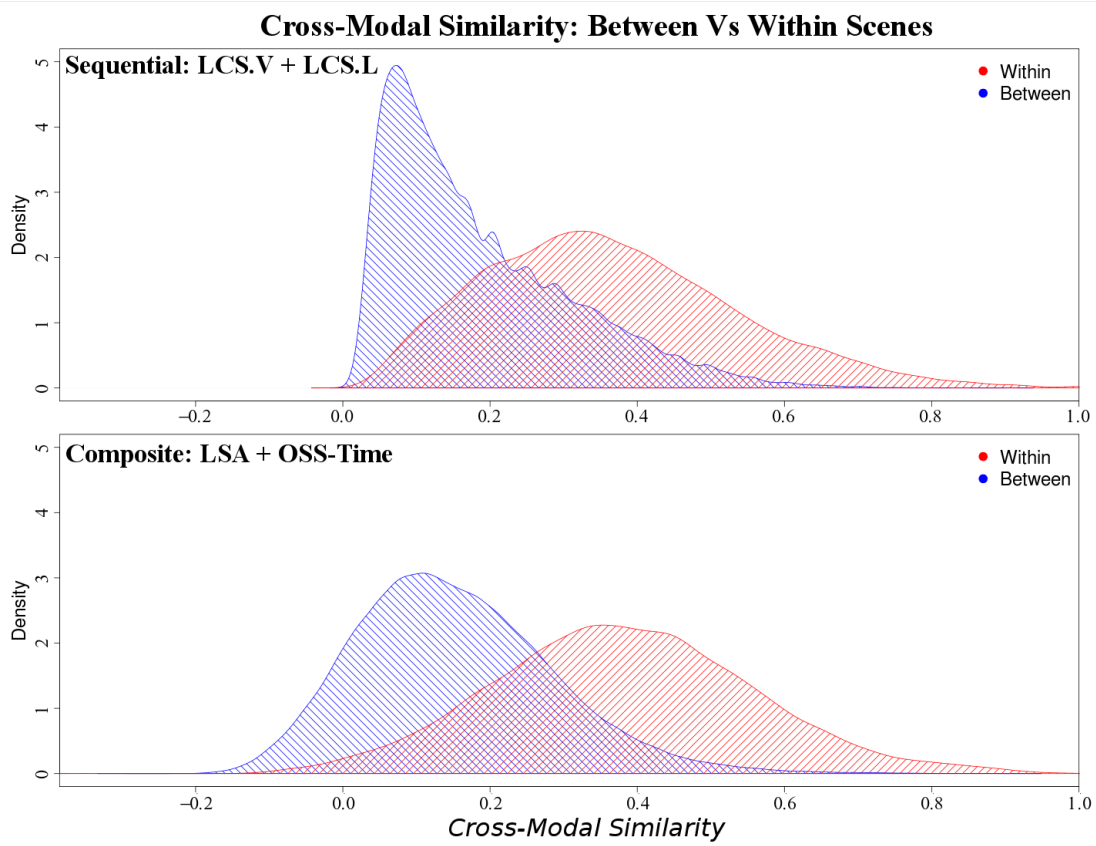


Figure 5.6: Density plot of cross-modal similarity. Cross-modal similarity is computed by summing the similarity scores obtained separately for the linguistic and visual measure and normalized to range between 0 and 1. In the upper panel, we show cross-modal similarity obtained aggregating the sequential similarity measures of LCS.L and LCS.V; whereas in the bottom panel we aggregate LSA and OSS-Time. The red line indicates cross-modal similarity within the same scene, whereas the blue line between different scenes.

similarity, compared to different scenes. Obviously, the more referential overlap there is, the more probable it is that similar sentences and scan patterns are generated. An interesting continuation for future research would be to look at the region of the distribution where cross-similarity between and within scenes overlaps. By investigating the referential information able to coordinate vision and language across different scenes, we should be able to 'rank' relevance of scene and sentence information.

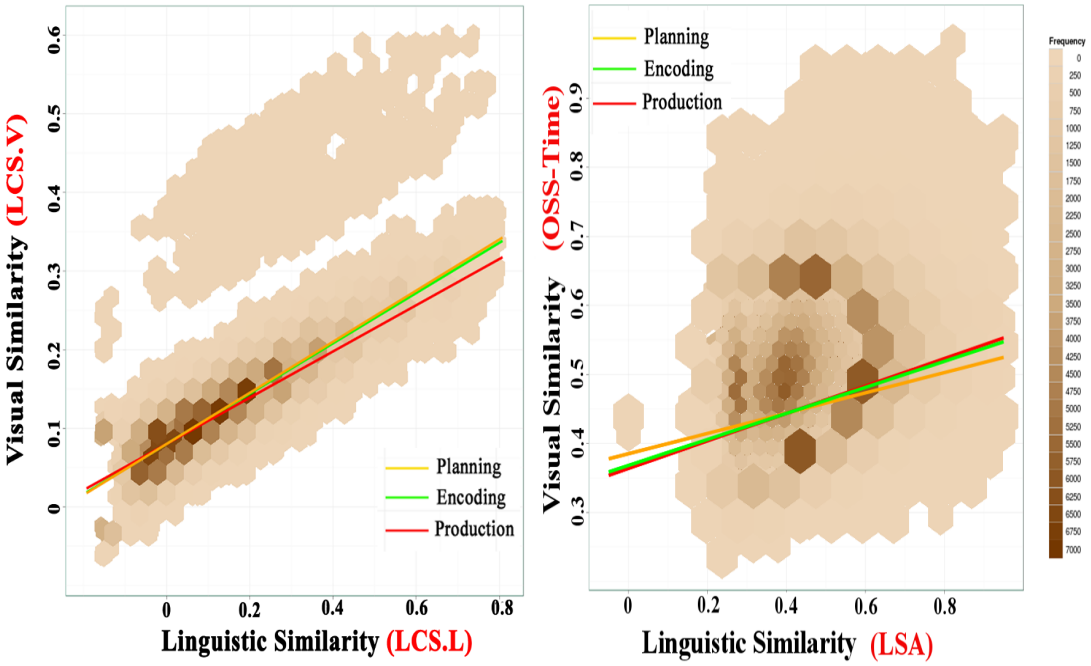


Figure 5.7: Hexagonal binning plots of predicted values of the linear mixed effects model: linguistic similarity predicted by scan pattern similarity and phases of the task. On the left panel, our dependent linguistic measure is LCS.L, and the scan pattern predictor is LCS.V; whereas on the right panel, the dependent measure is LSA, predicted by OSS-Time. The plot shows the observed data binned into hexagons. The colour of the hexagon reflects the frequency of observations within it: the more observations, the darker is the color. The solid lines overlaid represent the grand mean intercept for the different phases: Planning (orange), Encoding (green), Production (red).

5.5.2 Inferential analysis

In the first part of inferential analysis, we examine more closely, using linear mixed effect modeling, the global correlation between visual and linguistic similarities across the different phases of the task. In the second part, we explore the role played by clutter and animacy on cross-modal similarity, again across the different phases of the language generation task.

Patterns of global coordination Turning now to the inferential analysis, Figure 5.7 shows two plots of LME predicted values calculated globally for the sequential measures LCS.V/LCS.L (left panel) and combined measures LSA/OSS-Time (for the full list of LME coefficients refer to Table 5.3). We show plots for only these two pairs

of measures, as they represent the two most different combinations of similarity measures used. Both models closely follow the empirical patterns in Figure 5.4 with the observations being distributed in the expected positive direction: when visual similarity increases linguistic similarity also increases. However, when comparing the scatter obtained by using different measures, we observe some important differences. The first difference is that when only sequential measures without time (left panel) are combined, we find overall less similarity with most of the observations falling between 0 and 0.2 in the linguistic similarity (x-axis) and 0.1 in the visual similarity (y-axis). Instead, when LSA is used for sentences, and the scan patterns similarity measure includes temporal information (OSS-Time), we observe most of the observations falling around 0.4 on the x-axis and 0.5 on the y-axis. These results are consistent with Figure 5.6, where cross-modal similarity seems to have a higher mean, both within and between scenes, when LSA and OSS-Time were combined, compared to the other combination of measures.

In Table 5.3, we list the coefficients of the mixed models; as expected from Figure 5.7, we find a significant main effect of scan pattern similarity for both LCS.V and OSS-Time, for both the LCS.L and the LSA model. Moreover, confirming the correlation analysis, we observe a main effect of task's phase which is modulated by the inclusion of time, and the type of metrics combined. When time is not included and we are looking at sequential measures only, we find that, during *Planning*, sentence similarity is more strongly related to scan pattern similarity, compared to both *Encoding* and *Production* regions. By including time, instead, the coordination is strengthened during Production, especially when combined with LSA. Time seems to have a negative impact on coordination during Planning, compared to Encoding. When LSA is combined with LCS.V, we observe the same trend, but with smaller coefficients.

Furthermore, we observe interactions between region of analysis and scan-pattern similarity, which go along with the results observed for the main effects. When looking at sequential measures only, for the *Planning* region, the similarity between sentence and scan pattern has a steeper change, compared to *Encoding* and *Production*, where instead we observe a negative interaction. The inclusion of time favors Production, where similarity between sentences increases along with scan-pattern similarity, especially when LSA is the dependent measure used (see Figure 5.7 to visualize).

Table 5.3: LME coefficients. The dependent measures are: *LCS.L* and *LSA*.The predictors are: *Reg*ion (Planning; Production; Encoding, which is expressed at the intercept.) and the Scan Pattern (SP) LCS.V or OSS-Time. Each column shows which linguistic/scan patterns similarity measure is compared. **n.i.** stands for *not included* during model selection.

Predictor	LCS.L/LCS.V	LCS.L /OSS-Time	LSA/LCS.V	LSA/OSS-Time
Intercept	0.111^{***}	0.059^{***}	0.541^{***}	0.486^{***}
SP	0.311^{***}	0.433^{***}	0.087^{***}	0.135^{***}
Reg-Planning	-0.006^{***}	0.04^{***}	-0.006^{***}	n.i
Reg-During	-0.004^{***}	-0.08^{***}	-0.016^{***}	-0.100^{***}
SP:Reg-Planning	0.011^{***}	-0.089^{***}	0.018^{***}	-0.089^{***}
SP:Reg-During	-0.022^{**}	0.104^{***}	0.017^{***}	0.172^{***}

To summarize, it seems that coordination depends on the phase of task we are in, on the inclusion of temporal information, and partially on the combination of metrics used to quantify it. During each phase of the task, the interaction between vision and language is expected to change. In Planning, the visual system performs a search of the cued target object, by combining target properties and referential scene information (Castelhano & Heaven, 2010; Malcolm & Henderson, 2010). Here, sentence encoding hasn't started yet and visual attention is strongly driven by endogenous visual control. Thus, even if the scan patterns followed during the inspection of the scene contain similar sequences of visual referents to those found in the corresponding similar sentences, they vary on the temporal dimension. The visual material retrieved does not need yet to be integrated with linguistic processing. Thus, coordination is on the sequence of referents looked at, but it temporally varies across different participants. Once the target object has been identified, the linguistic processing begins controlling the visual system to retrieve visual material to be used during production. In Encoding, in fact, the referential information of the scene has to be integrated with the linguistic referents selected as arguments of the sentence. Coordination now depends more on the choices of linguistic processing, and visual endogenous control has to be integrated with linguistic control. Finally, in Production, the visual processor performs specialized routines of visual information retrieval uniquely based on the demands of sentence encoding. Here, linguistic control takes over the endogenous allocation of

visual attention. During this phase, time is crucial, as the linguistic material mentioned is time-locked with the visual referents scanned. In such a region, the combination of a linguistic measure of sentence similarity based on vector semantics (LSA) and a temporally informed measure of scan pattern similarity (OSS-Time) gives the best coordination, compared to the other regions.

Visual and linguistic factors on cross-modal similarity In Chapter 4, we explored the role of animacy and clutter during sentence production. We focused on the influence of these factors on the type of description generated (Experiment 4), and on the pattern of eye-movements observed during referential integration (Experiment 5), i.e. before and after the cued target object is mentioned. In line with previous psycholinguistic and visual cognition literature, we found that animate referents facilitate sentence production (Branigan *et al.*, 2008) and are inspected more frequently and more easily than inanimate referents (Fletcher-Watson *et al.*, 2008). Moreover, minimal visual information density facilitates target identification (Henderson *et al.*, 2009a). However, we also found several other interactions between visual information density and animacy of target. Especially, we observed that cluttered scenes are inspected during mentions of animate entities to support sentence encoding with contextually relevant visual referential information. Such a finding contrasts with previous accounts of the *eye-voice span* (Griffin & Bock, 2000), refer to Chapter 4 for details. Moreover, in contrast with visual search studies, where it has been found that the more cluttered a scene is, the more difficult is target identification, we find that inanimate targets are looked at more in cluttered scenes than in minimal scenes before they are mentioned. Here, we extend these results by looking at the impact that these factors have on cross-modal similarity, i.e., a measure to quantify the strength of referential association between pairs of sentences/scan patterns.

In Table 5.4 we report the LME coefficients estimates of two separate models for the cross-modal similarity measures, seen in section 5.5.1 (CROSSSEQ, CROSSTIME), as predicted by animacy, clutter, and phases of the task.

For both cross-modal similarity measures, we observe a main effect of animacy, where animate targets trigger more coordination than inanimate targets. Moreover, in minimal scenes we observe a better coordination than in cluttered scenes. When we look at the task phases, we find better coordination during the linguistic encoding in

Table 5.4: LME coefficients. The dependent measures are: *CrossSeq* (LCS.V/LCS.L) and *CrossTime*. The predictors are: *Region* (Production and Encoding are contrast coded with Planning, expressed at the intercept.); *Cue* (Animate/Animate, Inanimate/Inanimate are contrast coded with Animate/Inanimate—Inimate/Animate, expressed at the intercept); *Clutter* (Minimal/Minimal, Cluttered/Cluttered are contrast coded with Minimal/Cluttered—Cluttered/Minimal, expressed at the intercept); Each column shows which cross-modal similarity measure is used as dependent measure. **n.i** stands for *not included* during model selection.

Predictor	CrossSeq	CrossTime
Intercept	0.226^{***}	0.456^{***}
Animate	0.145^{***}	0.070^{***}
During	-0.054^{***}	0.009^{***}
Inanimate	-0.09^{***}	-0.045^{***}
Encoding	0.021^{***}	0.022^{***}
Minimal	0.023^{***}	0.011^{***}
Cluttered	-0.035^{***}	-0.016^{***}
Encoding:Minimal	-0.043^{***}	-0.052^{***}
Encoding:Cluttered	0.061^{**}	0.072^{***}
Animate:During	-0.11^{***}	-0.023^{***}
Animate:Encoding	-0.056^{***}	-0.013^{***}
Animate:Minimal	n.i	0.046^{***}
Animate:Cluttered	n.i	-0.065^{***}
During:Inanimate	0.073^{***}	0.007^{***}
Inanimate:Encoding	0.033^{***}	n.i
Inanimate:Minimal	n.i	-0.042^{***}
Inanimate:Cluttered	n.i	0.055^{***}

both measures. However, confirming what observed in section 5.5.1, during production we find that only CrossTime has positive coefficient. In fact, this measure includes temporal information (OSS-Time) and evaluates more precisely the semantics of sentences (LSA). During linguistic processing, visual attention and sentence processing are better coordinated.

When looking at the interactions, we find that during linguistic processing, i.e. encoding and production, there is more coordination in a cluttered scene than in a

minimal scene; especially when the target is inanimate. Moreover, comparing the two cross-similarity measures we find that we are able only with CrossTime to capture interactions between animacy of targets and visual density. Especially, we find that animate targets in cluttered scenes are less coordinated than inanimate targets, which instead show a positive relation with it. An animate target has to be contextualized within the scene, e.g. *the man is signing in*, thus the more referential information there is, i.e. high clutter, the more information can be inspected and utilized to describe such animate referent. An inanimate target, instead, has to be spatially located, e.g. *the suitcase is on the table next to the man*, and this requires a stricter and more precise selection of referential information to be integrated. Thus, the more visual information there is, the more precise coordination has to be to unambiguously situate the inanimate object within the scene.

5.6 General Discussion

A range of cognitive modalities are involved in everyday tasks, which raises the questions to which extent these modalities are coordinated. In chapter 4, we have seen that visual and linguistic factors of referential information have important influences on the way cross-modal information is integrated. In general, our hypothesis was that during task demanding synchronous processing, e.g. scene description, visual attention and sentence processing interact on a shared referential interface. This interaction has to be coordinated, i.e. visual attention retrieves referential information which has to be strongly associated with the sentence processing output.

In this chapter, we investigated the extent to which vision and language are coordinated, and how the visual, i.e. clutter, and linguistic, i.e. animacy, factors influence this coordination. In particular, we asked how visual attention (scan patterns) and linguistic processing (sentences) coordinate during description of naturalistic scenes. Our main hypothesis is that referential information about the scene, expressed as sequences of labels (words and objects), shared by vision and language, drives coordination. The main condition to observe coordination is to let the two modalities synchronously interact on a shared task: scene description. During interaction, the modalities have to be coordinated to allow integration between referential cross-modal information. We

can observe and quantify coordination by looking at cross-modal similarity: the more similar two sentences are, the more similar the corresponding scan patterns will be.

We tested this hypothesis using the dataset collected during the eye-tracking experiment (experiment 5) reported in chapter 4, in which participants had to describe photo-realistic scenes. Each sentence generated during the cued description of a scene is paired with the scan pattern that followed. Sentences and scan-patterns are sequential data, which develops over time. We used a simple method for biological sequence analysis (Longest Common Subsequence, LCS), to calculate similarity between sentences (LCS.L) and scan patterns (LCS.V). Time has been encoded in the scan pattern (fixation duration on objects), and similarity computed using a categorical measure of similarity (Order Sequence Similarity, OSS). Moreover, the similarity of sentences has been calculated using a compositional measure of vector semantics (Latent Semantic Analysis, LSA). Similarity is computed pairwise, i.e. each trial with every other trial.

Coordination has been analyzed over three task phases: planning (from scene onset until the cued target is found), encoding (from first fixation on target until beginning of utterance) and production (from start to end of scene description). Both descriptive and inferential analysis confirmed our hypothesis: if two trials involve similar scan patterns, then the sentences produced in these two trials are also similar, and vice-versa. This was true for all pairs of linguistic and scan pattern similarity measures, across all different region of analysis, at both global (across scenes) and local level of analysis (within the same scene). Significant correlations were found in both cases, which suggests that the coordination between sentences and scan patterns cannot be solely due to the referential information shared between an individual scene (objects) and the sentences (words) mentioned within it; but rather more global factors (visual and linguistic) shared across different scenes, are responsible for modulating the coordination. This conclusion is confirmed at the level of individual scenes, where the variability observed suggests that coordination can't be uniquely due to referential scene identity.

An important point that emerged during our analysis regarded the interaction between phases of task, temporal information and the methods used to calculate similarity. We found that sequential analysis on both vision and language (LCS.L, LCS.V) had the highest similarity correlation; getting smaller going from planning to production. During Planning, visual attention is purely driven by combined top-down information

of scene and target to predict the position of target. As visual attention is not yet synchronized with sentence processing, the time spent on individual objects can vary across different participants. This variability is not captured when the information of fixation duration is not included into the scan pattern representation. This explains the higher correlation of LCS.L and LCS.V found during planning, compared to LCS.L and OSS-Time. During Production, instead, temporal integration of cross-modal referential information is needed to synchronize the two modalities. In fact, we observe that when temporal information is included on the scan pattern (OSS-Time), production has a higher correlation and a steeper slope of change compared to planning. This is especially evident when OSS-Time is combined with LSA, i.e. a measure specifically designed to evaluate the semantic similarity of sentences; hence more suitable to capture a correlation during the act of sentence generation.

When looking at the impact of factors on cross-modal similarity, we found results in line with what was observed in experiment 5. Animate referents trigger more coordination than inanimate referents, especially when description is situated in a minimal scene. Moreover, the density of visual information shows significant interactions with the animacy of the target. When the target is animate, the denser the scene is, the more referential information can be used to contextualize the target within the scene. Therefore, cross-modal similarity is lower than in a minimal scene, as there are more ways to relate an animate referent with the surrounding context. However, when a target is inanimate, we observe an opposite effect. A cluttered scene triggers higher cross-modal similarity. An inanimate referent has to be spatially located, e.g. *the suitcase is on the counter next to the man*, therefore coordination has to be more precise on the referential information integrated, in order to unambiguously situate the inanimate object within the scene.

A limitation of the study is that it fails to explore more specifically how the co-occurrence of visual referential information, e.g. a mug is usually on counters, is associated with the corresponding sentence encoding. Moreover, a more stringent analysis of the aspects, syntactic, semantic, or contextual, involved in cross-modal similarity is needed.

Theories of active visual perception have provided significant evidence about endogenous top-down mechanisms guiding the allocation of visual attention during goal directed tasks. In this approach, the knowledge of a scene can be viewed in terms of

semantic relations between objects' references (e.g. mug, cup...). Referents have also a linguistic identity, which is found to be tightly coupled to the associated visual identity, when linguistic processing is situated in a visual context. We have unified these findings by showing that the endogenous allocation of visual attention (scan-patterns) is coordinated to linguistic processing (sentences). When vision and language interact on a task requiring cross-modal integration, i.e. scene description, they coordinate processing over a shared referential interface. Coordination is the key mechanism underlying the broader problem of multi-modal synchronous processing. By investigating it, we aim to discover more general cognitive mechanisms shared across modalities (e.g. vision and language), while beginning to explain cognition as a unified and integrated system.

5.7 Conclusion

During tasks demanding synchronous processing, vision and language have to be coordinated. It clearly emerges that the type of task, in which participants are engaged, plays a major role in the way modalities interact. In fact, depending on the type of task, we expect different levels of interaction between modalities. A visual search task, for example, requires contextual knowledge of the scene in order to quickly locate the target object, but once the target is found, no further processing is needed; whereas in a scene description task, once the target is found, visual attention and sentence processing have to tightly cooperate, in order to appropriately situate the target object in the context of the scene. Moreover, even across different situated sentence processing tasks, there might be sensible differences in the way cross-modal referential information is integrated. In a situated language comprehension task, there is a beginning phase of free viewing, where participants scan the scene trying to predict which objects are going to be mentioned, followed by a utterance mediated phase, where linguistic information is mapped against the referential information of the visual context. In a situated language production task, instead, visual attention has to actively retrieve referential information before sentence encoding can start.

