Internal Representations of Visual Objects and Their Retrieval during Situated Language Processing

Dissertation
zur Erlangung des akademischen Grades eines
Doktors der Philosophie
der Philosophischen Fakultäten
der Universität des Saarlandes

vorgelegt von
Emilia Ellsiepen
aus Düsseldorf

Saarbrücken, 2015

Dekan:	Prof. Dr. Ralf Bogner
Berichterstatter:	Prof. Dr. Matthew W. Crocker
	Prof. Dr. Bernd Möbius
Tag der letzten Prüfungsleistung	11.7.2014

Abstract

Human language processing often takes place in a surrounding that is inherently connected to the linguistic content because real objects or ongoing events are being referred to. Research in situated language processing has shown that visual attention in a concurrent scene is closely connected to language processing. Furthermore, visual information can disambiguate between linguistic interpretations. This thesis investigates whether these kinds of findings acquired in the Visual World Paradigm generalize to settings where the use of internal memory representations of visual objects is necessary. Moreover, the role of language processing on covert visual attention is examined.

To investigate whether and how internal memory representations may be utilized for situated language processing, we developed a variant of the blank screen paradigm that manipulates the order in which seven visual objects are presented before processing a spoken sentence. Building on the hypothesis that the nature and accessibility of an internal representation partly depends on its serial position in the presentation sequence, this design allows us to examine whether language mediated eye movements on the blank screen can rely on both shallow representations associated with working memory and rich representations containing conceptual information. The results suggest that these different representations become available at different stages of processing with only rich representations being the basis of eye movements during the processing of a restrictive verb, all relevant representations being accessible for anticipatory eye movements after the verb, and shallow but highly active representations being the best candidates for referential eye movements during a noun phrase.

We provide an analysis of these results that combines aspects of two existing accounts of situated language processing to characterize the role and nature of visual attention during sentence processing. From the featural overlap account described in Altmann & Kamide (2007), we adopt the idea that language mediated attention arises automatically as a by-product of linguistic processing and scene processing. In addition to this automatic process, we propose a top-down driven process to guide attention during prediction, similar to the mechanism described in the Coordinated Interplay Account in Knoeferle & Crocker (2006, 2007).

Further evidence for both top-down and automatic effects of language processing on visual attention is provided by two experiments using a variant of the Posner Paradigm. By manipulating the timing and the task given to the participants, we reveal that language can have an automatic influence on covert visual attention in that this influence arises very early and even if the linguistic stimulus is completely irrelevant or even obstructive to a concurrent task. On the other hand, the effect of language processing on the orienting of covert visual attention is enhanced by a task that encourages the use of the linguistic information.

This work highlights the necessity of including non-linguistic cognitive functions in a comprehensive model of situated language processing. We provide the outline of a model that includes a notion of memory and a specific visual attention mechanism that accounts for our experimental findings. In addition, our results support the conjecture that findings from the Visual World Paradigm and especially its Blank Screen version generalize to more naturalistic settings, as the reliance on internal representations is of particular importance in an immersive environment. On the other hand, the influence of a concurrent task on language-mediated eye movements emphasizes the importance of methodological details for the interpretation of existing results.

Zusammenfassung

Natürliche Sprache wird oft in Kontexten produziert und verarbeitet, in denen die umgebende Situation inhärent mit den linguistischen Inhalten verbunden ist. Dies ist z.B. der Fall, wenn das Gesprochene auf reale Gegenstände oder Vorgänge in der Umgebung referenziert. Forschung im Bereich der situativen Sprachverarbeitung hat ergeben, dass Augenbewegungen eines Zuhörers eng mit der Verarbeitung der sprachlichen Äußerung verflochten sind (Cooper, 1974; Tanenhaus, Spivey-Knowlton, Eberhard & Sedivy, 1995): Wird ein Gegenstand, der sich im Sichtfeld des Zuhörers befindet, genannt, wandert der Blick häufig innerhalb weniger hundert Millisekunden zum Gegenstand selbst, oder einer Abbildung des Gegenstandes. In einigen Fällen genügt die bloße Erwartung einer Referenz, zum Beispiel als Komplement eines restriktiven Verbes, um eine Augenbewegung zu einem passenden Objekt auszulösen. Weiterhin wurde gezeigt, dass Informationen aus dem visuellen Umfeld genutzt werden können um zwischen verschiedenen Lesarten zu disambiguieren (Tanenhaus et al., 1995; Knoeferle, Crocker, Scheepers & Pickering, 2005). Viele der Studien im sogenannten "Visual World" Paradigma nutzen stark eingeschränkte visuelle Kontexte, die etwa nur vier bis fünf Objekte umfassen, welche noch dazu gleichzeitig im Sichtfeld des Probanden liegen. In einer realen Umgebung dagegen sind schwerlich alle für die Sprachverarbeitung relevanten Gegenstände ständig vor Augen. Um auch solche Gegenstände, die außerhalb des Sichtfeldes liegen oder durch etwas anderes verdeckt werden auf ähnliche Weise in die Sprachverarbeitung einzubeziehen, müsste daher auf Gedächtnisinhalte zugegriffen werden. Ziel