In the next chapter, we explore how different tasks influence the pattern of cross-modal integration observed. Especially, we compare a visual search task, where participants are asked to find and count a cued target object in the scene, with the scene

description experiment previously reported. Our expectation is that the fewer modalities have to be synchronized, e.g. visual search, the less cross-modal interaction is observed. Moreover, we investigate how situated language production differ from comprehension. Here, we give participants a set of descriptions drawn from the production experiment, in order to allow a direct comparison between production and comprehension. We will be looking at the temporal differences found on the scan patterns when a sentence is mentioned compared to when is listened.

Chapter 6

The Influence of Task on Visual Attention: A Comparison of Visual Search, Object Naming, and Scene Description.

6.1 Introduction

The nature of the task defines the way cross-modal interaction is established. Each task entails different sub-goals, which determine the cognitive modalities engaged, and the pattern of their interaction. The task of grasping a mug, for example, implies the synchronous processing of visual attention and motor-action. Visual attention performs a search task to identify the MUG in the scene, while motor-actions are activated to direct the grasping action. Both modalities have to utilize top-down referential information to perform the task: scene information is used to predict the expected target location, while physical properties of the object, e.g. the MUG is full or empty, are used to plan the act of grasping.

Tasks can be distinguished by the degree of cross-modal interactivity required to perform them, that in turn directly relates to the way referential information is accessed and utilized. If we are just looking for a MUG but we have no intention of picking it up, there will be no interaction between motor-actions and visual attention. The absence of cross-modal interaction directly modulates the way referential information is utilized.

In fact, after the MUG is found, no further visual processing is needed to instruct motor actions.

In previous chapters, we have investigated situated language processing tasks, observing cross-modal interaction between visual attention and sentence processing. We found that a description, e.g. *the pen is on the table*, is generated by integrating information about the visual referents retrieved with the associated linguistic referents to be encoded. Referential information about scene (clutter) and objects (animacy) had direct implications on the patterns of visual attention and types of linguistic encoding observed.

We expect, however, a different pattern of referential information processing to emerge during single modality tasks, such as a visual search, where the goal is to find a cued target object embedded into a naturalistic scene. During search, visual attention utilizes referential information about the scene and target to build expectations about the locations where the target object is more likely to be found. The task ends when the target is found. During description instead, after the target is found, visual information is selected according to the underlying choices of sentence processor to produce a referring sentence. Search and description differ in the goals to be achieved, which in turn directly influence the access to referential information and its cross-modal integration.

To the best of our knowledge, it is yet unclear how the mechanisms of cross-modal interactivity modulate visual responses when purely visual, and linguistically driven tasks are performed. Thus, in this chapter we compare three different tasks (visual search, object naming and scene description), which vary in their degree of cross-modal interaction.

A search task implies only a shallow access to referential information, e.g. semantic relations between objects; thus the only possible interaction with sentence processing occurs at the level of scene context. Object naming and scene description, instead, demand a deeper cross-modal interaction on the visual and linguistic, referential information shared. In an object naming task, the scene is widely inspected to localize and name those visual referents which have also linguistic relevance; thus, it is a search task, but ‘weighted‘ by both visual and linguistic factors. And in scene description, only visual referents relevant to the sentence synchronously processed are going to be

observed; thus, visual attention and sentence processing have to tightly coordinate over the referential information selected to generate the description.

We hypothesize that in a task involving cross-modal interaction, i.e., scene description, participants should coordinate more on the way visual information is accessed, i.e. higher scan patterns similarity, compared to a search task. Moreover, such tasks should trigger a more complex visual processing, e.g. longer fixation duration, as visual referential information of the object fixated has to be integrated with the associated linguistic encoding.

This hypothesis is investigated in the context of the factors animacy and clutter already explored in previous chapters. We expect animate referents to be fixated longer when sentence processing is involved, as they carry a larger set conceptual structures than inanimate referents, and this boosts linguistic encoding. Moreover, an interaction is expected between clutter and tasks. In search and object naming, more clutter implies more difficult target identification, and more visual referents to name; whereas for description, clutter is a source of contextual information that facililates sentence encoding.

6.2 Background

The visual system is actively employed in many **tasks** of our daily life to support ongoing cognitive processes to achieve specific goals: e.g. finding a mug in a kitchen (Castelhano *et al.*, 2009; Findlay & Gilchrist, 2001; Henderson, 2003).

The active allocation of visual attention is driven by **top-down** knowledge-based information processing, which in this thesis has been more specifically discussed in terms of referential information processing. Thus, for example, a referent context KITCHEN usually includes other referents such as TABLE, PLATE, MUGS, etc. The relations, spatial, semantic or statistical, connecting referential information are assumed to guide visual attention, especially in goal directed tasks. In fact, only when precise goals have to be achieved, referential information has to be utilized.

During a search task, for example, referential information about target and scene is combined to predict the location of a cued target (Malcolm & Henderson, 2009; Neider & Zelinsky, 2006). If the target object is a MUG, and the scene context is a kitchen; expectations about related referents TABLE or COUNTER are activated to efficiently

allocate attentional resources (Brockmole & Henderson, 2006; Malcolm & Henderson, 2010; Torralba *et al.*, 2006). When a task doesn't entail any specific goal instead, e.g. free-viewing, bottom-up processes, which are based on the stimuli per se, e.g. saliency (Itti & Koch, 2000b), are more likely to steer attentional mechanisms. In the absence of goals, attention is captured by more general image-based information, rather than specific object-based referential information (Nuthmann & Henderson, 2010).

It is important to note that when referential information is accessed during a visual goal-directed task (e.g. search), only visual processing will be actively engaged. However, in previous chapters, we have observed that during situated language processing tasks, e.g. description, visual attention and sentence processing interact over a shared cross-modal referential interface; i.e. visual attention sources referential information about the scene according to the linguistic choices of encoding. In such a case, the access and integration of top-down referential information is mediated by the cross-modal interaction of vision and language, which have to synchronously cooperate during the course of the task. Obviously, cross-modal tasks are expected to have a different impact on referential information processing, compared to tasks where such cross-modal interaction is not present.

Research in visual cognition has compared visual tasks, finding that they differ on several eye-movement measures (Castelhano *et al.*, 2009). In Castelhano *et al.* 2009, search and memorization (memorize the scene in preparation for a recall phase) were compared. The authors found clear differences on spatial measures, e.g. total area of a scene inspected; but unclear evidence on the temporal measures, where they find longer total fixation duration during memorization than search (more time inspecting the scene across trials), but a similar mean gaze duration on the individual objects inspected, which indicates an overall similar temporal processing of object information.

In order to understand the reasons behind this difference, we contextualize the interpretation of eye-movement measures within two components of visual attention: spatial and temporal.

At the spatial component, we observe operations of referential **inspection**, i.e. which objects form the scene; and it is quantified by measures of spatial fixation distribution, e.g. the number of regions inspected in a scene. In relation to the task: the more the task requires a broad access to scene information, the more the spread of fix-

ation across the scene, i.e. more regions are fixated[1]. In memorization, more objects are inspected, compared to search where attention is allocated only on regions relevant to the cued target (Castelhano *et al.*, 2009).

Regarding the temporal component, instead, we can observe operations of referential **integration**, i.e. how much information is accessed on each object; this is quantified by measures of fixation duration. In memorization, during the first fixation visual and linguistic features [2] of the object fixated have to be accessed and memorized; whereas in search, the objects are fixated to verify their identity with respect to the cued target (Malcolm & Henderson, 2010). Therefore, the first fixation is found to be longer in memorization than search (Castelhano *et al.*, 2009).

Both memorization and search involve only visual attention. Thus, after the target is visually processed, e.g. memorized or verified, no further processing is needed; and this might explain the similarity of mean gaze duration found between memorization and search (Castelhano *et al.*, 2009): all fixations launched to target objects, after the first fixation, did not require any further processing of visual information. However, when the task demands cross-modal interaction, e.g. scene description, visual attention synchronously interacts with sentence processing, and this interaction modulates its spatial and temporal allocation; i.e. target objects are fixated in relation to the underlying linguistic processing, which changes during the course of the trial.

In chapter 4 and 5, we have investigated the cross-modal interaction of vision and language during a scene description task. We have shown that the pattern of fixation to the visual referents mentioned in the sentence are influenced by non-linguistic properties of the target, i.e. animacy, and scene, i.e. clutter. Furthermore, more generally, we have shown that cross-modal referential information is coordinated, i.e. similar sentences are associated with similar scan patterns. In order to complete our investigation on cross-modal referentiality, we explore how its activation varies across tasks demanding different degrees of cross-modal interaction.

In this chapter, we compare three different tasks, search, naming and description, which vary by the cross-modal interactivity required to perform them; with (1) search,

[1] Spatial access can also be measured by looking at the spread of eye-movement distribution over the scene (Pomplun *et al.*, 1996); more details in section 6.3.3.2.

[2] Especially if the recall phase was done on a word identifying the target object, rather than its visual appearance.

expected to activate only visual attention, (2) naming, demanding also a partial activation of sentence processing, and (3) description, requiring visual attention and sentence processing to be highly synchronized, as visual retrieval sources information for linguistic encoding. Our main hypothesis is that the more referential information has to be integrated across modalities, the more visual processing is needed (leading to, e.g., longer fixation durations). The cross-modal interactivity is expected to impact both spatial and temporal components of visual processing; with an emphasis on the temporal measures after the first fixation, e.g. mean gaze duration.

In particular, in the spatial component, we expect wider spatial distibution of fixations in search and naming compared to description. In description, visual attention focuses on the objects to the description; whereas in search, fixations are launched to find as many objects as possible corresponding to the cue, and in naming the linguistic relevance of most of objects forming the scene has to be evaluated, in order to select the most relevant targets to be named. On the temporal component, we expect longer fixation duration when visual attention interacts with sentence processing, as referential information has to be integrated across modalities.

We divide our analysis in two different experiments (7 and 8). In experiment 7, we compare search and description, where beside animacy and clutter, we also investigate the impact of number of targets (1,2,3). In experiment 8, we compare naming, with a subset of the data collected in search and description. We keep the conditions of animacy and clutter, but we focus on cases where the number of targets is two across all scenes.

6.3 Experiment 7: Visual search and scene description

In experiment 7, we compare the influence of cross-modal interaction on two tasks, *Search* and *Description*. We observe how spatial and temporal components of visual attention are influenced by the cross-modal interactivity of the task, and how it interacts with properties of the target and the scene. Our main hypothesis is that the cross-modal interactivity of a description task would trigger more visual processing, e.g. longer fixation duration, compared to search, where only visual attention is actively engaged. Especially, description should demand more inspections and longer fixations compared to search. Moreover, in line with previous literature, animate targets should facilitate

search, e.g. shorter fixation (Fletcher-Watson *et al.*, 2008); but it would require more processing during description, e.g. longer fixation, to integrate referential information across modalities. The clutter of the scene is expected to impair search and interact with animacy of the target (Henderson *et al.*, 2009b), e.g. inanimate objects would be more difficult to locate than animate ones, especially if the scene contains only one target.

6.3.1 Design and Material

The experimental setting, identically designed for both experiments, crossed three factors: *Cue* (Animate/Inanimate), *Target* (1,2,3) and *Clutter* (Low/High) (see Figure 6.1).

Each scene contained either 1, 2 or 3 visual *Targets* corresponding to the cue. The *Cue* was either Animate, i.e. person, or Inanimate, i.e. laptop, and for more than 1 Target, it was referentially ambiguous in respect with the scene, i.e. three PERSON depicted. Moreover, we computed the Feature Congestion (Rosenholtz *et al.*, 2007) (*Clutter*) visual density measure, and used it to divide the scenes in two density classes (Low/High). We created 72 experimental items using photo-realistic scenes drawn from 18 different scenarios (four scenes per scenario): 9 indoor (e.g., Bathroom, Bedroom), 9 outdoor (e.g. Street, Mountain). In each scene, we inserted the objects (animate and inanimate), which correspond to the two Cue conditions, for the different Target condition (1,2,3) using Photoshop.

6.3.2 Method and Procedure

In both search and description, we cued participants with a word corresponding to the target object, i.e. laptop, and we asked them to **describe** (see Figure 6.1 for an example of a sentence, and refer to chapter 2 for details) the target in the context of the scene (description), or **count** how many instances of the target were in the scene (search). Forty-eight (24 per task) native speakers of English, all students of the University of Edinburgh, were each paid five pounds for taking part in the experiment. Each participant saw a randomized list of the 72 trials. An EyeLink II head-mounted eye-tracker was used to monitor participants eye-movements with a sampling rate of 500 Hz. Images were presented on a 21" multiscan monitor at a resolution of 1024 x 768 pixels.

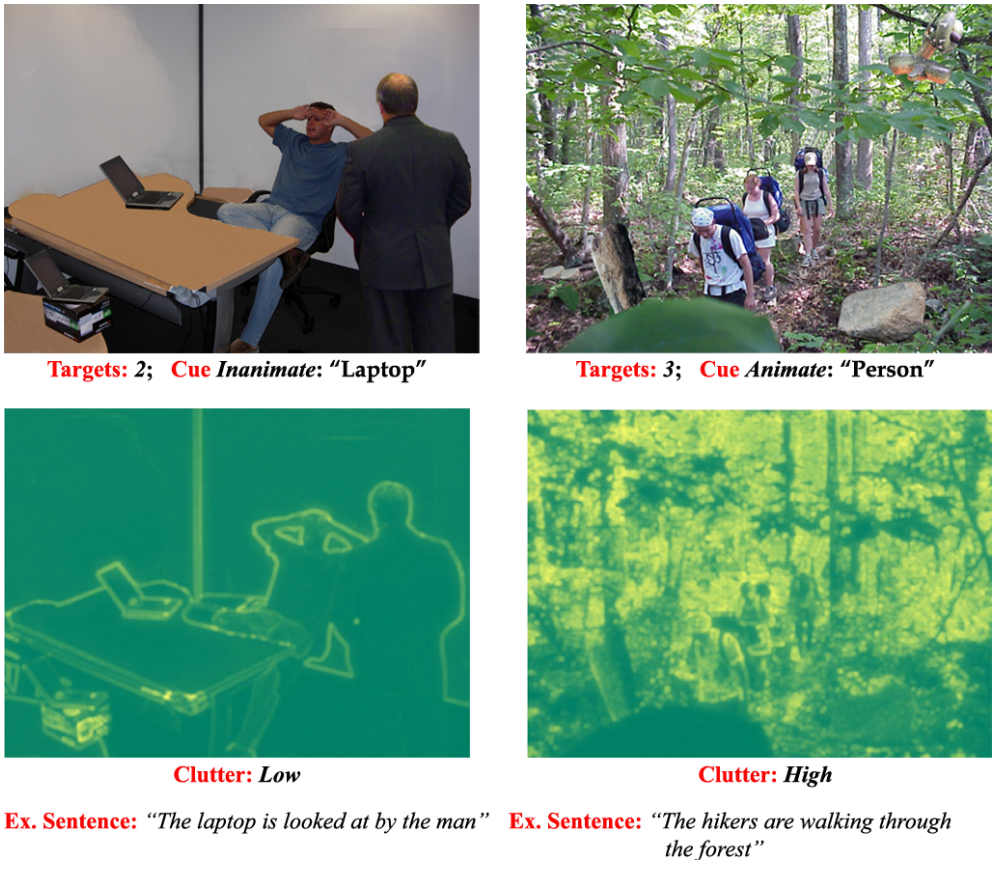


Figure 6.1: On the upper row, an example of scene and cues used as stimuli for the visual search and production task. On the bottom row, density maps of corresponding scenes are computed using feature congestion (Rosenholtz *et al.*, 2007): Low and High clutter.

Participants sat between 60 and 70 centimeters from the computer screen, which subtend a region $\approx$ 20 degree of visual angles. Only the dominant eye was tracked. A cue word appeared for 750 ms at the center of the screen, after which the scene followed and the search or description task began[1]. A 9 points randomized calibration was done at the beginning of the experiment, and repeated every $\approx$ 24 trials. Drift correction was performed at the beginning and between each trial. Once every four trials, during the Search task, a comprehension question about the number of target objects present in the scene was asked. Participants had to respond by pressing a button on the control pad which corresponded to the number of targets (1, 2 or 3). At the beginning of

[1] In description, a lapel microphone was activated to record the descriptions generated.

each experiment, there were four practice trials to familiarize the participants with the task. There was no time limit for the trial duration and to pass to the next trial participants pressed a button on the response pad. The experimental task was explained using written instructions and took $\approx$ 30 minutes to complete.

6.3.3 Data Analysis

We compare tasks using standard eye-movement measures, which have been extensively applied in the visual cognition literature. Eye-movement behavior is analyzed both on the temporal, e.g. first pass fixation duration, and spatial component, e.g. total number of inspected objects. We perform both a descriptive analysis of observed data to show the empirical trend, and inferential analysis based on linear mixed effect models (LME) (Baayen *et al.*, 2008) to quantify the effects of design factors.

6.3.3.1 Pre-processing

As a pre-processing step, each scene has been fully annotated with labelled polygons drawn around the objects of the scene (Russell *et al.*, 2008). We divided the scenes into two groups, according to their level of Feature Congestion (FC) (Rosenholtz *et al.*, 2007) clutter[1] (Low, High).

On a side note, we expected a positive relation between number of objects and level of clutter: i.e. the higher the clutter, the more the objects. Instead, we observe partial independence between clutter and number of objects. Low cluttered scenes had a mean density of 3.10 ± 0.22 and mean number of objects 27.42 ± 9.93; whereas in high cluttered scenes, the mean density is 3.90 ± 0.24, and the number of objects 28.65 ± 11.30. It seems that the density of a scene does not directly inform on the number of referents contained. Probably, on one hand, objects differ by their visual density, and on the other hand not all objects can be labeled. This trade off between labels and density makes the relation between clutter and referents partially independent.

[1] We compute Feature Congestion on each scene to calculate its clutter. Then, we use the mean value of clutter to divide the different scenes into two classes: Low, i.e. lower than the mean, and High, i.e. higher than the mean.

6.3.3.2 Measures of eye-movement behavior

We define the target objects (1, 2 or 3) by their relative Position in the scene following a left-to-right order (Left, Middle, Right). We report percentage of misses (the target has not been fixated during the trial), and position of first target fixated. Outliers that were 2 s.d. from the mean, and first fixation after onset of scene, are removed. In order to evaluate the role of object area on fixation duration, we compare description and search by looking at mean fixation duration as a function of area, which we bin in blocks of increasing size.

Following Malcolm & Henderson 2010, we define three task phases during the first pass on the target object: **initiation**, **scanning** and **verification**. Initiation is the time spent before generating the first eye-movement. Scanning is the number of objects inspected before landing on the target[1], we include in the count also re-fixation on the same object. Verification is the fixation duration on the first target object, during the first pass. We also report total measures of verification and scanning. Total verification is the sum of fixations over all target objects during the whole trial. Total scanning, instead, is the sum of all inspections between passes on target objects. In each phase, we compare search and description on the factors Cue and Target, for the two levels of Clutter. Research in visual search has focused on single target objects; in our design, instead, we have up to three target objects, which allows us to test the impact of fixation *Order* (First, Second and Third) during scanning and verification. We look at the impact of order of fixation on scanning and verification during the first pass on target objects.

In order to quantify the spatial distribution of eye-movements, we compute attentional landscapes for each scene (Pomplun *et al.*, 1996), to compare search and description. Notice that the frequency of inspected objects and spatial distribution of fixations do not represent the same measure. In fact, the same few objects can be inspected multiple times (narrow spatial distribution), or many different objects can be inspected only once (wide spatial distribution). The landscapes are created by generating 2D Gaussians on the x-y coordinates for each fixation, with the height of gaussian

[1] In Malcolm & Henderson 2010, the scanning period is based on fixation duration, from the first saccade after beginning of target until first fixation on target. We will test this version of scanning in experiment 8.

weighted by fixation duration, and radius of 1 degree of visual angle (roughly 27 pixels), to approximate the size of the fovea. A fixation map is generated for each subject. Then, all maps obtained on the same scene, across all subjects, are summed, and normalized to be a probability distribution. We use the attentional landscapes to quantify the difference in spatial distribution between Search and Description. On the landscape, we compute Entropy[1] which conceptually represents the spread of information, i.e. the more entropy, the more spread fixations are on the scene. The entropy is calculated on each map:

$$\sum_{x,y} p(L_{x,y}) \log_2 p(L_{x,y}) \tag{6.1}$$

where $p(L_{x,y})$ reflects the normalized probability of fixation at a point x,y in the landscape L. We also compute the Jenson-Shannon (JS) divergence (Dagan *et al.*, 1997), a symmetric version of the Kullback-Leibler divergence (MacKay, 2003), used to measure the distance between two probability distributions. We use JS to calculate how distant are two different tasks in their fixation landscape. JS is calculated using the following formula:

$$JS(p_A||p_B) = \frac{1}{2}(KL(p_A||p_{avg}) + KL(p_B||p_{avg})) \tag{6.2}$$

where p_A and p_B, are the probability maps of fixation for the two tasks compared, $p_{avg} = (p_A + p_B)/2$ is a point-wise average of p_A and p_B and $KL(.)$ is the Kullback-Leibler divergence, calculate as following:

$$\sum_{x,y} (p(A_{x,y}) log(p(A_{x,y})/p(B_{x,y})) \tag{6.3}$$

where $p(A_{x,y})$ is the probability of fixation at point x,y in task A, and $\log(p(A_{x,y})/p(B_{x,y}))$ is the log-ratio between the two distributions (A and B). With entropy, we capture the distribution of visual attention and we can test whether search is more narrow than production or they show a similar distribution. With JS, we can test how similar are fixation landscapes across different tasks. We report JS results in the next experiment only, section 6.5.2.3, where the three different tasks are compared.

[1] An application of entropy measure for eye-movement data could be found in Frank *et al.* 2009

6.3.3.3 Inferential analysis

As in previous chapters, we inferentially analyze our observed data using Linear Mixed Effect Modeling (Baayen *et al.*, 2008), and report the coefficients of those experimental factors retained after model selection. Each eye-movement measure is fit in a separate model. The centered predictors used are Cue (Animate/Inanimate), Clutter (continuous variable) and Target (1,2,3), which is orthogonal coded (target 3 as reference level). The random effects are Subjects, Trials and Positions [1]. To control for effects of object area on our eye-movement measure, we residualize its effect on our eye-movement measures in a simple linear regression model, e.g. $depM \sim Area$, and we take the residuals obtained as dependent measure for the LME modeling. When order of fixation is a predictor for the first pass measures of scanning and verification, we cross Order and Target in the random effects, as the levels of order are directly associated to the number of targets. We select our best model following a step-wise forward procedure based on log-likelihood comparison of nested models (for details refer to chapter Methodology)

6.3.4 Results and Discussion

We begin by showing descriptive statistics of target search accuracy for search and description. Moreover, we investigate whether there any unexpected effects, which we have to control for during our LME analysis. In particular, we focus on whether the position of the target is predictive of specific routines of visual inspection, e.g. left to right scanning; and if the object area influences fixation duration. These results are supported by LME analysis which we discuss in the text.

In Table 6.1, we show percentages of missed targets comparing the two tasks. A miss is counted when no target object has been fixated during the trial. So, for 2 or 3 Targets, it means that none of the targets has been fixated. When 3 Targets were present in the scene, right and left target were roughly equidistant from the middle target. We find a main effect of Task ($\beta_{Search} = 0.03; p < 0.05$); more targets are missed in search compared to description. The description task forces participants to be more accurate during their visual inspection, therefore less trials are skipped compared to search. If

[1] Position indicates the location of the target (Left, Middle, Right)

Table 6.1: Percentage of target missed, comparing *Task*, divided by number of *Target*(columns), and *Cue* (rows).

Task	Cue	1 Target	2 Targets	3 Targets	Total
Description	Animate	2.61	0.44	0.90	
	Inanimate	6.44	5.44	4.11	
					19.56
Search	Animate	4.08	0.63	0.50	
	Inanimate	7.19	7.07	4.11	
					23.13

we break down the percentages by the number of Targets and Cue type, we find that more targets are missed when Cue is Inanimate[1] ($\beta_{Inanimate} = 0.28; p < 0.05$) and when 1 Target is in the scene ($\beta_{Target1} = 0.18; p < 0.05$). Animate targets are located more easily than Inanimate. Also, the more targets in the scene, the more likely it is that at least one target is found. In order to unravel the influence of target position on visual attention when more than one target was present in the scene (2 and 3 Target), we calculate how many times (in percentage) a target is fixated on the different positions (Left, Middle, Right) for the first time. If participants follow a reading-like behavior of picture scanning, we should see a preference of starting from the leftmost target.

In Table 6.2, we observe that when 2 Targets are present in the scene, there is a preference of looking at the leftmost first ($\beta_{Left} = 0.05; p < 0.05$), more prominently when the Cue is Inanimate ($\beta_{Left:Inanimate} = 0.07; p < 0.05$). However, we do not find a main effect of Task. When 3 targets are present, there is a preference for looks at the center of the screen ($\beta_{Middle} = 0.15; p < 0.05$). Moreover, we find an interaction between Task and Position ($\beta_{Left:Search} = 0.15; p < 0.05$), with a preference of starting from the leftmost target during a Search task. A search task triggers a more automatic, left to right routine of visual inspection, compared to description where the target has to be found and linguistically contextualized in the scene.

[1] Since we do not include a table for this data, in order to simplify the interpretation, we report the value of coefficient uncentered for the factor we want to discuss.

Table 6.2: Percentage of first looks on target given its position, comparing *Task* by the number of *Target* (2,3)

Number of Targets	Position	Production	Search
2 Targets	Left	54.04	56.27
	Right	45.95	43.72
3 Targets	Left	25.91	29.98
	Middle	51.06	45.16
	Right	23.01	24.85

Turning to the effect of object area, we investigate it as a function of fixation duration, see Figure 6.2.

We observe an effect of task, description has longer fixation duration than search, but there is no interaction with the area. We correlate fixation duration with area using Spearman ρ to quantify the direction trend. We find a significant negative correlation in both tasks $\rho_{Search} = -0.044; (p < 0.05)$ and $\rho_{Description} = -0.054; (p < 0.05)$. The bigger the area of the object, the shorter the duration on it. Nevertheless, the strength of the correlation is very small.

Two interesting points arise from this analysis: 1) Production demands more visual processing than Search, with longer fixation duration on objects[1], and 2) objects are not fixated proportionally to the space they occupy in the scene, but rather according to the referential information they carry. Since object area is marginal to the issues discussed in the current study, we residualized it on all measures of Scanning and Verification, prior to LME modeling.

6.3.4.1 First Pass: Initiation, Scanning, Verification

We divide the first pass, which is the time elapsed from the onset of the trial to the end of first fixation on the first target, into three phases: initiation, scanning, and verification (Malcolm & Henderson, 2010).

[1] This issues is explored further in the next sections

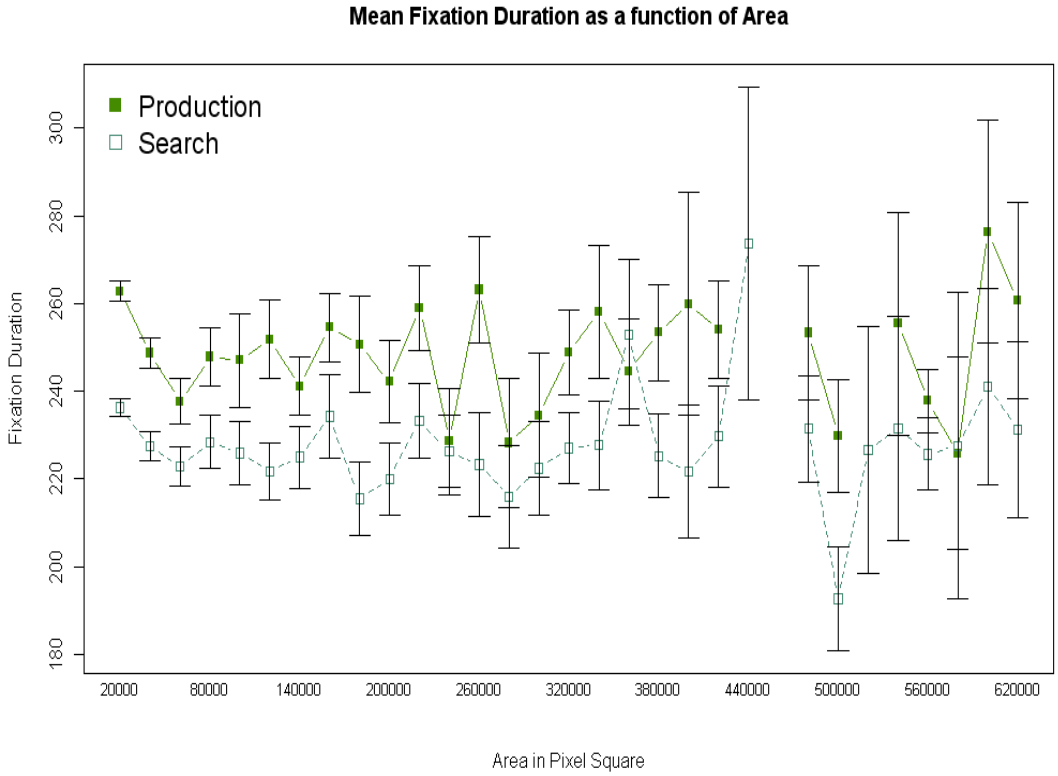


Figure 6.2: Fixation Duration as a function of object area. Comparing search and description. The green solid line represents description. The aquamarine dotted line instead is search.

In Figure 6.3, we show the results for the initiation phase. We find only a significant effect of Targets, initiation is slightly faster when there are 2 Targets compared to 3 Targets (refer to Table 6.3). Probably, the fact that less targets are identified during gisting might boost initiation of visual processing.

In Figure 6.4(a), we show results for the scanning phase. We find a main effect of task: in description more objects are inspected compared to search. A description, compared to search, requires the target object to be linguistically contextualized in the scene, thus more inspections are needed to retrieve visual material situating it[1]. We also observe a main effect of Targets. In particular, more objects are inspected when

[1]Notice that number of consecutive inspections does not imply more objects inspected: the same objects could be re-fixated. We will come back to this point when analyzing the spatial distribution.

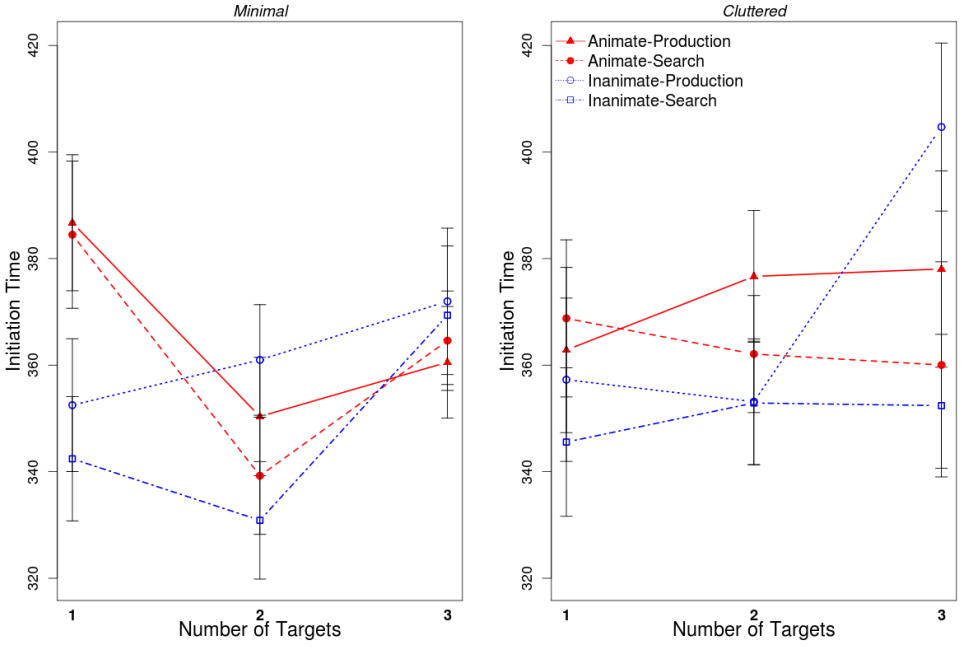
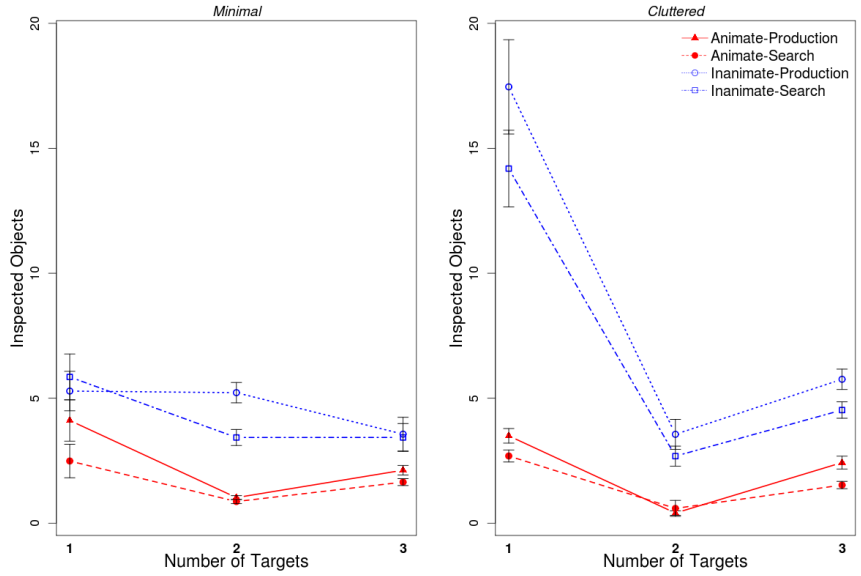


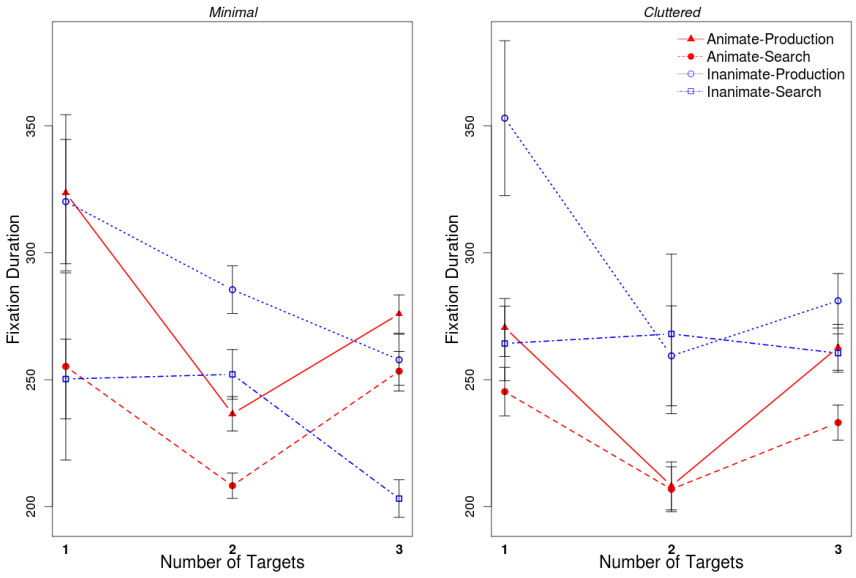
Figure 6.3: Initiation: the time spent to program the first saccadic movement. On the left panel, we plot results from low cluttered scenes (Minimal), on the right panel for high cluttered scenes (Cluttered). The Targets (1,2,3) are displayed on the x-axis. The colors represent the two factors of Cue: red is animate, blue is inanimate. The line and point types represent the 4 different condition compared to help visualization.

only 1 Target is in the scene, thus less likely to be fixated compared to 2 or 3 targets; and this difficulty increases when the object is Inanimate. Animate targets are more quickly detected than Inanimate targets, especially when only 1 Target is present. 2 Targets are easier to locate but, interestingly, when the Cue is Animate, we observe more inspections prior to target identification compared to 3 Targets. It seems that 2 Targets have a special status, compared to 3 Targets, which might due to the fact that 2 Targets are found more in connection with Indoor scenes, than the Outdoor ones[1]. In contrast with visual search studies, where an inanimate object is more difficult to find in a cluttered scene (Henderson *et al.*, 2009b), we do not find a main effect of clutter. Probably, an interaction could be found between clutter and target-one, however the model selection adopted discards all possible interactions of a factor which is not found

[1]The search experiment is designed as a follow up of the description one, where the targets were all depicted in Indoor scenes.



(a) Scanning: the number of objects inspected before landing the first time on the target object.



(b) Verification: the fixation duration during the first pass on the first target object.

Figure 6.4: Measures of First Pass. On the left panel, we plot results from low cluttered scenes (Minimal), on the right panel for high cluttered scenes (Cluttered). The Targets (1,2,3) are displayed on the x-axis. The colors represent the two factors of Cue: red is animate, blue is inanimate. The line and point types represent the 4 different conditions compared to help visualization.

Table 6.3: LME coefficients. The dependent measures are: *Initiation*, *Scanning* and *Verification*.The predictors are: *Target* (1;2;3), target 3 is expressed at the intercept, Cue (*Animate* -0.4, vs *Inanimate* 0.6), Task (*Search* 0.5, vs *Description* -0.5), and Clutter.

	Initiation	
Predictor	Coefficient	p
Intercept	360.14	0.0001
Target 2	−16.58	0.02
	Scanning	
Predictor	Coefficient	p
Intercept	0.66	0.1
Task	−0.82	0.001
Target 1	4.86	0.001
Cue	4.01	0.0001
Target 2	−2.16	0.01
Cue:Target 2	−7.84	0.0001
Task:Cue	−0.82	0.01
Target 1:Cue	11.22	0.01
	Verification	
Predictor	Coefficient	p
Intercept	6.233	0.1
Task	−33.12	0.001
Target 1	41.22	0.01
Target 2	−25.67	0.01

as main effect (see Chapter 2 for details).

In Figure 6.4(b), we show results for the verification phase. Similar to scanning, we observe a main effect of Task: in description, the first fixation is longer than in search (see Table 6.3 for coefficients). A description task requires more interaction between visual and linguistic processing compared to a search task. In fact, the cross-

modal integration between visual and linguistic referential information demands longer fixations. In search instead, once the object is found, it does not need to be further integrated with other ongoing cognitive processes. We also find a main effect of number of targets. When only 1 Target is in the scene, it is fixated longer. Since no other target objects are competing for visual attention, the verification phase takes longer. On the contrary, for 2 Targets, we find shorter fixation duration compared to 3 Targets. The verification time is decreased by the referential competition of 2 targets. Probably, 2 targets generate more competition than 3 targets, as fixations are more tightly launched to discriminate the pair of competitors. This issue is discussed in more details in the next section.

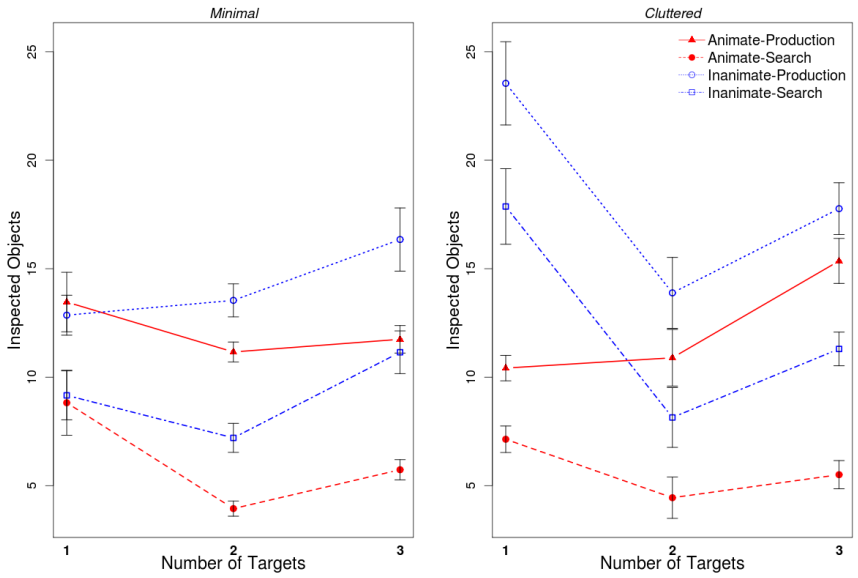
6.3.4.2 Total

In order to have a picture of eye-movement behavior during the whole trial, we analyze total verification and scanning[1]. Total verification is the sum of fixation duration on the target objects across the whole trial. Total scanning is the frequency of inspected objects between passes on the target objects, across the whole trial.

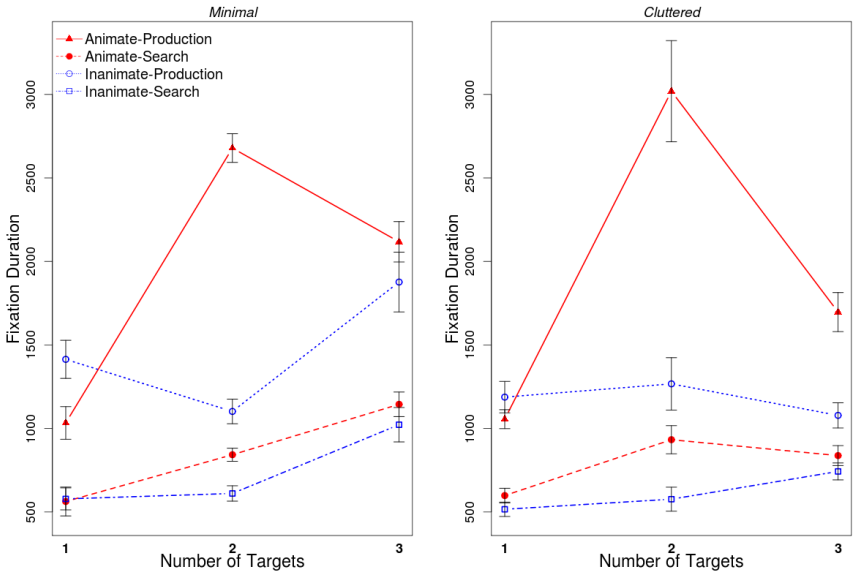
In Figure 6.5(a), we show frequency of total inspections. We find a main effect of Task, whereby during description there are more inspections compared to search (refer to Table 6.4 for coefficients).

During linguistic processing, the target object has to be visually contextualized in the ongoing description, thus objects are fixated/re-fixated in order to clearly establish associations between the observed visual referents with the linguistic referents mentioned. For search, instead, visual processing stops once the object has been found. We confirm the main effect of Cue, where Inanimate cues trigger more inspections compared to Animate ones, especially in cluttered scenes, where more regions are inspected to identify all target objects. Similarly to first pass, we observe a main effect of number of target, but this time only for 1 Target. A single target object triggers overall more inspections. Participants want to be sure that they did not miss any target, especially during a search task.

[1] Initiation time can only be computed for first pass.



(a) Total Scanning: the frequency of inspected objects between passes on the targets.



(b) Total Verification: the total sum of fixation duration on the target objects.

Figure 6.5: Total Measures.On the left panel, we plot results from low cluttered scenes (Minimal), on the right panel for high cluttered scenes (Cluttered). The Targets (1,2,3) are displayed on the x-axis. The colors represent the two factors of Cue: red is animate, blue is inanimate. The line and point types represent the 4 different condition compared to help visualization.

In Figure 6.5(b), we show total verification time across all targets. Also in total verification time, we find a main effect of Task: description, differently from search, requires the integration of visual and linguistic information; which results in longer fixations. Differently from first pass, we find that animate targets are overall more inspected than inanimate ones, especially during a description task, where animate referents carry important conceptual information to source the underlying sentence encoding. When looking at effects of Target, we find that a single target receives less looks during the course of a trial than 3 targets. Despite the fact that in Figure 6.5(a) we observe more looks to 2 animate targets during description, we do not find this effect to be significant after model selection. The reason is that as main effect, 2 targets is not significantly different than 3 targets, thus the associated interactions are not considered (see Chapter 2). Probably, however, an explanation for the effect observed is that if two visual targets share the same linguistic referent, they both can appear in the same description through coordination, thus triggering visual competition. For 3 targets this situation is less likely, as a sentence containing three identical referents connected through two coordinators, *the man is drinking and another man is walking and another one is greeting*, is less frequent than a sentence containing two referents connected with a single coordinator, *the man is drinking and another one is walking*.

Table 6.4: LME coefficients. The dependent measures are: *Total Verification* and *Total Scanning*. The predictors are: *Target* (1;2;3), target 3 is expressed at the intercept, Cue (*Animate* -0.4,vs *Inanimate* 0.6), Task (*Search* 0.5, vs *Description* -0.5) and Clutter.

	Total Scanning	
Predictor	Coefficient	p
Intercept	-0.2716	0.7
Task	-6.01	0.001
Clutter	41.32	0.01
Animacy	1.53	0.04
Target 1	0.92	0.5
Clutter:Cue	122.71	0.0003
Task:Target 1	2.77	0.006
	Total Verification	
Predictor	Coefficient	p
Intercept	-3.581	0.7
Task	-1033.341	0.0001
Target 1	-402.99	0.001
Cue	-265.612	0.001
Task:Cue	850.10	0.0001

6.3.4.3 Ordered Targets

The order in which targets are fixated might have an impact on visual responses. We calculate first pass scanning and verification on each target object based on the Order of fixation (First, Second and Third). For reason of conciseness, we do not show plots but only report and discuss the results obtained with LME analysis.

In Table 6.5, we report LME coefficient of scanning and verification predicted by a model where order of fixation is included as fixed effect. Together with Subject and Trials, we include a random slope (Order—Number of Targets), as the number of targets is crossed with order of fixation[1]. We confirm the main effects of task and cue found in section 6.3.4.1. Description demands more inspections than search. It takes longer to identify a target object when it is Inanimate, compared to when it is Animate. We also find a main effect of Order, where more objects are inspected before landing on the first target, compared to the number of objects inspected before landing on the third target. This effect is particularly prominent when the target object is Inanimate. In the model, we also observe that more objects are inspected before landing on the second target, compared to the third one but the effect doesn't reach significance. The more visual attention has gathered information about the scene, the less inspections are needed to locate subsequent target objects.

Table 6.5: LME coefficients. The dependent measures are: *Verification* and *Scanning*. The predictors are: *Order* (First;Second;Third), order third is expressed at the intercept, Cue (*Animate* -0.4,vs *Inanimate* 0.6), Task (*Search* 0.5, vs *Description* -0.5) and Clutter.

	Scanning	
Predictor	Coefficient	*p*
Intercept	0.5010	0.6
First	2.24	0.0001
Cue	3.69	0.1
Task	−0.80	0.002
Second	0.32	0.8
First:Cue	3.48	0.0001
	Verification	
Predictor	Coefficient	*p*
Intercept	3.072	0.6
First	−49.24	0.01
Second	0.29	0.9
Task	−30.71	0.001
Task:First	−29.23	0.03

On the verification time, we confirm the main effect of task observed in 6.3.4.1, where descriptions have a longer first pass compared to search; We find shorter verifi-

[1]The order of fixation is associated with the number of targets: if there is only 1 target depicted, there is only 1 possible order of fixation.

cation on the first target compared to the third one; but not during description, where instead the first target is fixated longer than a third target.

The trend of verifications seems to suggest that the fixation duration increases along with the number of targets observed. Probably, the more visual objects with the same linguistic reference are found, the more visual information is required to distinguish them, which leads to longer fixation duration. However, during description a different trend emerges. In this case, the target first fixated is probably selected as a linguistic referent to be mentioned, and fixation duration has to be longer, in order to identify its distinctive visual features, before passing onto the surrounding scene information to contextually situate it.

6.3.4.4 Spatial Distribution

The spatial distribution of eye-movements indicates how many regions of a scene are inspected during task performance. Different tasks trigger different spatial distributions, i.e. visual search has been shown to have a narrower spread than memorization (Castelhano *et al.*, 2009).

The more spread out the spatial distribution is, the more referential information of the scene is retrieved. We want to emphasize here that spread of fixations and number of inspections might be correlated, i.e. the more inspections, the more objects are scanned, but this cannot be generalized. For example, in the case of multiple Animate Targets, fixations can go back and forth between them. Our hypothesis is that a description task imposes a more stringent allocation of visual attention; which in turn results into a narrower selection of scene referents scanned. A description task constraints visual retrieval on those referents selected during sentence encoding. In search, instead, after an initial phase of contextually driven visual attention to locate the first target, it follows a second and more spread out phase, to find other potential targets embedded in the scene.

We have generated an attentional landscape (see Figure 6.6) of each scene for both description and search, and computed Entropy, to quantify the spread of spatial distributions. We find that search has a lower entropy compared to description; which nevertheless varies according to the other factors involved (see Table 6.6 for coefficients). Inanimate targets trigger higher entropy, especially during search; whereas

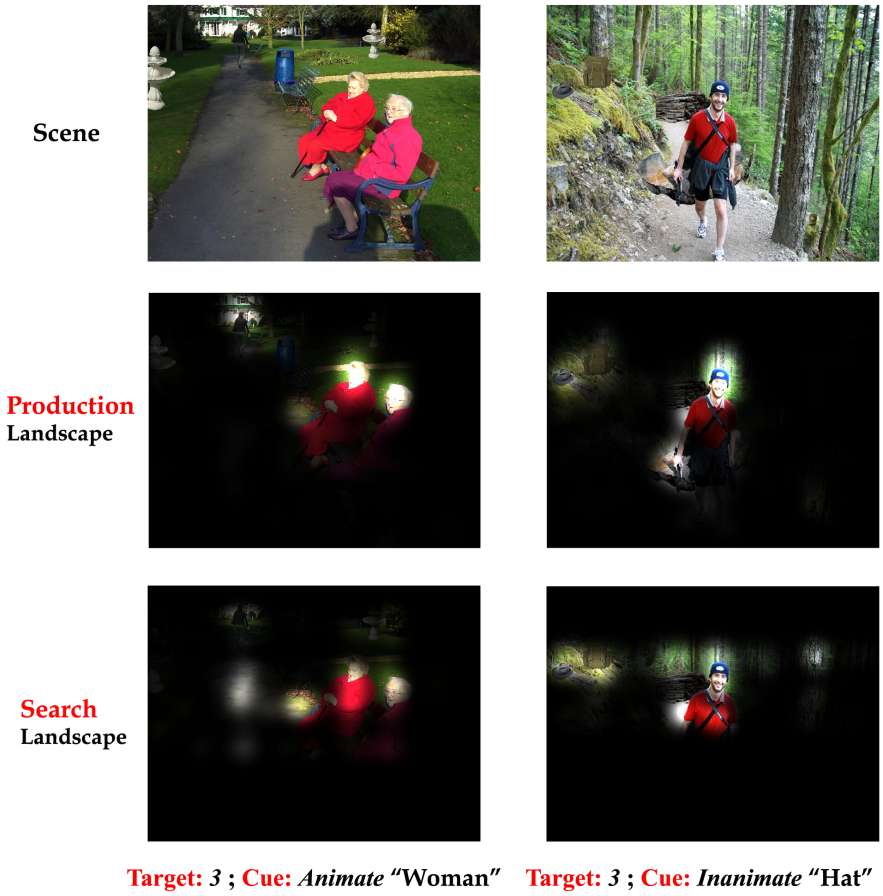


Figure 6.6: Attentional Landscapes. Comparing the spatial distribution of fixations of Search and Production.

during description, the same effect is found associated with animate referents. In line with previous findings, search is faster for animate objects (Fletcher-Watson *et al.*, 2008), thus fixation are less spread out across the scene. The opposite effect instead is observed on animate targets during description. Confirming results in Chapter 4, visual attention spreads on cluttered regions to retrieve scene informtion supporting the ongoing sentence generation. Clutter is also more generally found as a main effect, where the more visual information is available, the higher the entropy. We find also a main effect of target, where a single target leads to higher entropy than 3 targets, in particular when a search task is performed. The possibility that the scene might have up to 3 targets led the participants to inspect larger regions of the scene to look for

more targets, especially when only 1 target was present. We observe, instead, lower entropy for 2 targets compared to 3 targets. A reason might be that 2 targets requires less inspection to be found. Moreover, we believe that implicit learning might have developed during the course of the experiment by the participants, helping them to quickly discriminate between trials with 2 or 3 targets[1].

6.4 General Discussion

The aim of experiment 7 was to investigate the role of task interactivity on the active allocation of visual attention by comparing a purely visual task, i.e. search, with a multi-modal task, i.e. description.

We defined the notion of interactivity relative to the concept of cross-modal referentiality. We assumed that tasks differ by the amount of cross-modal interaction required to perform them. This cross-modal interaction depends on whether referential information has to be integrated, or not, across modalities. In a search task, visual attention utilizes referential information of scene and target to build a cognitive relevance model and optimally allocate attentional resources (Malcolm & Henderson, 2010). In such tasks, only visual referential information has to be accessed and utilized; thus, after the target object is verified, no further processing is needed. The situation is rather different, however, when the task performed demands synchronous processing between modalities. As seen in chapters 2 and 3, if a situated language processing task is performed, e.g. description, visual attention

Table 6.6: LME coefficients. The dependent measure is *Entropy*. The predictors are: *Target* (First;Second;Third), target 3 is expressed at the intercept, Cue (*Animate* -0.5, vs *Inanimate* 0.5), Task (*Search* 0.5, vs *Description* -0.5) and Clutter (*Minimal* 0.35 vs *Cluttered* -0.65. Scenario (Indoor -0.25 vs Outdoor 0.75) is included as random effect to control for differences in clutter.

	Entropy	
Predictor	Coefficient	p
Intercept	11.88	0.0001
Cue	0.20	0.0001
Clutter	−0.12	0.006
Target 2	−0.30	0.0001
Task	−0.07	0.001
Target 1	0.15	0.03
Cue:Task	0.21	0.0001
Task:Target 1	0.14	0.03

[1] 2 Targets are found more often in Indoor Scenes. Although, the result holds even after including Scenario as a random effect in our LME models.

and sentence processing interact over a shared cross-modal referential interface, which allows modalities to coordinate the synchronous processing flow. Our main hypothesis was that the cross-modal coordination demanded by a description task influences both components[1], temporal and spatial, of visual attention differently than search. Especially, during description we expected more temporal processing compared to search, e.g. longer fixation duration, because visual and linguistic referential information has to be integrated across modalities. Moreover, we expected task interactivity to be modulated by non-linguistic factors of the target, i.e. animacy, and the scene, i.e. clutter. In line with the search literature, clutter was expected to impair visual search performance (Henderson *et al.*, 2009b), and animate targets to make it faster (Fletcher-Watson *et al.*, 2008). In line with results shown in Chapter 4, during a description task we expected sentence encoding to benefit from clutter, especially in relation to animate targets, as a cluttered scene provides more contextual information to linguistically situate the targets.

We compared search and description on a range of standard eye-movement measures, covering both spatial and temporal components of visual processing, over two main levels of granularity: first pass, i.e. until the target is found, and total, i.e. the whole trial. Results on first pass and total confirm our main hypothesis: the more cross-modal interactivity is needed, the more visual processing occurs.

On the spatial component, we find that description triggers more inspections than search. When looking at their distribution, furthermore, we find that conceptual factors, e.g. animacy, play an important role in which regions are inspected. Especially, inanimate targets lead to a more spread distribution during search than description, whereas the opposite is found for animate targets. A description of an inanimate object requires it to be anchored in the context of another ground object, e.g. *the pen is on the table*, hence narrowing the span of visual attention. In search, instead, the scene is more widely inspected, as other target objects can be found at different locations, e.g. a PEN on a COUNTER. A completely different pattern is observed for animate objects. In fact, descriptions of animate objects need contextual information, e.g. HOTEL, MAIN HALL, to linguistically situate the referent, e.g. *the man is signing in*; thus, visual at-

[1]We have introduced this distinction to give a better and more contextualized interpretation of the eye-movement measures used, see section 6.2 for details.

tention tends to spread more compared to search, where instead animate referents are quickly identified (Fletcher-Watson *et al.*, 2008).

On the temporal component, we observed longer first pass and total verification during description compared to search. Again, non-linguistic factors have been shown to have a crucial influence on visual responses for both tasks. In particular, fixation duration was found shorter on animate objects in search than description. Moreover, when investigating the impact of order of inspection, we find that the first inspected target is more fixated during description than during search. Probably during description, the first target inspected is selected as linguistic referent of the sentence, thus visual information about it has to be extracted; whereas in search, the first target is quickly verified, in order to proceed with the search, as other targets can still be embedded in the scene.

Description and search primarily differ in the cross-modal interactivity required to perform them. In description, sentence processing and visual attention are synchronously activated, and mechanisms of cross-modal referentiality are needed to integrate information across modalities. In search, instead, only visual attention is actively involved. The need for cross-modal interactivity, however, gradually changes across different tasks. An object naming task, for example, can be imagined as intermediate with respect to search and description. In such a task, a search is performed to identify objects which are interesting given their visual features, however, since they have to be named, their relevance has to be linguistically evaluated. In Experiment 8, we compare search and description with an object naming task. We expect to find similarity with both search and description in the visual responses associated to this task.

6.5 Experiment 8: Cross-modal interactivity across tasks

In experiment 8, we investigate how a task requiring an ‘intermediate‘ degree of cross-modal interaction, i.e. object naming, compares to a single modality task, i.e. search, and a fully multi-modal task, i.e. description. Our main hypothesis is that object naming should share patterns of visual responses with both search and description. In particular, we expect naming and search to behave similarly on the spatial component

of visual processing; as search and naming involve a wider scanning of the scene, i.e. looking for cued targets or finding interesting objects to name, compared to a description, which instead focuses on the visual objects associated to the sentence generated. On the temporal component, instead, we expect more similarities between naming and description. As observed in Experiment 7, the cross-modal integration of referential information should result into longer fixation duration measures in both naming and description compared to search.

In relation to the factors of clutter and animacy; we expect again object naming to share similarities with both search and description. We have observed that the more the clutter, the more the objects inspected, especially during search for inanimate targets. Here, we expect object naming to show patterns similar to search. In fact, the more the clutter, the more objects can be named; and this effect would be independent from the animacy of targets.

Finally, in order to quantify the overlap of referential information processed during the different tasks, we investigate similarities between scan patterns. We expect more similarities of scan patterns to arise between tasks involving cross-modal interaction, i.e. naming and description, as in both tasks, the objects are fixated according to their linguistic relevance, compared to search where instead targets are fixated merely according to their contextual relevance.

6.5.1 Method

In an eye-tracking experiment, participants were asked to name the most relevant 5 objects embedded in photo-realistic scenes. A subset of the material used in this experiment contained the 24 experimental scenes, in the two versions Minimal and Cluttered, used in the scene description experiment presented in Chapter 4, and the search experiment of this chapter. Our analysis will focus on these 24 scenes, as they are shared across the three experiments and therefore directly comparable. These scenes contained always 2 referentially ambiguous targets, animate (*man*) and inanimate (*clipboard*), thus there is no condition involving the number of targets.

In the analysis, we use the eye-movement measures discussed in section 6.3.3.2. Notice that, in an object naming task, participants were not cued to a particular object, and therefore looks to animate or inanimate targets are not controlled. However, in

order to keep animacy of objects as an explanatory variable, in object naming we calculate eye-movement measures based on the cued objects (animate, inanimate) given in the description and search experiments.

We divide the first pass in three phases: initiation, scanning and verification. For scanning, instead of reporting the number of inspected objects, we use the correlated measure of latency, which tells us how much time elapses from beginning of trial, until the first fixation on the target object. We use latency to make our measures of first pass directly comparable to Malcolm & Henderson 2010. As total measures, we report the total verification on targets, and a more general measure based on proportion of fixations for all animate and inanimate objects contained in the scene. In addition, we consider the measure of mean gaze duration, i.e. the average fixation duration across all objects, to test how much temporal processing is devoted during a fixation for the different tasks.

We calculate entropy and JS-divergence on attentional landscapes generated for the different scenes (see section 6.3.3.2, for details). Moreover, on the same scene, we calculate pairwise scan pattern similarity[1] performing two types of analysis: (1) within task, i.e. comparing different participants performing the same task; and (2) between tasks, i.e. different participants from different tasks. The within task analysis investigates how much similarity there is in the referential information visually processed across different participants when performing the same task. The between task analysis, instead, investigates how much similarity is shared across different tasks.

We model our eye-movement measures using linear mixed effect models. The predictors are Clutter (Minimal/ Cluttered), Cue (Animate/Inanimate), Task (Search, Naming, Description), coded using treatment coding. The reference level is chosen according to the observed data, always taking the factor which allows a richer interpretation[2].

The random effects are Subjects and Trials. To avoid unexpected effects due to object size, we residualize object area on each dependent measure.

[1] We report results using OSS measure, see Chapter 2 for details.

[2] The choice of the reference level affects the interpretation of the results for each individual coding variable; however, it does not change the overall effect of the model fit and related statistics. Source: Statistical Tutorial at McGill, http://wiki.bcs.rochester.edu:2525/HlpLab/StatsCourses

6.5.2 Results and Discussion

We begin discussing the results found on the measures of first pass: initiation, scanning and verification. We proceed with total verification, mean gaze duration and proportion of fixation. Then, we explore the entropy and JS-Divergence of spatial distribution across the different tasks, and finish our discussion looking at the scan pattern similarities.

6.5.2.1 First Pass: Initiation, Scanning and Verification

In Figure 6.7, we show results for the initiation phase across the three different tasks.

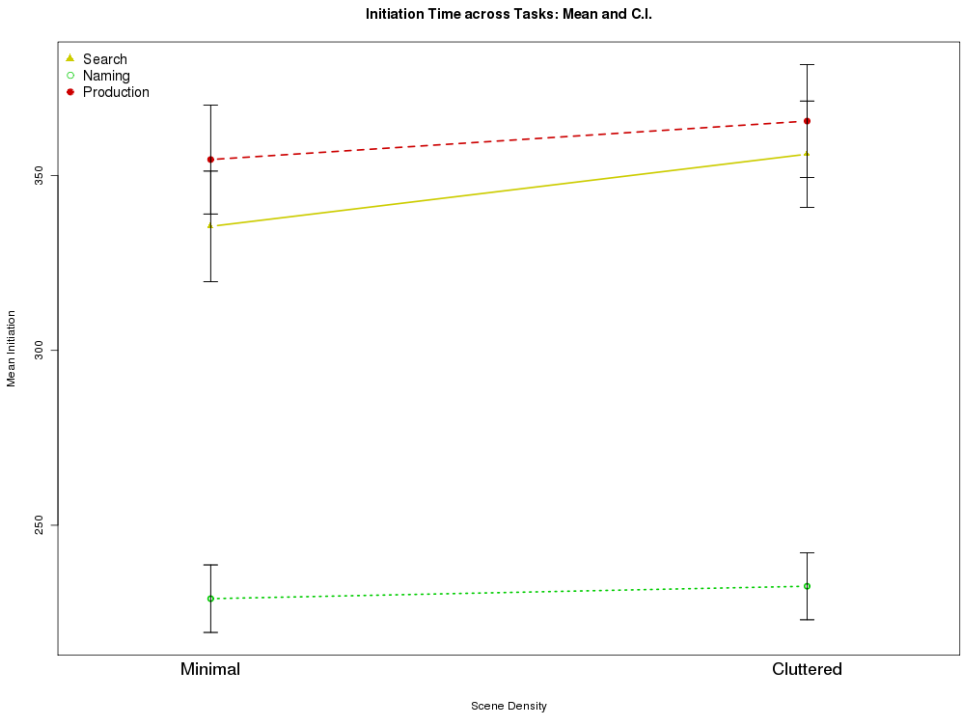


Figure 6.7: Initiation: the time spent to program the first saccadic movement.

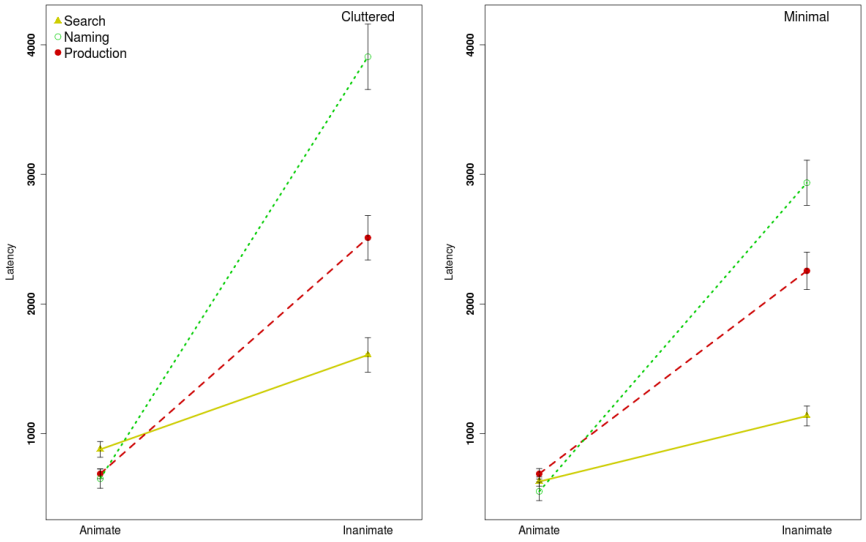
We find only a significant effect of Task, where a naming task has a faster initiation time compared to search (see Table 6.7, for coefficients). No difference is found between search and description. In search and description, an expectation template is

generated on the basis of the cue, thus initiation takes longer than in naming where no cue is given.

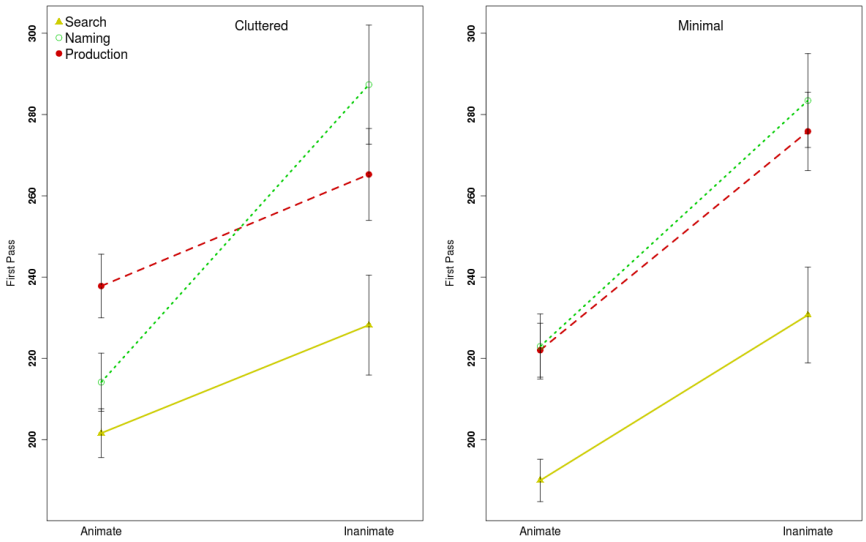
Table 6.7: LME coefficients. The dependent measures are: *Initiation*, *Scanning* and *Verification*.The predictors are: Task (*Search*, *Naming* and *Description*) with search used as a reference level, Clutter (*Minimal* 0.5, *Cluttered* -0.5) and Cue (*Animate* -0.5, vs *Inanimate* 0.5)

	Initiation	
Predictor	Coefficient	p
Intercept	312.19	0.0001
Naming	−122.2	0.001
	Scanning	
Predictor	Coefficient	p
Intercept	61.13	0.4
Cue	1973.29	0.0001
Naming	684.07	0.0001
Clutter	−294.1	0.0001
Description	331.85	0.0001
Cue:Naming	2168.43	0.0001
Clutter:Naming	−264.74	0.03
Cue:Production	1050.21	0.0001
Cue:Naming:Clutter	−734.93	0.002
	Verification	
Predictor	Coefficient	p
Intercept	−0.2281	0.9
Cue	48.60	0.0001
Description	37.63	0.0002
Naming	35.97	0.003
Cue:Naming	30.90	0.003

In Figure 6.8(a), we show results for the scanning phase. We find a main effect of task: in description, and especially naming, it takes longer to fixate the target for the first time; the effect mostly regards inanimate targets and it is more prominent for naming compared to description. During naming the participants were not cued, thus the inanimate objects all had the same chance of being looked. Moreover, in a scene there are more inanimate objects than animate ones, thus participants took longer to



(a) Scanning (Latency): the time elapse from the beginning of the trial until the target object is fixated for the first time.



(b) Verification: The fixation duration during the first pass on the target object.

Figure 6.8: Measures of First Pass. On the left panel, we plot results from low cluttered scenes (Minimal), on the right panel for high cluttered scenes (Cluttered). Cue (Animate, Inanimate) are displayed on the x-axis. The three tasks are plotted using different colors, line types and points: Search (yellow, full line, triangle); Naming (green, small dotted lines, empty circle) and Description (red, large dotted lines, full circle).

fixate at a specific inanimate object. As expected, the more the clutter, the longer it takes to identify the object in the scene; in particular if the object is inanimate and the task is to name it. As just said, the naming task is not triggered by cueing to particular objects of the scene.

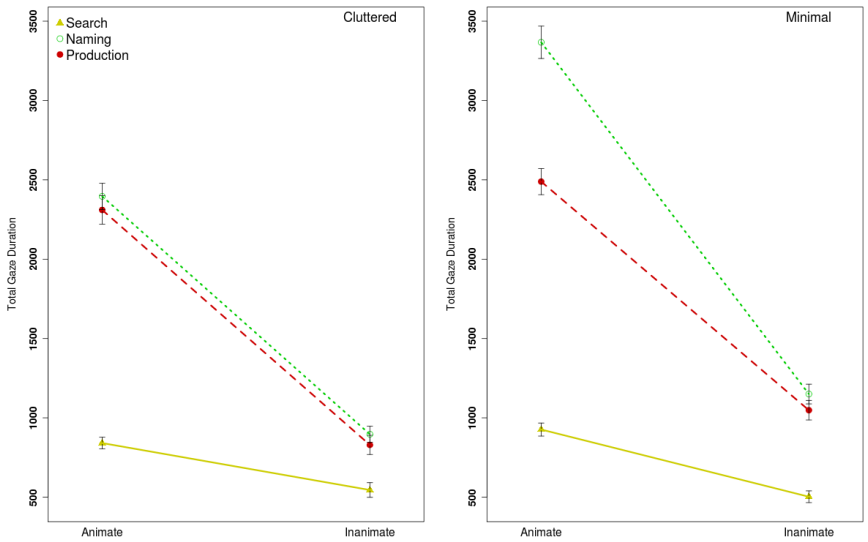
In Figure 6.8(b), we show results for the verification phase. Similar to scanning, we observe a main effect of Task: in description and naming, the first fixation is longer than in search (see Table 6.7 for coefficients). Confirming results of section 6.3.4.1 of experiment 7, a cross-modal task implies a longer temporal processing, as referentiality has to be integrated across modalities. In line with previous literature, animate referents are fixated less than the inanimate one, especially during a naming task, where the retrieval of associated name is boosted for animate objects (Branigan *et al.*, 2008).

6.5.2.2 Total

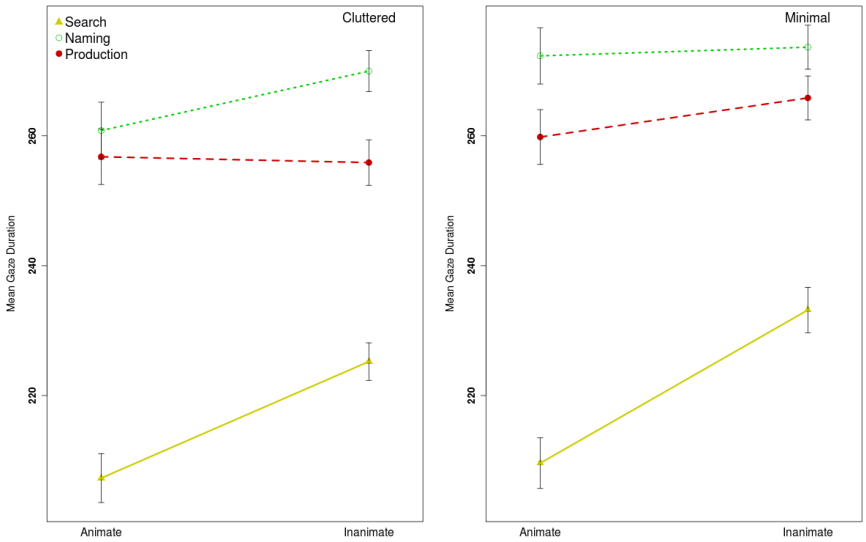
In this section, we report results related to the whole trial. Thus, we analyze total verification, i.e. the sum of fixation duration on targets, mean-gaze duration, i.e. the average fixation duration across all objects, and proportion of fixation, i.e. proportion of fixation on all animate and inanimate objects of a scene.

In Figure 6.9(a), we show total verification. We find a main effect of Cue, whereby animate referents are fixated more than inanimate referents, especially during the two cross-modal tasks, description and naming, which overall have longer total fixation compared to search (refer to Table 6.8 for coefficients). The conceptual richness of animate referents makes them more likely to be fixated during a linguistically mediated task. Furthermore, the animacy of the referent interacts with the density of visual information: the less the clutter, the more attention focuses on the animate referents, and this is especially true during a naming task, where the scarcity of inanimate referents to name, pulls attention more prominently on the animate referents.

In Figure 6.9(b), we show mean gaze duration on all objects of a scene, divided by their animacy. Confirming previous finding, we find a main effect of task: cross-modal tasks require longer mean gaze compared to a single modality task, i.e. search. The cross-modal integration of referential information demands longer temporal processing. We find also a main effect of cue, which differs from that observed in total verification. Inanimate objects have a higher mean gaze compared to animate objects,



(a) Total Verification: the the total sum of fixation duration on the target objects.



(b) Mean Gaze: the average fixation duration across all objects inspected.

Figure 6.9: Total Measures.On the left panel, we plot results from low cluttered scenes (Minimal), on the right panel for high cluttered scenes (Cluttered). Cue (Animate, Inanimate) are displayed on the x-axis. The three tasks are plotted using different colors, line types and points: Search (yellow, full line, triangle); Naming (green, small dotted lines, empty circle) and Description (red, large dotted lines, full circle).

especially in a search task, as suggested by the negative interaction found between cross-modal tasks and inanimate objects (see Table 6.8 for full list of coefficients). In a search, it is crucial to verify the identity of objects fixated in respect of the cued target. The operation of verification is more difficult on inanimate objects, as their visual features might be shared with other visual objects embedded in the scene; which makes their recognition more ambiguous compared to animate referents.

In Figure 6.10 we observe a significantly higher proportion of looks to inanimate objects across all tasks; especially during naming and search, although this interaction is not find significant in the LME analysis. The reason is that tasks are not significantly different as main effects, but only in interaction with the cue. However, given our model selection procedure, these interactions will not be considered, as they violate the subset assumption (see Chapter 2 for details). In fact, when we manually build a model, containing Cue as main effect and in interaction with Task, we find the interactions to be significant in the direction suggested by Figure 6.10 (coefficients are reported in Table 6.9).

During search and naming the scene is widely inspected to find targets, or name linguistically relevant objects. Thus, inanimate objects have a more important role than animate ones. During description, instead, animate objects play a key role for sentence encoding, as sentences usually have animate referents as subjects.

Table 6.8: LME coefficients. The dependent measures are: *Total Verification*, *Mean Gaze* and *Proportion of Fixation*. The predictors are: Task (*Search*, *Naming* and *Description*; with search used as a reference level), Clutter (*Minimal* 0.5, *Cluttered* -0.5) and Cue (*Animate* -0.5, vs *Inanimate* 0.5)

	Total Verification	
Predictor	Coefficient	p
Intercept	−15.64	0.8
Cue	−1357.2	0.0001
Clutter	328.31	0.0001
Naming	1470.34	0.0001
Description	1132.99	0.0001
Clutter:Naming	689.69	0.0001
Cue:Naming	−1489.06	0.0001
Cue:Clutter	−263.07	0.001
Cue:Description	−1082.91	0.001
Clutter:Description	166.43	0.07
Cue:Clutter:Naming	−621.33	0.0001
	Mean Gaze	
Predictor	Coefficient	p
Intercept	−0.0826	0.9
Cue	9.72	0.007
Clutter	7.01	0.0001
Naming	49.99	0.0001
Description	40.30	0.0001
Cue:Description	−17.90	0.0001
Cue:Naming	−15.54	0.003
	Proportion of Fixation	
Predictor	Coefficient	p
Intercept	−0.2281	1
Cue	0.28	0.0001

Table 6.9: LME coefficients for Proportion of Fixation from a manually constructed model. The dependent measures is: *Proportion of Fixation*. The predictors are: Task (*Search*, *Naming* and *Description*) with description used as a reference level, and Cue (*Animate* -0.5, vs *Inanimate* 0.5)

	Proportion of Fixation	
Predictor	Coefficient	*p*
Intercept	0	0.9
Cue	0.28	0.0001
Cue:Naming	0.17	0.0001
Cue:Search	0.07	0.0001

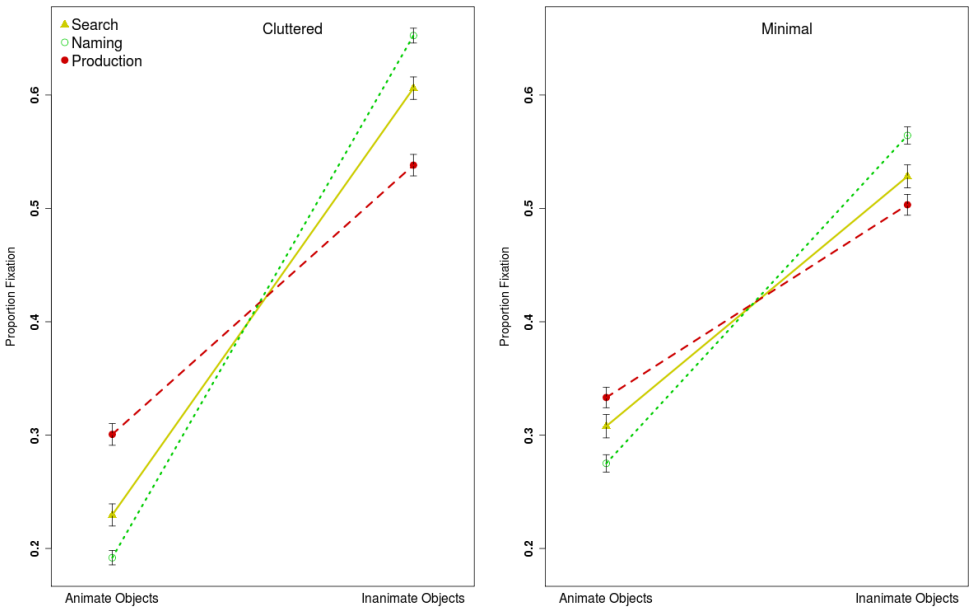


Figure 6.10: Proportion of Fixation spent on animate or inanimate object of a scene during a certain trial.

6.5.2.3 Spatial distribution

In this section, we compare the three different tasks on their attentional landscapes, thus directly addressing how visual processing is spatially allocated. We look at two measures of spatial distribution, entropy and JS-Divergence (see section 6.3.3.2 for details).

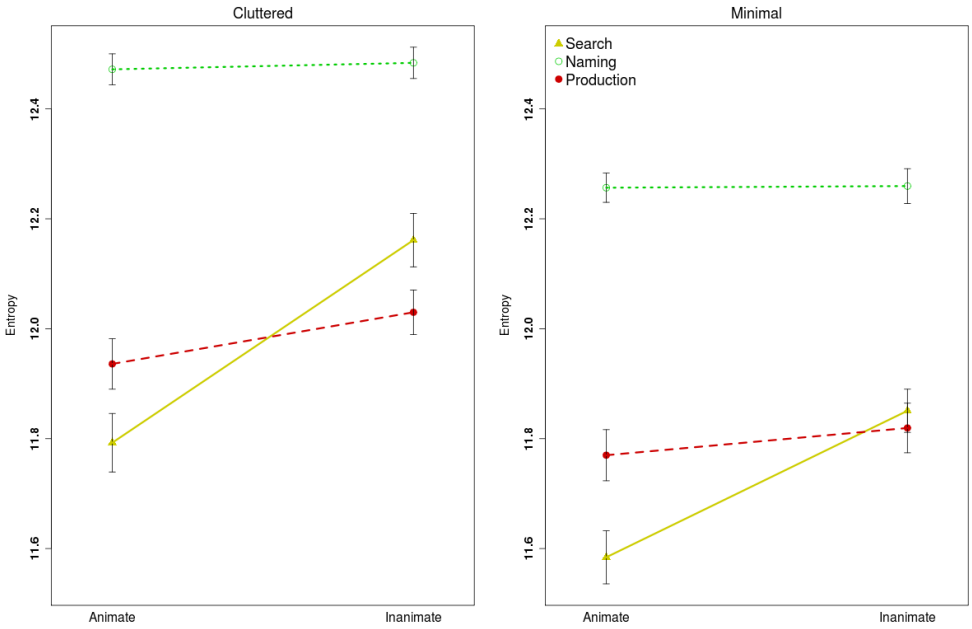


Figure 6.11: Entropy of fixation landscape across the three different tasks.

In Figure 6.11, we show how spatial fixation entropy changes across the different tasks, given the factors of animacy and clutter. We find that the spatial distribution for both search and description has a smaller entropy compared to naming (refer to Table 6.10 for coefficients). In a naming task, a scene is more widely inspected to ensure that linguistically relevant objects are not missed; whereas in search and description, visual attention has to focus either on the objects contextually appropriate to identify the target location, or semantically relevant to the sentence encoded.

In Figure 6.12, we show JS-Divergence of spatial distribution of fixations for the different tasks' comparison (refer to section 6.3.3.2 for details).

We find that naming and search have the highest divergence compared to the description/search and description/naming (see Table 6.10 for coefficients). Confirming what was observed in the analysis of entropy, during naming participants inspect more objects, which results into a more scattered fixation distribution. We also observe a main effect of Clutter, where the more the clutter, the more the divergence observed. Interestingly, there is no significance difference in the divergence between description/search and description/naming. In a description task, fewer visual referents (re-

Table 6.10: LME coefficients. The dependent measures are: *Entropy*, and *JS-Distance*. For entropy as dependent measure, the predictors are: Task (*Search*, *Naming* and *Description*) with naming used as a reference level and Clutter (*Minimal* 0.5, *Cluttered* -0.5). For JS-Divergence as dependent measure, instead of Task, we have task comparison (description/search, description/naming, naming/search), the reference level used is naming/search.

	Entropy	
Predictor	Coefficient	p
Intercept	12.03	0.8
Search	-0.52	0.0001
Clutter	-0.22	0.0001
Description	-0.47	0.0001
	JS-Divergence	
Predictor	Coefficient	p
Intercept	0.1526	0.0001
Description/Search	-0.0309	0.0001
Clutter	-0.0222	0.0001
Description/Naming	-0.0253	0.0001

lated to the ongoing encoding) are fixated, which might also be fixated during naming and search for their linguistic and visual relevance. In future research, we plan to investigate which objects are commonly fixated across different tasks and at what time during the trial.

6.5.2.4 Scan Pattern Similarities

In this section, we compare scan patterns generated by the participants while performing the different tasks. Beside its standard definition, a scan pattern can be imagined as an overt representation of how referential scene information is processed over time, during the performance of a given task. By computing similarity scores between scan patterns, we can observe how similar participants are while processing scene information within the same task, and between different tasks. The within task analysis measures how consistently referential information is processed, during a certain task, across participants. The between task analysis, instead, measures how similar the choices of referential information processing are between tasks.

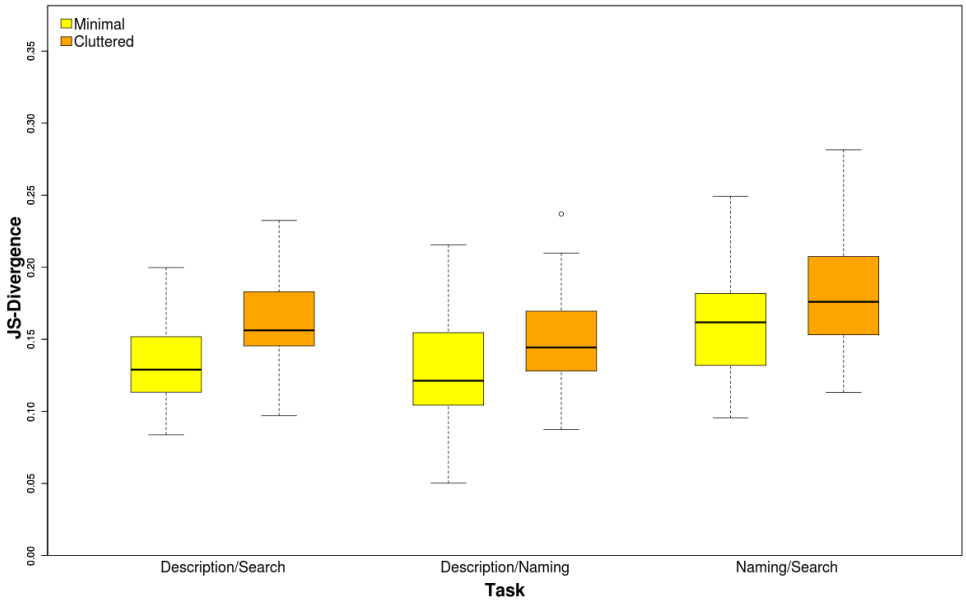
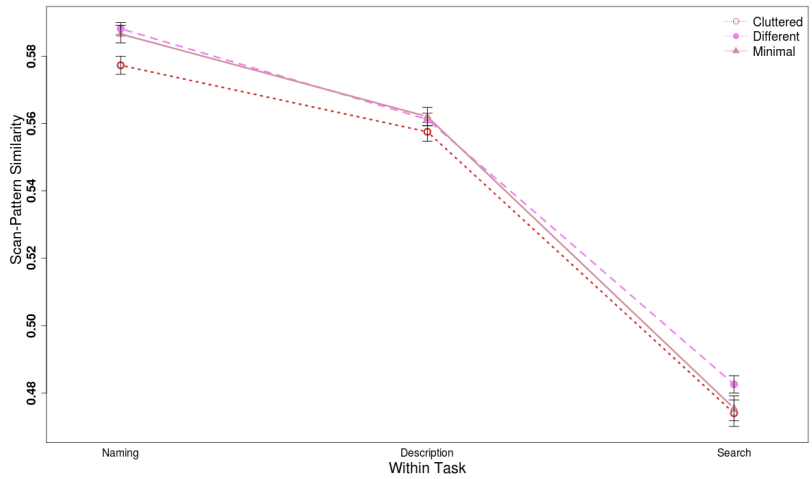


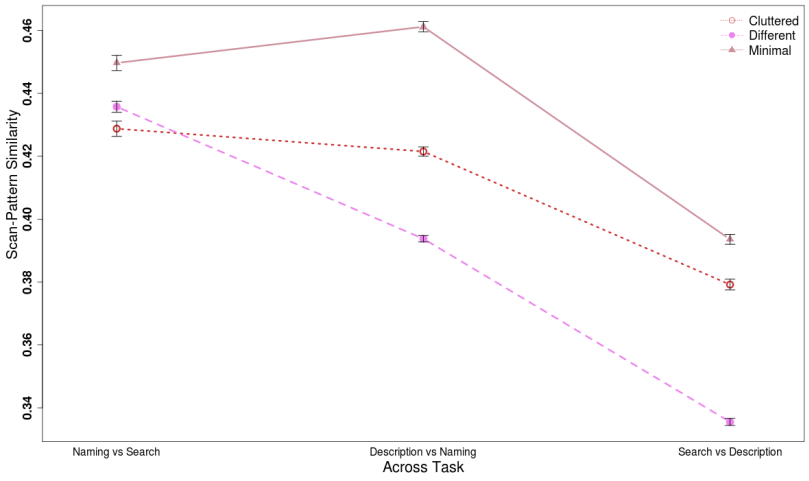
Figure 6.12: JS-Divergence box plot. On the x-axis, we show the different task comparison (description/search; description/naming; naming/search). On the y-axis, we plot JS-Divergence. The colors of the boxes refer to the conditions of clutter (Minimal - yellow; Cluttered - orange)

In Figure 6.13(a), we plot the scan pattern similarity within the same task. We find that in a naming task, participants have a higher scan pattern similarity compared to both description and search, which has the lowest similarity (see Table 6.11 for coefficients). This result is intriguing, as we observed that naming has a higher entropy, i.e. more objects are inspected, compared to search and description. Probably, however, the combined activation of object-based visual prominence together with a linguistic evaluation of object relevance has strengthened guidance of visual attention, thus making participants look at visual objects in a more similar order. In description, as seen in Chapter 5, the similarity of scan patterns is associated with the linguistic content of the sentence; thus if sentences are dissimilar also the associated scan patterns will be. In search, instead, after an initial effect of object-based guidance, since visual attention is not in synchronous processing with other modalities, it loses this referentially based control hence allowing for more variability across participants.

In Figure 6.13(b), we plot the scan pattern similarities across different tasks. We



(a) Within Task: scan pattern similarities on the same scene across participants during the same task.



(b) Between Tasks: scan pattern similarities on the same scene across participants between different tasks.

Figure 6.13: Scan-Pattern Similarity. Scan-patterns are compared pairwise, the same scene can be both minimal and cluttered; thus, Different refers to those cases.

find that a naming task shares similarity with both description and search. Despite the higher entropy of fixation observed in section 6.10, we find that the visual referents fix-

ated during naming are similar to those observed in the other two tasks. Moreover, we find that linguistically driven tasks are more similar, than when a visual modality task is compared with a linguistically driven task, i.e. description vs search. The cross-modal activation of visual and linguistic referential information allows synchronization of visual attention allocation. Low cluttered scenes allow a better synchronization of scan patterns between tasks compared to cluttered scenes; especially during linguistically driven tasks. The less visual information is available, the less linguistically relevant visual objects there are in the scene.

6.6 General Discussion

In experiment 8, we have extended our investigation of cross-modal interactivity by comparing object naming, i.e. a task demanding an intermediate level of synchronous activation between visual attention and sentence processing, with search, i.e. a single modality task, and description, i.e. a highly synchronized multi-modal task. We expected naming to share similarities of visual processing with both search and description.

In particular, on the spatial component of visual processing, we expected naming to resemble a search task, as the scene is widely inspected to have a full understanding of potentially interesting objects embedded that can be named. This similarity was expected to emerge especially when search is cued with an inanimate target, as it forces the visual inspection to span more broadly the scene to identify potential targets. We find that naming and search have a larger entropy of fixation distribution than description, where visual attention focuses on animate referents and the relation they have with the surrounding scene context. However, naming is a linguistically driven task; which implies that the objects of the scene are inspected relative to their linguistic relevance. The direct consequence of this fact is the overlap with the referential information processed during description. We find on JS-divergence, that description shares a similar fixation distribution with naming, i.e. similar objects are inspected, but naming is more spread, probably more objects are fixated, especially when compared to search. However, interestingly, we find that a more spread out distribution does not imply more variability of visual responses in naming than in description or search. On the contrary, we find that within tasks, scan patterns across participants are more

similar during naming than during description and search. Moreover, when comparing different tasks, we find that the order of objects fixated during naming is more similar to both search and description, than when search and description are compared. During naming, mechanisms of active scene exploration are weighted by linguistic judgments of the referential relevance of the objects in the scene. Taken together, these results suggest that naming is an intermediate task between search and description.

On the temporal component of visual processing, we expected naming to be more similar to description than search. Both description and naming demand cross-modal referential integration, whereas in search only visual integration, in the form of cue verification, is needed. In line with results from experiment 7, we find longer temporal processing, e.g. longer mean fixation duration, in naming and description than in search. This effect was true on all measures of fixation duration considered, with the exception of initiation time; where instead, we observed shorter initiation for naming compared to search and description. The longer initiations during search and description are a result of the integration between cued target and initial scene, i.e. gist, information. Since the naming task is not cued, the initiation time is faster.

Table 6.11: LME coefficients. The dependent measures are: *Similarity Within* and *Similarity Between*. For Similarity Within; the predictors are: Task (*Search*, *Naming* and *Description*) with search used as a reference level and Clutter (*Minimal*, *Different* and *Cluttered*), with Different used as a reference level. For Similarity Between, instead, the Task predictor is (Naming vs Description, Description vs Search, Search vs Naming), with Description vs Search used as reference level.

	Within Task	
Predictor	Coefficient	p
Intercept	0.59	0.0001
Naming	0.096	0.0001
Description	0.061	0.0001
Minimal	0.0022	0.6
	Between Tasks	
Predictor	Coefficient	p
Intercept	0.40	0.0001
description vs naming	0.05	0.0001
Minimal	0.04	0.0001
Clutter	0.021	0.001
Minimal:naming vs search	−0.04	0.0001
Clutter:description vs naming	−0.016	0.0001
Minimal:description vs naming	0.009	0.01
Clutter:naming vs search	−0.05	0.0001

When looking at the conceptual factors of target (i.e. animacy) and scene (i.e. clutter) that were manipulated, we largely confirmed what was observed in experiments 7-8 and previous chapters. Animate targets are identified faster than inanimate ones, especially during search; but

they are looked at more during description and naming, as they carry conceptual information highly relevant for a linguistic task. Moreover, animacy of targets interacts with scene clutter in several task dependent ways. The less the density, the easier is target identification. However, when scene information is used to drive a linguistic task, the higher density triggers linguistic facilitation as more referential information is available. In linguistic tasks, in fact, the low density of the scene pulls attention to the animate objects, as in the absence of a richer inanimate context, they carry the most relevant linguistic information that could be used during naming or description.

6.7 Conclusion

Theories of active visual perception have stressed the importance of task on defining the patterns of visual attention allocation (Castelhano *et al.*, 2009; Findlay & Gilchrist, 2001; Henderson, 2007; Yarbus, 1967). The research conducted in visual cognition has mainly focused on visual tasks, i.e. only visual attention is actively engaged. Nevertheless, several other tasks, e.g. scene description, require the synchronous interaction of different modalities, e.g. sentence processing. To the best of our knowledge, it is largely unknown how such linguistically driven tasks, which demand cross-modal interaction, compare to standard tasks in the visual cognition literature, and whether they are subject to the same visual biases. Moreover, in the context of this thesis, by understanding how visual processing reacts to the cross-modal demands of the task, we can develop a more general theory of cross-modal referential processing.

We extend previous literature in several significant ways. We confirm that referential object-based scene information is actively used during goal directed tasks to guide visual attention (Malcolm & Henderson, 2010; Nuthmann & Henderson, 2010; Schmidt & Zelinsky, 2009). Moreover, we find that the way referential information is visually processed is related to the nature of task, and the degree of cross-modal interactivity required. We observed that linguistically driven tasks share a similar pattern of referential information processing, where the joint activation of visual and linguistic information about the objects yields a more tight synchronization, compared to purely visual tasks.

Based on the nature of the eye-movement measure, we have distinguished between a spatial and temporal component of visual processing. This distinction has helped

the interpretation of the results by marking a difference between inspection, e.g. spatial distribution of fixation, and integration, e.g. mean fixation duration, processes. Interestingly, we have shown that naming and search are more related on the spatial component, i.e. wider scene inspection; whereas on the temporal component naming is more related to description, i.e. cross-modal integration.

In line with the previous chapters, we observed that the animacy of the target and the clutter of the scene have important influences on the pattern of visual responses observed, across the different tasks. Especially, inanimate objects are more difficult to identify in cluttered scenes during search, while they are an important source of referential information during naming. Animate objects, instead, are crucial during linguistically driven tasks, and this effect is especially evident in low density scenes. In such cases, the scarcity of inanimate objects to name or describe pulls visual attention to animate objects, which carry important conceptual information.

Overall, the nature of the task has important theoretical and modeling consequences. Our results support a theory of object-based allocation of visual attention during goal directed tasks (Nuthmann & Henderson, 2010; Zelinsky & Schmidt, 2009). However, this view does not spell out how the sub-goals of a task (e.g. formation of a target template, integration of gist information, localization of cued object, etc.) interact with the referential information about the objects, i.e. a CUPBOARD might contain a PLATE. This interaction of task sub-goals and object-based information is further complicated when visual attention interacts with sentence processing. In such cases, object information is relevant if it is also linguistically relevant, and its activation is bound to the constraints imposed by the linguistic structures comprehended or produced. From a modeling perspective, our study suggests that for each task different routines of spatial and temporal visual processing have to be accounted for. These routines are, furthermore, intimately connected to the degree of cross-modal interactivity required. Thus, we believe that models based on a combination of contextual and image-based information (Torralba *et al.*, 2006) would not be able to extend beyond search tasks.

Overall, theories of active visual perception, beside visual factors, must now include also cross-modal factors. Moreover, our results suggest that the modeling framework adopted has to be task specific, and must be able to utilize the referential information of objects to achieve the different sub-goals required.

Chapter 7

Conclusion

In this thesis, we investigated how referentiality is formed, maintained and shared across vision and language during their synchronous processing. In a range of different behavioral experiments, we unraveled the mechanisms underlying cross-modal referentiality by exploring the visual and linguistic factors involved, and the pattern of their interaction.

7.1 Contributions

Cognition is a highly integrated system, which emerges as a result of the interaction between mechanisms of different cognitive modalities. The conclusion we arrived at in this thesis is the existence of a cross-modal referential interface allowing the different modalities to communicate, share and integrate information. Our work focused on the interaction of vision and language during tasks demanding synchronous processing. We found that properties of visual referential information directly modulate processes of situated language understanding and production. And likewise, the responses of visual attention closely relate to the linguistic referential information concurrently processed.

Previous literature in situated language processing (Visual World Paradigm, VWP) has focused on linguistic phenomena, largely underestimating the active contribution of mechanisms specific to visual processing (e.g. Altmann & Kamide 1999; Knoeferle & Crocker 2006; Tanenhaus *et al.* 1995). Our first contribution was to show how image-based low-level visual information, i.e. *saliency* (Itti & Koch, 2000b; Parkhursta

et al., 2002), is activated during situated language understanding. In particular, we showed that the saliency of visual objects is used to predict upcoming linguistic referents of the sentence. When the linguistic information processed is not sufficient to make a full prediction about upcoming arguments, sentence processing resorts to visual saliency to anticipate this information. This effect is, in fact, observed around the verb site, where visual information can be used to anticipate its following arguments (e.g. direct object). This finding, beside showing a clear interaction between visual and linguistic information, also challenges evidence in the visual cognition literature, where saliency information is observed guiding visual attention only when the task performed is not goal directed, i.e. free viewing (Henderson *et al.*, 2007). In a situated language understanding task, the linguistic information processed mediates visual attention incrementally (Crocker *et al.*, 2010). Thus, when linguistic information is not sufficient to generate a prediction about all the arguments involved in the sentence (as visual referents), visual attention is relatively unconstrained, and image-based effects emerge to fill in this gap. So, more generally, we argue that image-based information is utilized when other sources of top-down information, e.g. linguistic information, are not sufficient to guide visual attention.

In a situated language understanding task, the patterns of visual attention are mainly reactions to linguistic stimuli; and this implies a rather passive contribution of visual mechanisms during the course of the task. In order to explore the active involvement of visual mechanisms during sentence processing, we moved on to situated sentence production tasks. The main advantage of studying production processes is that we can observe how the visual information of a scene is selected for linguistic encoding; thus allowing us to observe natural associations between visual and linguistic information. During production, in fact, we are able to disentangle the cross-modal factors modulating the synchronous association between linguistic, i.e. sentences, and visual referential information processed, i.e. scan patterns.

An important change, which we are the first to introduce within the VWP approach, is to situate language production tasks in photo-realistic scenes. This change is crucial in order to explore with a finer granularity cross-modal integration of visual and linguistic referential information. The complexity of a photo-realistic scene, compared to the commonly used object arrays or clip-art pseudo-scenes (e.g. Arai *et al.* 2007; Spivey-Knowlton *et al.* 2002), allows visual attention to be more realistically driven

by image (e.g. color, luminosity) and object based (e.g. co-occurrence) information. In particular, we focused on two factors shown to have an influence both on visual attention and sentence processing: scene density, i.e. *clutter* (Rosenholtz *et al.*, 2005), and object semantics, i.e. *animacy* (McDonald *et al.*, 1993). The clutter of a scene is a general indicator of visual complexity (Rosenholtz *et al.*, 2007), which, investigated during search task, is negatively correlated with target identification (Henderson *et al.*, 2009b). The animacy of objects, instead, marks a conceptual division of real-world entities, which has important implications for sentence processing, e.g. word order and grammatical function assignment (Branigan *et al.*, 2008); as well as on visual attention, i.e. target identification and temporal fixation processing (Fletcher-Watson *et al.*, 2008). We re-evaluated and unified these findings by investigating the impact of animacy of objects and clutter of the scene during situated sentence production. Beside confirming previous literature, we provided an unified explanation of their cross-modal interaction. We showed that clutter of the scene and animacy of objects are intimately connected both in visual and linguistic responses. In particular, the encoding of animate referents is facilitated by higher clutter: the more the clutter, the more referential information can be used to situate the description, i.e. more looks to the scene context occur during linguistic mention; whereas a lower clutter forces visual attention to resort to the animate referent itself to source the generation process, i.e. more looks to the animate referent during its mention. Also inanimate referents tend to benefit from higher clutter during description, but in a different way compared to animate referents. A cluttered scene provides more ground objects to spatially relate the inanimate referents. Once the linguistic encoding has started, visual attention narrows around those visual referents forming the description to avoid competition with surrounding objects; whereas in minimal scenes, the lackness of competing ground objects makes visual attention more spread.

A mechanism assumed to explain how visual and linguistic referential information is combined during situated language production is the *eye-voice span* (Griffin & Bock, 2000; Qu & Chai, 2008, 2010), which states that a visual referent is looked at shortly before its linguistic mention. We showed that this mechanism doesn't hold in photo-realistic scenes; instead, we observed the eye-voice span to be modulated by the clutter of the scene and the animacy of target object. During linguistic mention of animate referents, we found for example, that visual attention is captured by contextual scene

information rather than the referent itself. More generally, we assumed the eye-voice span to be a sub-routine of a larger process of cross-modal *coordination* which occurs only when certain visual and task related factors are satisfied, e.g. minimal scenes. Evidence of coordination has been observed on scan patterns during multi-modal tasks, such as dialogue (Richardson *et al.*, 2007) and motor-action (Land, 2006): participants align their scan patterns to synchronize the interaction between cognitive processes. We argued, more generally, that coordination emerges to synchronize cross-modal interaction. So, our hypothesis is that coordination should be observed during situated language processing, as a result of cross-modal interaction occurring between visual and linguistic referential information. Thus, we expected that during scene description, the linguistic information mentioned (sentences) is coordinated with associated patterns of visual inspection (scan patterns). Moreover, we assumed this relation to be partially independent from the eye-voice mechanism, i.e. visual objects can be fixated before or after their linguistic mention: what matters is that the sequence of objects fixated relates to the sequence of words produced. By performing a cross-modal similarity analysis, where we correlate similarity between sentences and scan patterns, we are able to show their coordination: i.e. similar sentences are associated to similar scan patterns. Crucially, we are able to show that the coordination holds both within the same scene and across different scenes, i.e. the similarity between sentences produced in different scenes positevely correlates with similarity of associated scan patterns. We conclude that the coordination observed across scenes goes beyond the known scene-based factors (bottom-up/top-down) driving scan pattern similarities (Humphrey & Underwood, 2008; Itti & Koch, 2000b), and suggests a deeper process of synchronous cross-modal alignment, which allows multi-modal cognitive processes to be organized.

A situated language processing task requires the cross-modal interaction between visual and linguistic mechanisms. However, cross-modal interaction emerges only in relation to the nature of the task performed. In fact, each task entails different subgoals, which are decisive on the cognitive modalities engaged, and the pattern of their interaction. Research in visual cognition has mainly compared visual tasks, e.g. search and memorization, observing significant differences in the eye-movement responses (Castelhano *et al.*, 2009). Both search and memorization are single modality tasks, in that only visual attention is actively engaged. However, in order to better understand

the influence of task, and its relation with cross-modal interactivity, a comparison between tasks demanding different degrees of cross-modal interactivity was needed. We compared three tasks: (1) search (i.e. find and count a cued target), (2) object naming (i.e. name the most important five objects), and (3) description (i.e. describe a cued target in relation to the scene). These tasks vary by the degree of cross-modal (visual and linguistic) interaction required. During search only visual attention is activated; whereas in object naming and description also sentence processing is involved but with a different prominence. In naming, visual attention is partially driven by the linguistic relevance of objects, whereas in description, visual attention is tightly coordinated with the structure of the sentence concurrently processed. We tested, and supported, the hypothesis that cross-modal interaction triggers a more complex pattern of visual referential processing by looking at several eye-movement measures, e.g. first pass fixation duration, and scan pattern similarities within/between tasks. Moreover, motivated by previous results, we argued that density of scene and animacy of objects should modulate visual responses according to the task being performed. In order to rationally interpret eye-movement measures in relation to the visual processing triggered, we introduce a distinction between the temporal and spatial components of visual processing. Measures for the spatial component look at *inspection*, e.g. spatial distribution of fixation, and indicate how wide visual sampling is for a certain task. Measures for the temporal component look at *integration*, e.g. mean fixation duration, and refer to the complexity of visual processing needed to perform a certain task. We showed that the cross-modal interaction needed during naming and description, require more complex temporal processing, as visual and linguistic referential information has to be integrated. Moreover, the cross-modal interaction of these tasks results in a higher scan pattern similarity compared to search, a single modality task. The joint activation of visual and linguistic relevance allows for a stronger guidance of visual attention, which makes different participants more coordinated in their scan patterns. On the spatial component, we find that naming has a wider visual sampling compared to search and description. Overall, the goal of a task, e.g. name the most relevant five objects, has a direct impact on the spatial component, whereas its cross-modal interactivity is reflected by the complexity of temporal processing needed to integrate referential information across modalities. In light of these findings, theories of active visual perception (e.g. Malcolm & Henderson 2010; Schmidt & Zelinsky 2009) and

situated sentence processing (e.g. Altmann & Mirkovic 2009; Tanenhaus *et al.* 1995) must now move toward a more unified framework, which is able to explain the influence of both visual and linguistic factors in relation to the type of task performed, and its goals.

7.2 Future work

In this thesis, we have demonstrated the existence of a cross-modal referential interface upon which scene understanding and sentence processing can be coordinated. The coordination on reference allows different cognitive processes to be synchronized. More importantly, coordination permits an efficiently organized communication, as information have to be 'standardized' in order to be optimally conveyed. So, the more similarly we behave, the more coordinated the resulting communication will be. Importantly, we demonstrated that such similarity crosses the domain of a single modality, and it extends over the different modalities engaged by the task performed.

The main consequence of our demonstration is that even if modalities might be governed by independent mechanisms, they nevertheless share correlated patterns of processing. So, any theory of cognition that aspires to integrate multi-modal processing within the same account, has to explain: (1) the formation of correlated cross-modal patterns, (2) the mechanisms modulating the strength of such correlation, and (3) the range of multi-modal factors involved. The work presented in this thesis lays the foundation to systematic research on cross-modal processing; but its theoretical reach is limited to a general understanding on how cross-modal processing can be experimentally examined and computationally quantified. More specific questions elucidating the interplay between different components (conceptual, representational and computational) involved in cross-modal similarity are needed.

In particular, during our exploration of the coordination between vision and language, we observed a non-linear (sigmoid-like) trend of cross-modal similarity. This trend displayed an extended intermediate plateau region with constant values of correlated similarity, and extremes with sharp changes (positive and negative), where a minimum change of similarity in one modality was correlated to sensible changes on the other modality. This finding is particularly intriguing, as it suggests the presence

of different weighting factors, both visual and linguistic, mediating coordination. As a first diagnostic, we propose to divide pairs of sentences/scan patterns into classes, according to their cross-modal similarity value; thus, we can distinguish intermediate values from extremes. Then, we suggest to explore more in depth which factors could be implicated (positively or negatively) on the cross-modal similarity value for the different classes. These factors might be conceptual, as we have shown with animacy, but they can also relate more generally to the accessibility and processing of event knoweldge, i.e., some scenes foster coordination more than others. In situated language production, in fact, accessibility of an event strongly depends upon the visual material displayed in the scene: how such material is spatially organized, which objects are contained, how many can be recognized or classified, their familiarity and the contextual relations they entertain with one another (just to mention a few). Thus, we suggest that through a series of different experiments, spanning both behavioral and brain imaging, testing the role of event context, it would be possible to provide a taxonomy of the factors involved in cross-modal similarity, and quantify their relative weights.

Obviously, the accessibility of the event is not isolated from the production task performed. We investigated production cued on specific targets. This decision was taken to bound the generation within the information displayed by the scene, and limit the actual sentence production to mere descriptions. However, in other production tasks, such as free production, we expect a different pattern of coordination to arise. Especially, we expect a sensible increase in the variability of both sentences and the associated scan patterns, resulting from the use of associative and episodic memory (i.e., an object could be described relative to a fictitious past [1]), and freedom of linguistic selection both of the referents, and their syntactic construction (e.g., active vs passive). Even if a free production task is expected to show less coordination than a cued task, it would be crucial to unravel their underlying similarities; as the pool of sentences shared by both tasks would represent 'optimal descriptions' of the corresponding scenes, i.e., descriptions generated both with and without explicit cueing.

[1] Hints suggesting this idea come from a follow-up free production web-experiment, which has not been reported in this thesis.

A basic question about cross-modal similarity of visual and linguistic processing, and more generally about their synchronous interaction, regards the way visual and linguistic information are represented. Both types of information has syntactic organization and semantic form; but they are quite distinct for the two modalities. Thus, an important step to bring us closer to the understanding of cross-modal similarity is to unify the syntactic and semantic representation of visual and linguistic information. In order to address this issue, the first challenge is to assign a syntactic representation to visual information. At the state of art, we can compare semantic information of visual and linguistic processing in the form of referential sequences (see Chapter 5), but syntactic information of sentences does not extend to visual information, as a verb phrase like *eat* is represented by a **configuration** of visual objects such as a MAN with the MOUTH open, and his HAND holding an APPLE. Thus a 'syntactic' representation shared between visual and linguistic information must be able to incorporate these patterns of visual configurations with the identity of the linguistic constituent. By finding a syntactic representation shared between vision and language, we would also be able to solve the problem of one-to-one multimodal mapping between visual and linguistic information, when modeling techniques, such as Hidden Markov Models, are applied.

The studies presented in this thesis have focused on the English language. However, if cross-modal coordination is a general mechanism of cognition, we should be able to observe it across different languages. The advantage of studying coordination across languages is that we can investigate phenomena occurring at the syntactic and semantic component of sentence processing by looking at cross-linguistic similarity. A concrete extension of these ideas could be a cross-linguistic comparison between languages which have a different syntactic ordering, like English and Japanese; and a test whether similarity of scan patterns emerges in relation to sentences with identical semantics but a different syntax. If such similarity is found, we can conclude that visual attention organizes its referential representation independently of syntactic mechanisms, but based on the meaning of the event. In order to actually quantify such similarity, we suggest to include in the cross-linguistic comparison another language, like Italian, which has a similar syntactic ordering to English, and could be used as a reference baseline.

Throughout the thesis, we have performed statistical regression modeling to quantify the factors involved during cross-modal processing, using the descriptive approach.

In order to situate the cross-modal interaction between sentences and scan patterns in a generative framework we might approach it in terms of graphs, and focus on simulating a specific task, e.g. object naming. A scan pattern can be represented as a directed graph, where the nodes are the visual objects looked at, which can be enriched with more information such as fixation duration, and the edges connecting the nodes are the transition probabilities of saccading to another object. A starting representation for a sentence might be a semantic network where the content words are the nodes, and the edges connecting them are co-occurrences probability of other content words related to them. The main challenge, however, is to join the two resulting graphs to perform measures of their connectivity, while attempting to simulate sentences from scan patterns, and vice-versa. In this generative approach, a combination of methods from graph theory, graphical models, and markov processes can be used to align visual and linguistic processing.

In general, our work has shown that it is possible to unify findings across different independent fields, such as human sentence processing and visual cognition. By identifying the common mechanisms allowing cross-modal interaction, we aim to provide a more integrated understanding of the architecture of cognition.

Chapter 8

Experimental Material

Sentences used in Experiments (1-3)

The boy will put the pillow on the table in the box
The girl will put the orange on the tray in the bowl
The boy will put the pasta on the plate in the colander
The girl will put the griddle in the oven on the table
The boy will put the bottle in the freezer on the shelf
The girl will put the sausage in the pot on the platter
The boy will move the apple on the towel in the box
The girl will move the pen on the folder in the box
The boy will move the cake on saucer in the bowl
The girl will move the key on the envelop in the closet
The boy will move the salt-shaker on the envelop in the drawer
The girl will move the flower on the newspaper in the freezer
The boy will place the pencil on the erase in the cup
The girl will place the spoon on the napkin in the bowl
The boy will place the coin on the cash in the cup
The girl will place the coin in the vase on the paper
The boy will place the flower in the glass on the newspapers

The woman will place the apple in the colander on the napkin
The boy will lay the pencil in the vase on the cash
The girl will lay the salt-shaker in the briefcase on the tray
The boy will lay the spoon in the bin on the platter
The girl will lay the pineapple in the briefcase on the carpet
The boy will lay the pie in the luggage on the plate
The girl will lay the ruler in bin on folder
The boy will put the lighter on the jacket in the wardrobe
The girl will put the agenda on the lantern in the bag
The boy will put the parsley on the fish in the aquarium
The girl will move the nuts in the jug on the cooker
The boy will move the chopsticks in the pitcher on the microwave
The girl will move the shrimp in the can on the chair
The boy will lay the candle in the candelabra on the cabinet
The girl will lay the socks on the blanket in the dryer
The boy will lay the lobster on the rug in the basin
The girl will place the corn in the jar on the desk
The boy will place the clip on the curtain in the hamper
The girl will place the burger in the basket on the bench

Set of Images used in Experiment (1-3)

Sentences produced in Experiment 5

The woman is weighing herself
The woman is sitting on the bed
The man is cutting the pie
The man is paying for his room
Kid is playing the drum
The man is sorting things out
The sponge is next to the bath
The towel is on the basket case
The orange juice is in glasses
The suitcase, suitcases, are in the reception of the hotel
Th.. The knives are on the table
The book is lying open on the table
There are two babies in the bathroom
Man is lying down
The man is writing notes
Man is sitting in the chair
The man is cutting up meat
The man is tired at work
The bucket is in the blue basket
The teddy is in the girl's harms
The juice is in jugs on the table
The flowers are on the table
The fruit is on the table
The phone lies next to the screen
A man is washing the bath
The girl is hugging a teddy bear
The woman is looking at a pile of paper
The woman is paying an hotel bill
The woman is making a salad
The woman is talking on the telephone
The woman is using the scale to weigh herself

The shoe is on the floor
The men are drinking soup
The lamp is next to the reception desk
The baby is playing with the apple
The mug is on top of the in-tray
The woman is reading some letters
The kid is sleeping
The woman is eating breakfast
The man is checking out of the hotel
The man is carving a chicken
A woman is cleaning the floor
The baby is playing with the toilet-paper
The mobile is on the bed
The man is writing on a clipboard
The telephone is on the desk
Void
The man is working on his laptop
The baby sat on the toilet
The man laid upon the bed while his friend read the newspaper
The man filled out an application
A man sits in a rocking chair
The man prepares food in the kitchen
The man sat exasperated at his computer speaking to his friend
The man cleaned the tub rinsing his rug in a bucket
The girl resting upon the bed held her teddy tightly while another teddy sat on the ground
As the two women thought the pens sat neatly on the table next to them
The hotel reception was decorated with flower buds
The woman cuts the fruit
The woman at the desk spoke on the phone while another phone sat unused next to her
The woman weight herself on the scale while her friend redid her face
This catholic glad woman sat on the edge of the bed talking to her friend
The man ate dinner at his table

The man spoke to the employ at the registration desk at the hotel
The kid banged his drumstick against the floor
Void
The woman used the sponge to clean the counter top
The towel was placed next to the child who was laying upon the bed
The female drinks juice with her breakfast while speaking to a female friend
As the man checked into the hotel his suitcases sat by his feet and another sat on the counter
The man dissected the roast chicken with his knife
The book sits open on the desk
A woman washing glass
A kid on a bed
A woman sitting in a table drinking orange juice
A man standing at an hotel counter
A man preparing chicken
Woman cleaning the floor
Baby playing with toilet paper
A mobile phone sitting on a bed
Soldier writing on a clipboard
Telephone on a pedestal
A man preparing a waffle
A laptop on a desk
A man washing clothes in a bathtub
A girl holding a bonny rabbit
A woman talking on the phone
A woman at the counter
A woman preparing food
Woman in an office
A woman standing on a scale
A shoe in a box
A man eating soup
A lamp on a table
An apple on a plate

A mug on a conference room table
The woman stands on the scale
The woman is sitting on the bed
The man is standing at the dinner table
The man is at reception
The kid plays the drum
The man is on the desk and counting out some leaflets
The sponge is next to the bath
The towel is on the end of the bed
There is juice on the table
There is a suitcase on reception and on the floor
The knife is on the chopping board
The book is on the table
The baby is on the toilet
The man is lying on the bed
The man is sitting on the sofa
The man is at the reception
The man is in the kitchen
The man is in his office
The bucket is on the floor
The child is holding a teddy
There is juice on the mantel piece and on the table
The flower is sitting on the reception
There is fruit on the table
The woman is on the phone
The man is bringing his clothes in the bathtub oh no cleaning the bathtub
The girl holding the teddy bear is sitting on the bed and the girl in the nighty is next to the bed
The woman is sitting in a chair in the corner
The woman is chatting to the reception
The woman is having coffee and the other woman is arranging the food on the table
The woman approaches the receptionist who is on the phone
The woman is standing on the scale in the bathroom

The running shoes are scattered in the bedroom and the woman is wearing black heel shoes
The soup is on the dining room table
The lamps are on the side tables
The apple is in the toddler's bowl and the other apple is on the counter
The two mugs are on the desk holding stationery
The women are in the bathroom
The kids are in the bedroom playing and sleeping
The woman is having breakfast with another woman at the table
The man is signing papers at the lobby
The men are sitting at the kitchen counter, wow, one man is sitting at the kitchen counter and the other man is carving a chicken
The woman is sitting on the chair
The children are playing with the toilet-paper in the bathroom
The mobile is open on the bed
Two clipboards are on the table of a living room
One telephone is on an elaborate side table and the other telephone is on the counter
The chef on the left is making a waffle
The laptop is on the office desk
The baby is in the bathroom
The man lies in the bed
The man is filling a paper
The man sits waiting in the chair
The man is working in the kitchen
The man is at the office
There is a bucket in the bathroom
Teddy is on the ground
The woman has a pen in her hand
On the table sits the beautiful flowers
The woman puts fruit on the table
The woman is on the phone
The woman is on the scale
The woman sits on the bed

The man is drinking wine
The man is working at the hotel
The kid is in the kitchen
The man is working in the office
The sponge is used for cleaning
The towel lays bordered on the bed
The woman drinks orange juice for breakfast
A small suitcase is being carried by the man
The knife sits on the cutting board
The book is on the shelf
Two, there are two, there is one woman cleaning bathroom taps while another woman is doing something that looks like the same
One kid is sleeping as the other kid is beside him playing
There is a woman drinking orange juice and another woman speaking to her
One man waits for another man to sign in to, fill out the registration form for a hotel
There is a man preparing a turkey while another man is smelling it
There is a woman sitting in a chair as well as a woman cleaning the waiting area
One child is wasting toilet-paper
A man laid down on his bed with his mobile beside him
One man writes on a clipboard, as another man observes a piece of paper
There is a telephone in front of a man, who is checking another man into a hotel
Two, one chef is preparing a waffle
There is a man scratching his head in front of his laptop
One man is washing something in the sink while another man is observing
There is a girl hugging a teddy bear while another girl observes her
One woman sits in the chair relaxing as another woman sits on the phone
There is a woman ahead waiting to check in an hotel room and she is being served by another woman
There is a woman standing up and drinking coffee while another woman prepares appetizers
There is a woman talking with someone at the phone as another woman waits for her attention
A woman is standing on the scale and another woman behind her

There is a shoe on the floor beside two women
A man is about to serve another man soup
There is man checking into a hotel in front of a lamp
Children playing in the kitchen
There is a mug on the desk behind the man who is sorting out files
There is a woman weighing herself on the scales
There is a woman sitting on the bed
There is a man drinking wine, waiting for his food
There is a man in a pink shirt serving another gentleman
There is a kid playing the drums and a kid playing with a fruit bowl on the floor
There is a man unpacking a box
The woman is cleaning the sink with a sponge
There is a towel as well as other towels folded on the bed
There are two glasses of juice in this picture
There is a suitcase on the desk and a suitcase by the gentleman's feet
A gentleman is cutting chicken with a knife
There is a book sitting open on the desk
There is a baby playing with the toilet-roll
There is a man resting on the bed as another man unpacks his clothes
There is a man taking notes
There is a man relaxing in a chair
There is a man chopping meat on the counter
The man is overworked
There is a blue bucket in the basket
The little girl is sitting on the bed cuddling her teddy
There is a jug of juice sitting on the table
There is a very pretty yellow flower in the vase
There is a lot of different kind of fruit in the fruit bowl
The woman is talking on the telephone
The man cleaned the bathtub as his friend watched
The girl hugged her teddy
The two women sat in the empty room
A woman arrived at her hotel and was greeted kindly

The woman tried some of dinner she just made for her friend
The woman went to the help desk
The scale in the bathroom was stood on by the woman
The girl has tried on her new shoes
The soup was made and placed on the table
The lamps in the foyer were old but beautiful
The apple on the counter went unnoticed
The mug was new
The two women cleaned the bathroom together
The kid played while his baby brother slept
The woman enjoyed her breakfast while the nanny arrived for the day
The man at the hotel desk helped the tourist check-in
The man made lunch for his friend
The woman cleaned the floor in the office
The baby used too much toilet-paper when using the toilet
The man's mobile phone rang
The empty clipboard laid beside the two soldiers
The telephone in the hotel lobby began to ring
The waffle was made especially that morning
The laptop stopped working
The baby is playing with the toilet-paper
The man is in police uniform
The man is writing
The man is waiting at the reception
The man is cooking
The man is tired
A man is cleaning the bucket
The girl is holding a teddy
The pen is lying on the sofa
The flower is in the vase
The fruits are lying on the table
The girl is talking on the phone
The woman is weighing herself

The woman is looking at the toys
The man is making a meal
The men are at the reception
The kids are playing
The man is in an office
The girl is cleaning with the sponge
The towel is yellow, coloured
The juice is on the table
The suitcase is lying on the ground
The knife is lying on the table
The book is lying opened
There is a woman scrubbing a bathroom sink with a sponge, and another woman who looks to be dusting a wall
There is a little kid passed out on his mum's and dad's bed, and a kid standing next to him with some sort of holes in the crouch of his pants
There is a woman sitting at the table drinking a juice, laughing, and an older woman facing her, she looks like she is about to say something
A man is standing behind a reception desk facing directly forward, there is another man to his right signing in a form on the desk
There is a man preparing what looks to be a chicken or a turkey to eat and another man sitting across from him
There is a woman who looks to be mopping a floor pushing a broom in the left hand corner and another woman sitting at ehh in a chair at a table eating toast
There are two small children playing with toilet-paper in the bathroom
There is a man asleep on a bed and an open mobile next to him, there is another mobile perched on the ehh perched on the wall, sorry, charging
There is a clipboard sitting on a coffee table and another clipboard next to it on which a US army man is writing
There is some sort of reception counter and a telephone on the reception counter next to a man, there is another telephone on a pedestal to his left
There is a Japanese man, I think, in a kitchen cooking waffles
There is a man leaning back sitting on a desk whereas a laptop is opened in front of him there is another opened laptop ehh further to the front of the picture

There is a man leaning over into a bathtub, he looks like is dipping a cloth into the water and there is another man with his back to the scene near many cleaning products
There is a girl sitting on her bed hugging her teddy bear and another little girl next to her, she looks like she wants the teddy bear
There is a woman in a party jazz lying back against a chair and another woman directly across from her on the phone in a living room, with a lot of liquor on the mantel
There is a woman checking into a hotel she is handing her credit card to another woman that is standing behind the desk
There is a woman making a platter of fruit in the kitchen and another woman drinking a cup of tee and talking to her
There is a woman talking on the phone behind her desk and another woman who looks to be waiting to see her when she is done
There is a woman standing on a scale in a bathroom facing the sink and the mirror
There is one shoe in a filing box next to a woman on a bed in a bedroom and another shoe on the floor near another woman also in the bedroom
There is a man serving soup to his friend
There is a man at the desk in what looks to be an hotel and there is a bed side table with a lamp on top of it next to him ahh there is also another lamp in a sitting area to his right
There is an apple sitting on a plate on a kitchen counter and two little boys playing on the floor with some toys
There is a mug sitting on an office desk with a bunch of pencils and rulers and office supplies in it, and another mug across from it with the same contents, there are also two larger mugs that look like travel mugs on the desk
The woman is on the scales and the woman is standing
The woman is sitting on the bed and standing
The man lifts the dish from the table
The men are standing
The kids are sitting on the ground playing
The man is standing reading and the man is sitting
The sponge is being used to clean the bathroom
The towel is on the bed and on the sh.. drawer
The women are drinking the juice

The suitcases on the desk and on the floor
The knife is on the bread board
The book is on the table
The babies are in the bathroom playing
The man is reading the newspaper and lying on the bed
The men are sitting
The man is sitting and the man is standing
The men are chopping food
The man is in his office leaning back
The bucket in the picture is next to the cleaning products and in a basket
The teddy is on the floor next to the bed and being held by the girl
The juice is in two jugs on the table and the mantel
The flowers are in a vase on the desk and on the table with the chairs
The fruit is on the table in two baskets
The woman is talking on the phone
The man washing the clothes in the bath
The little girl on the bed cuddling a teddy, two little girls
One woman is on the phone while the other is sitting relaxing in her seat
Two women talking at the front desk
The woman standing drinking tea
There is a woman standing by the desk and another like woman on the phone
A set of weighing scales, two sets of weighing scales in the bathroom
Two shoes on the cha.., on the bed and on the floor
The man dishing soup to his guest
There are two lamps on, one on the table and one on the cabinet
There is an apple on the bowl, on the floor with the child and an apple on the work surface
There are two mugs on a desk with stationary in them
There is one woman cleaning in shorts and one woman polishing glass
There is one little kid asleep on the bed and another standing, playing
One woman is sitting at the table eating breakfast and another is standing up
There is a man signing papers at the front desk and another man sit.. ehmmm that works at the front desk

There is one man preparing chicken and one sitting
There is one woman eating and one woman standing, cleaning I think
The child playing with the toilet-paper on the loo seat
A mobile phone charging at the wall
The man is writing information, soldier I guess or marine, was writing information on the paper on the clipboard
One telephone on the stand and one telephone on the desk
The chef is making waffles and there is waffles in the microwave
The laptop on the desk
The baby is playing with toilet-paper
The man is lying in bed
The man is signing a paper
The man is near a fax machine
The man is wearing a yellow apron
The man is tired
A bucket is blue
The girl is holding a teddy bear
The pen is in the woman's hand oh no the pen is on the table
The flowers are yellow
The woman is picking a piece of fruit
The woman is on the phone
The woman is weighing herself
The woman is sitting on the bed
The man is serving dinner
The man is checking in
The kid is playing with sticks
The man is emptying the box
The sponge is on the edge of the bathtub
The towel is folded on the bed
The woman is having juice for breakfast
The suitcase has red edges
The knife is sitting on the cutting board
The book is open on the table

A woman and her friend busy cleaning a bathroom
A kid asleep at the foot of a double bed and another kid staring at the camera
A woman drinking orange juice sitting on a seat speaking to another woman holding an handbag
A man stands behind a reception desk while a second man scribbles some notes in a pad
A man sat at the table watching another man prepare a chicken
A woman cleaning a public area while another woman sits with a plate holding toasts
Toilet-paper has been unravelled by an unruly toddler on top of a lavatory
A mobile phone occupies its hoister on the wall of a hotel room while a second mobile phone lies at the hand of a Chinese male
A man jots notes on a clipboard while his friend observes a second clipboard and a third clipboard lies unused on a coffee table
A man sits behind... stands behind a counter with a telephone while a second telephone rests a top a pedestal
A chef preparing a waffle while a separate chef carves a chicken
Two laptops are currently unused in an office space
A man cleaning a bath and another man possibly chatting to him
A girl sits cross legged on a bed clutching a teddy bear while another girl looks thoughtfully at the room
A woman is relaxing at home while another woman, possibly her mother, speaks on the telephone
A woman behind a reception desk is speaking to another woman checking in to the hotel
A woman sips coffee while another woman idly interacts with food
A woman speaking on the phone and another woman waiting at her desk
A woman weights herself upon a scale while a second scale lies to her right
A woman wearing high heel shoes sits on a bed next to a box of new shoes
Two men sit preparing to eat soup
A lamp a top a side table in the lobby of a guest house
A child sitting on the floor playing with a pot holding an apple
A mug on an office table
The woman is on the scales

The woman is on the bed
The man is serving food
The man is at the desk
The kid is playing with the drum
The man is standing
The sponge is on the bath
Towel is on the bed and on the chest
Woman is drinking juice
The suitcase is on the floor
The knife is on the board
The book is on the table
Baby is on the toilet
The man is on the bed
The man is writing
The man is sitting in the chair
The man is cooking
The man is tired
There is a bucket next to the bath and a bucket next to the door
The girl is holding a teddy
Juice is in the jug
The flower is on the table
Fruit is on the table
Woman is on the phone
There is a man watching another man in the bathroom
There is a girl holding a teddy on the bed
There is a woman on telephone
The woman is booking into reception
There is a woman who is preparing the fruit-salad
There is a woman answering the telephone
There is a woman on top of the scales
There is a shoe in a box
The man is preparing the bowl of soup
There is a lamp beside the reception

The child has an apple in the pot
There is a mug filled with stationary
There is a woman who is scrubbing down the worktop
There is a kid asleep on the bed
There is a woman eating her breakfast
There is a man signing off papers at the reception
There is a man preparing a chicken
There is a woman cleaning the floor
The child is playing with the toilet-paper
There is a mobile phone open on the bed
There is an army office writing on the clipboard
There is a telephone behind the receptionist
The chefs are preparing waffles
There is a laptop open on the desk
The baby is sitting on the toilet
The man is lying on the bed
The man is filling out paperwork
The man is sitting in the lobby
The men are cooking
The man is stretching
The man is using a bucket to clean with
The girl is holding the teddy
The pen is sitting on the chair
There are flowers on the side table
The woman is eating fruit
The phone is being used by the lady
The woman is on the scale
The woman is looking at the teddy bear
The man is drinking wine
There is a man behind the desk
The kid is playing the drums
The man is working
The lady is cleaning with the sponge

There are two towels on the bed
The lady is drinking juice
There are two suitcases in the lobby of the hotel
The man is using a knife to cut the turkey
There is a book on the table
The woman and her friend clean the bathroom
The kid lies on the bed
The woman drank a glass of orange juice
The man stood behind the desk
The man used the knife to cut the meat
The woman cleaned the floor
The baby used all the toilet-paper and trashed it over the floor
The man used his mobile to call his friend
The man wrote on his clipboard
There was a telephone on the polished desk
The chef prepared a waffle
The laptop was on the desk
The man washed in the bath
The girl sat on the bed hugging her teddy
The woman sat on the chair and looked at her wine collection
The woman greeted the receptionist
The woman drank a cup of tea
The woman was greeted by the receptionist
The woman stood on the scales
The girl looked at the new shoe she bought
The man enjoyed his soup
There was more than one lamp in the well lit reception area
There was an apple on the counter
There were many mugs on the desk
The woman is in the bathroom, one woman weights herself and another woman is moisturising
The woman is sat on the bed next to another woman
The man is sat at the table with another man

The man books into the hotel
The kid is playing with another kid in the kitchen
The man is sorting two files
The sponge is by the bath
There are two towels on the bed and two towels on the basket
The woman drinks the juice in the morning
There is a suitcase in the hotel, the customer has two suitcases
The knife is being used in cooking
There is a book on the side
The baby is playing with the toilet-roll
There is a man sleeping in an hotel room and another man reading a newspaper
The man is signing papers in the army
The man is relaxing in a hotel lobby
The man is cooking with another man
The man is tired of work
There is a bucket in the bathroom being used in cleaning the bathroom
The girl has lots of teddies
There are two jugs of juice in the room
There are flowers in the reception of the hotel
They have fruits on the table
The phone is being used by a receptionist
One man watches another man wash an item in the bath
One girl watches another girl sitting on the bed
One woman talks on the phone whilst another sits on the chair
One woman checks on with another woman
One woman drinks tea whilst another prepares dinner
One woman waits to speak to another woman who she is in the phone
A woman weights herself on the scales
One shoe is in a box on the bed another shoe is beside of the bed
There is soup in the bowl
There is a lamp to the left of the man, another lamp is beside the stairs
There is an apple on the counter, another is in a bowl
There are two mugs containing stationary on the desk

A woman cleans the bathroom, another woman helps
One kid sleeps on the bed whilst another plays
One woman stands while another woman sits and eats her breakfast
One man is standing at the desk the other is signing a paper
Void
One woman eats whilst another woman cleans
The child plays with the toilet-paper
The mobile is on the bed next to the man
Two men hold, clipboard each, another clipboard rests on the table
There is a telephone on the desk, another telephone is on a stand
The chef prepares some waffles
A laptop sits on the desk, another laptop is on top of a box
There is a baby learning how to walk, within a toilet, while another baby is potty training himself
There is a man asleep on a bed, with his phone open, while another man stumbles reading a paper in a hotel room
A man signing a form in army costume next to an older member possibly a senior
There is a man sitting in the lobby of a reception while another man is on the computer
There is man preparing food in a kitchen with another man also doing the same
There is a tired man at work being chatted to by his boss in a very blank bare office
There is a man cleaning the bath with two buckets
There is two girls in a bedroom, one of which hugging a teddy the other has been left by end of the bed
A pen has been left on a table while two women sit in a very empty living room
A woman is handing another woman room-key for the hotel next to a set of lovely flower
A woman is about to have some fruit while chatting to her friend in a kitchen
There is a woman on a phone at reception with a customer waiting
There is a woman weighing herself at the sink within bathroom while her friend does some cleaning
There is woman sat on the bed and a woman standing up within a bedroom of, furniture
There is a man serving food to an empty table where one man with wine is about to sit

down
There is a man at reception talking to another man behind the counter within a hotel lobby
There is kid pretending to play the drums with his little brother playing with toys in the kitchen
There is man standing in an office, unpacking a box, while his colleague does some work on a computer
There is a sponge by the bath with two women cleaning it
A folded towel was left on a bed next to an unattended baby with his older brother playing with the toy
A woman is drinking juice at a coffee table chatting to her friend
A man with a red suitcase is signing in to a hotel while the receptionist poses for the camera
A knife is being left on a board, one man uses another knife to prepare a kitchen for his friend
A book is being left open on a table while a woman is eating some food in a chair with a cleaner behind her
One woman is cleaning the bathroom while the other one cleans the glasses
One kid is sleeping on the bed and the other one is playing
The woman is drinking orange juice
The man is checking out from the hotel
One man is cutting the chicken while the other one looks at him
The woman is eating while she waits
The kids are playing with the toilet-paper
The person on the bed was holding the mobile
The military is writing on his clipboard
There are two telephones in the lobby of the hotel
The chefs are cooking ... one of the chef is cooking waffles
The laptop is on the desk
The man is washing the tube while the other man looks at him
One girl is on the bed hugging a teddy bear and the other one is on the floor looking at her
One of the women is talking on the phone while the other one is looking at the flowers

The woman is checking in to the hotel
One of the women is eating fruit while the other one drinks tea
The woman waits while the other one is on the phone
There are two scales in the bathroom
There is one shoe right next to the bed and one shoe inside a box
The waiter is serving soup to the guest
There are several lamps in the room
The apple is on the counter
There are several mugs in the office

Set of Photo-realistic Scenes used in Experiment 5

Minimal

Cluttered

Bibliography

AGRESTI, A. (2007). *An introduction to categorical data analysis*. John Wiley and Sons, Ltd. 30

ALEXANDER, R., ZHANG, W. & ZELINSKY, G. (2010). Visual similarity effects in categorical search. In *R. Catrambone & S. Ohlsson (Eds.), Proceedings of the 32th Annual Conference of the Cognitive Science Society, Portland*. 10

ALTMANN, G. & KAMIDE, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition*, **73**, 247–264. 3, 7, 8, 18, 40, 43, 58, 131, 201

ALTMANN, G. & MIRKOVIC, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science*, **33**, 583609. 3, 5, 41, 206

ANDERSON, D.R. (2008). *Model Based Inference in the Life Sciences: A primer on evidence*. Springer. 31

ARAI, M., VAN GOMPEL, R. & SCHEEPERS, C. (2007). Priming ditranstitive structures in comprehension. *Cognitive Psychology*, **54**, 218–250. 8, 18, 22, 41, 202

BAAYEN, H.R. (2008). *Analyzing linguistic data. A practical introduction to statistics using R.*. Cambridge press. 31

BAAYEN, R., DAVIDSON, D. & BATES, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, **59**, 390–412. 30, 100, 111, 140, 165, 168

BADDELEY, R. & TATLER, B. (2006). High frequency edges (but no contrast) predict where we fixate: A bayesian system identification analysis. *Vision Research*, **46**, 2824–2833. 9

BAILEY, K. & FERREIRA, F. (2007). The processing of filled pause disfluencies in the visual world. In *R. Van Gompel & M. Fisher & R. Hill & W.Murray (Eds.), Eye movements: a window on mind and brain (Vol. 2003)*, Elsevier. xi, 40, 46, 48, 60, 61, 62, 70, 73

BAR, M. (2004). Visual objects in context. *Nature Reviews Neuroscience*, **5**, 617–629. 6

BARR, D. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of memory and language*, **59**, 457–474. 18, 29, 36, 52, 111

BARSALOU, L. (1999). Perceptual symbol systems. *Behavioral and brain sciences*, **22**, 577–660. 5

BRANIGAN, H., PICKERING, M. & TANAKA, M. (2008). Contribution of animacy to grammatical function assignment and word order during production. *Lingua*, **2**, 172–189. 93, 96, 97, 107, 123, 124, 132, 150, 189, 203

BROCKMOLE, J. & HENDERSON, J. (2006). Recognition and attention guidance during contextual cueing in real-world scenes: Evidence from eye movements. *Quarterly journal of experimental psychology*, **59**, 1177–1187. 7, 9, 129, 160

BURNHAM, K. & ANDERSON, D. (2002). *Model Selection and multimodel inference*. Springer. 31

CASTELHANO, M. & HEAVEN, C. (2010). The relative contribution of scene context and target features to visual search in real-world scenes. *Attention, Perception and Psychophysics*, **72**, 1283–1297. 129, 149

CASTELHANO, M., MACK, M. & HENDERSON, J. (2009). Viewing task influences eye-movement control during active scene perception. *Journal of Vision*, 1–15. 3, 9, 16, 90, 129, 130, 159, 160, 161, 179, 199, 204

CLARK, H. (1973). The language as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, **12**, 335–359. 29

CLIFTON, C., STAUB, A. & RAYNER, K. (2007). Eye movements in reading words and sentences. In *R. Van Gompel & M. Fisher & R. Hill & W.Murray (Eds.), Eye movements: a window on mind and brain (Vol. 2003)*, Elsevier. 17

COCO, I., M. & KELLER, F. (2009). The impact of visual information on referent assignment in sentence production. In *In N.A. Taatgen and H. van Rijn (Eds.), Proceedings of the 31th Annual Conference of the Cognitive Science Society, Amsterdam*. 14, 21

COCO, M. & KELLER, F. (2010a). Scan patterns in visual scenes predict sentence production. In *R. Catrambone & S. Ohlsson (Eds.), Proceedings of the 32th Annual Conference of the Cognitive Science Society, Portland*. 14

COCO, M. & KELLER, F. (2010b). Sentence production in naturalistic scene with referential ambiguity. In *R. Catrambone & S. Ohlsson (Eds.), Proceedings of the 32th Annual Conference of the Cognitive Science Society, Portland*. x, 14, 18, 21, 133

CRAWLEY, M. (2007). *The R book*. John Wiley and Sons, Ltd. 31

CRISTINO, F., MATHOT, S., THEEUWES, J. & GILCHRIST, I. (2010). Scanmatch: A novel method for comparing fixation sequences. *Behaviour Research Methods*. 25, 136

CROCKER, M., KNOEFERLE, P. & MAYBERRY, M. (2010). Situated sentence comprehension: The coordinated interplay account and a neurobehavioral model. *Brain and Language*, **112**, 189–201. 3, 8, 41, 46, 202

DAGAN, I., LEE, L. & PEREIRA, F. (1997). Similarity-based methods for word sense disambiguation. In *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, 56–63, Columbus, Ohio. 167

DEMBERG, V. & KELLER, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, **101**, 193–210. 17

DURBIN, R., EDDY, S., KROGH, A. & MITCHISON, G. (2003). *Biological sequence analysis. Probabilistic models of proteins and nucleid acids*. Cambridge University Press, Cambridge,UK, 2nd edn. 25, 135

EHINGER, B., HIDALGO-SOTELO, B., TORRALBA, A. & OLIVA, A. (2009). Modeling search for people in 900 scenes: a combined source of eye-guidance. *Visual Cognition*, **17**, 945–978. 7

EINHUSER, W., SPAIN, M. & PERONA, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision*, **8**, 1–15. 10

ELAZARY, L. & ITTI, L. (2008). Interesting objects are visually salient. *Journal of Vision*, **8**, 1–15. 10, 47

EVANS, K. & TREISMAN, A. (2010). Natural cross-modal mappings between visual and auditory features. *Journal of Vision*, **10**, 1–12. 44, 81, 130

FERREIRA, F., CHRISTIANSON, K. & HOLLINGWORTH, A. (2001). Misinterpretations of garden-path sentences: Implications for models of sentence processing and reanalysis. *Journal of Psycholinguistic Research*, **30**, 570–595. 17

FINDLAY, J. & GILCHRIST, I. (2001). Visual attention: The active vision perspective. In *M. Jenkins & L. Harris (Eds.), Vision and attention*, 83–103, Springer-Verlag, New York. 3, 9, 42, 129, 159, 199

FLETCHER-WATSON, S., FINDLAY, J., LEEKAM, S. & BENSON, V. (2008). Rapid detection of person information in a naturalistic scene. *Perception*, **37**, 571–583. 96, 107, 125, 150, 163, 180, 182, 183, 203

FORSTER, K. & MASSON, M. (2008). Introduction: Emerging data analysis. *Journal of memory and language*, **59**, 387–556. 30

FRANK, M., VUL, E. & JOHNSON, S. (2009). Development of infants' attention to faces during the first year. *Cognition*, 160–170. 167

GALLEGUILLOS, C. & BELONGIE, S. (2010). Context-based object categorizaton: A critical survey. *Computer Vision and Image Understanding*, **114**, 712–722. 6

GIBSON, J. (1977). The theory of affordances. In *R. Shaw & J. Bransford (Eds.), Perceiving, Acting, and Knowing: Toward an Ecological Psychology.*, 67–82, Hillsdale, NJ: Lawrence Erlbaum. 4

GOMEZ, C. & VALLS, A. (2009). A similarity measure for sequences of categorical data based on the ordering of common elements. *Lecture Notes in Computer Science*, **5285/2009**, 134–145. 25, 27, 135, 136

GORNIAK, P. & ROY, D. (2007). Situated lanugage understanding as filtering perceived affordances. *Cognitive Science*, **30**, 197–231. 5

GRIFFIN, Z. & BOCK, K. (2000). What the eyes say about speaking. *Psychological science*, **11**, 274–279. xii, 17, 91, 92, 94, 108, 113, 116, 122, 123, 124, 125, 150, 203

GUSFIELD, D. (1997). *Algorithms on strings, trees and sequences: computer science and computational biology*. 1st edn. 25, 135

HENDERSON, J. (2007). Regarding scenes. *Current Directions in Psychological Science*, **16**, 219–222. 9, 129, 130, 132, 142, 199

HENDERSON, J., BROCKMOLE, J., CASTELHANO, M. & MACK, M. (2007). Visual saliency does not account for eye-movements during visual search in real-world scenes. *Eye movement research: insights into mind and brain*. 42, 47, 202

HENDERSON, J., MALCOLM, G. & SCHANDL, C. (2009a). Searching in the dark: Cognitive relevance drives attention in real-world scenes. *Psychonomic Bulletin and Review*, **16**, 850–856. 9, 58, 86, 90, 95, 150

HENDERSON, J.M. & HOLLINGWORTH, A. (1999). High-level scene perception. *Annual Review of Psychology*, **50**, 243–271. 3

HENDERSON, J.M., CHANCEAUX, M. & SMITH, T.J. (2009b). The influence of clutter on real-world scene search: Evidence from search efficiency and eye movements. *Journal of Vision*, **9(1)**, 1–8. xii, 96, 100, 107, 122, 124, 132, 163, 172, 182, 203

HENDERSON, M., J. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences*, **7**, 498–504. 3, 129, 159

HUETTIG, F. & ALTMANN, G. (2007). Visual-shape competition during language-mediated attention is based on lexical input and not modulated by contextual appropriateness. *Visual Cognition*, **15**, 985–1018. 10, 41

HUMPHREY, K. & UNDERWOOD, G. (2008). Fixation sequences in imagery and in recognition during the processing of pictures of real-world scenes. *Journal of Eye Movement Research*, **2**, 1–15. 129, 130, 138, 204

HWANG, A., WANG, H. & POMPLUN, M. (2009). Semantic guidance of eye movements during real-world scene inspection. In *N.A. Taatgen & H. van Rijn (Eds.), Proceedings of the 31th Annual Conference of the Cognitive Science Society, Amsterdam*. 22, 129, 136

IORDANESCU, L., GRABOWECKY, M., FRANCONERI, S., THEEUWES, J. & SUZUKI, S. (2010). Characteristic sounds make you look at target objects more quickly. *Attention Perception and Psychophysics*, **72**, 1736–1741. 130

ITTI, L. & KOCH, C. (2000a). Computational modelling of visual attention. *Nature Reviews Neuroscience*, **2**, 194–203. 6

ITTI, L. & KOCH, C. (2000b). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, **40**, 1489–1506. 9, 42, 57, 124, 138, 160, 201, 204

JAEGER, T. (2008). Categorical data analysis: Away from anovas (transformation or not) and towards logit mixed models. *Journal of memory and language*, **59**, 434–446. 29, 100

JUDD, T., EHINGER, K., DURAND, F. & TORRALBA, A. (2009). Vector-based models of semantic composition. In *IEEE 12th International Conference on Proceedings of Computer Vision, 2009*, 2106–2113. 7

KAMIDE, Y., SCHEEPERS, C. & ALTMANN, G. (2003). Integration of syntactic and semantic information in predictive processing cross-linguistic evidence from german and english. *Psycholinguistic Research*, **32**, 37–55. 29

KELLER, F., GUNASEKHARAN, S., MAYO, N. & CORLEY, M. (2009). Timing accuracy of web experiments: A case study using the WebExp software package. *Behavior Research Methods*, **41**, 1–12. 99

KNOEFERLE, P. & CROCKER, M. (2006). The coordinated interplay of scene, utterance and world knowledge. *Cognitive Science*, **30**, 481–529. x, 3, 8, 21, 40, 46, 130, 201

KNOEFERLE, P. & CROCKER, M. (2007). The influence of recent scene events on spoken comprehension: Evidence from eye movements. *Journal of Memory and Language*, **57**, 519–543. 41

LABOV, W. (2004). The boundaries of words and their meanings. In *Aarts, B. & Denison,D. & Keizer,E. & Popova,G. (Eds.), Fuzzy grammar: a reader*, Oxford University Press. 4, 5

LAND, M. (2006). Eye movements and the control of actions in everyday life. *Progress in Retinal and Eye Research*, **25**, 296–324. 4, 9, 13, 129, 204

LANDAUER, T., FOLTZ, P. & LAHAM, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, **25**, 259–284. 135, 136

LEVELT, W., ROELOFS, A. & MEYER, A. (1999). A theory of lexical access in speech production. *Behavioral and brain sciences*, 1–75. 96, 125

MACKAY, D. (2003). *Information theory, inference and learning algorithms.*. 1st edn. 167

MALCOLM, G.L. & HENDERSON, J. (2009). The effects of target template specificity on visual search in real-world scenes. *Journal of Vision*, **9(11)**, 1–13. 9, 17, 122, 129, 159

MALCOLM, G.L. & HENDERSON, J. (2010). Combining top-down processes to guide eye movements during real-world scene search. *Journal of Vision*, **10(2)**, 1–11. 9, 90, 129, 142, 149, 160, 161, 166, 170, 181, 185, 199, 205

MCDONALD, J., BOCK, J. & KELLY, M. (1993). Word and world order: semantic, phonological and metrical determinants of serial position. *Cognitive Psychology*, **25**, 188230. 12, 96, 203

MIRMAN, D., DIXON, J. & MAGNUSON, J. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language*, **59**, 475–494. 38

MITCHELL, J. & LAPATA, M. (2008). Vector-based models of semantic composition. In *Proceedings of ACL-08: HLT*, 236–244, Columbus, Ohio. 137

MITCHELL, J. & LAPATA, M. (2009). Language models based on semantic composition. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 430–439. 136

MOORE, T. (2006). The neurobiology of visual attention: finding sources. *Current Opinion in Neurobiology*, **16**, 159–165. 9

NEIDER, M.B. & ZELINSKY, G. (2006). Scene context guides eye movements during visual search. *Vision Research*, **46**, 614–621. 9, 129, 159

NOTON, D. & STARK, L. (1971). Eye movements and visual perception. *Scientific American*, **224**, 34–43. 9, 23, 129

NOVICK, J., THOMPSON-SCHILL, S. & TRUESWELL, J. (2008). Putting lexical constraints in context into the visual-world paradigm. *Cognition*, 850–903. 29, 48

NUTHMANN, A. & HENDERSON, J. (2010). Object-based attentional selection in scene viewing. *Journal of Vision*, **10**, 1–20. 3, 6, 9, 90, 160, 199, 200

OLIVA, A., TORRALBA, A., CASTELHANO, M. & HENDERSON, J. (2003). Top-down control of visual attention in object detection. *Image processing 2003, Proceedings ICIP 2003*, **1**, 253–256. 91

PARKHURSTA, D., LAWB, K. & NIEBUR, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, **42**, 107–123. 9, 42, 201

PINHEIRO, J. & BATES, D. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer-Verlag. 30

POMPLUN, M., RITTER, H. & VELICHKVOSKY, B. (1996). Disambiguating complex visual information: Towards communication of personal views of a scene. *Perception*, **25**, 931–948. 161, 166

POTTER, M. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory*, **2**, 509–522. 129

PULVERMULLER, F. (2005). Brain mechanisms linking language and action. *Nature Reviews Neuroscience*, **6**, 576–582. 5

QU, S. & CHAI, J. (2008). Incorporating temporal and semantic information with eye gaze for automatic word acquisition in multimodal conversational systems. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP). Honolulu*. xii, 18, 91, 92, 93, 108, 110, 114, 122, 123, 125, 203

QU, S. & CHAI, J. (2010). User language behavior, domain knowledge, and conversation context in automatic word acquisition for situated dialogue. *Journal of Artificial Intelligence Research*, **37**, 247–277. 92, 203

RATNAPARKHI, A. (1996). A maximum entropy model for part-of-speech tagging. In *In Proceedings of the conference on empirical methods in natural language processing*, 133–142. 99

RAYNER, K. (1984). Visual selection in reading, picture perception, and visual search: A tutorial review. In *H. Bouma & D. Bouwhuis (Eds.), Attention and performance (Vol. 10)*, Hillsdale, NJ: Erlbaum. 16

RAYNER, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, **124**, 372–422. 16

RAYNER, K., T.J., S., G.L., M. & J.M., H. (2009). Eye movements and visual encoding during scene perception. *Psychological Science*, **20**, 6–10. 7

REINHART, T. (1983). *Anaphora and semantic interpretation*. Taylor and Francis. 7

RICHARDSON, D., DALE, R. & KIRKHAM, N. (2007). The art of conversation is coordination: common ground and the coupling of eye movements during dialogue. *Psychological science*, **18**, 407–413. 13, 204

RICHTER, T. (2006). What is wrong with anova and multiple regression? analyzing sentence reading times with hierarchical linear models. *Discourse processes*, **41**, 221–250. 29

RIZZOLATTI, G. & ARBIB, M. (1998). Language within our grasp. *Trends in neurosciences*, **21**, 188–194. 5

ROSENHOLTZ, R., MANSFIELD, J. & JIN, Z. (2005). Feature congestion, a measure of display clutter. *SIGCHI*, 761–770. 95, 203

ROSENHOLTZ, R., LI, Y. & NAKANO, L. (2007). Measuring visual clutter. *Journal of Vision*, **7**, 1–22. xiv, 12, 22, 93, 95, 107, 123, 124, 163, 164, 165, 203

ROY, D. (2005). Semiotic schemas: A framework for grounding language action and perception. *Artificial intelligence*, **167**, 170–205. 5

RUSSELL, B., TORRALBA, A., MURPHY, K. & FREEMAN, W. (2008). Labelme: a database and web-based tool for image annotation. *International Journal of Computer Vision*, **77**, 151–173. xiii, 23, 133, 134, 165

SAEED, J. (2008). *Semantics*. Oxford, Wiley-Blackwell. 89

SCHEEPERS, C., KELLER, F. & LAPATA, M. (2008). Evidence for serial coercion: A time course analysis using the visual-world paradigm. *Cognitive Psychology*, **56**, 1–29. 40

SCHMIDT, J. & ZELINSKY, G. (2009). Search guidance is proportional to the categorical specificity of a target cue. *Quarterly Journal of Experimental Psychology*, **62**, 1904–1914. 3, 9, 17, 41, 91, 129, 199, 205

SNEDEKER, J. & TRUESWELL, J.C. (2003). Using prosody to avoid ambiguity: effects of speaker awareness and referential context. *Journal of Memory and Language*, **48**, 103–130. xi, 40, 43, 44, 48, 59, 60, 72, 73, 84, 87, 131

SNEDEKER, J. & YUAN, S. (2008). Effects of prosodic and lexical constraints on parsing in young children (and adults). *Journal of Memory and Language*, 574–608. 8, 40, 59, 60, 61, 63, 66, 72, 73

SPIVEY-KNOWLTON, M., TANENHAUS, M., EBERHARD, K. & SEDIVY, J. (2002). Eye movements and spoken language comprehension: Effects of syntactic context on syntactic ambiguity resolution. *Cognitive Psychology*, 447–481. x, 3, 7, 17, 19, 33, 44, 48, 53, 202

STEEDMAN, M. (2002). Plans, affordances, and combinatory grammar. *Linguistics and Philosophy*, **25**, 723–753. 4

STEVAN, H. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, **42**, 335–346. 5

STURT, P. (2002). Semantic re-interpretation and garden path recovery. *Cognition*, **105**, 477–488. 17

TANENHAUS, M., M.K.AND SPIVEY-KNOWLTON, EBERHARD, K. & SEDIVY, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 632–634. xi, 3, 7, 12, 17, 40, 44, 45, 46, 48, 57, 58, 61, 66, 86, 130, 201, 206

TINKER, M. (1958). Recent studies of eye movements in reading. *Psychological Bulletin*, **55**, 215–231. 17

TOMASELLO, M. (2003). *Constructing a language: a usage-based theory of language acquisition*. Harvard University Press. 6

TORRALBA, A., OLIVA, A., CASTELHANO, M. & HENDERSON, J. (2006). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*, **4**, 766–786. 7, 14, 42, 91, 129, 160, 200

VAN DUREN, L. & SANDERS, A. (1995). Signal processing during and across saccades. *Acta Psychologica*, **89**, 121–147. 16

VO, M. & HENDERSON, J. (2010). The time course of initial scene processing for eye movement guidance in natural scene search. *Journal of Vision*, **10**, 1–13. 50, 129

WALTHER, D. & KOCH, D. (2006). Modeling attention to salient proto-objects. *Neural Networks*, **19**, 1395–1407. 6, 9, 42, 47, 50

WHITTINGHAM, M., STEPHENS, P., BRADBURY, R. & FRECKLETON, R. (2006). Why do we still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecology*, **75**, 1182–1189. 31

WITTGENSTEIN, L. (1921). *Tractatus Logico-Philosophicus*. Project Gutenberg. 4

WOLFE, J. (1998). Visual search. *Attention*, 13–73. 95

YANG, H. & ZELINSKY, G. (2009). Visual search is guided to categorically-defined targets. *Vision Research*, **49**, 2095–2103. 9, 129

YARBUS, A. (1967). *Eye movements and vision*. Plenum: New York. 9, 16, 199

ZELINSKY, G. & SCHMIDT, J. (2009). An effect of referential scene constraint on search implies scene segmentation. *Visual Cognition*, **17**, 1004–1028. 3, 200

ZELINSKY, J., G. & MURPHY, G. (2000). Synchronizing visual and language processing. *Psychological science*, **11**, 125–131. 130

ZELINSKY, W., G. ZHANG, YU, B., CHEN, X. & SAMARAS, D. (2008). The role of top-down and bottom-up processes in guiding eye movements during visual search. In *Y. Weiss & B. Scholkopf, & J. Platt (Eds.), Advances in Neural Information Processing Systems (Vol. 18)*, 1569–1576. 7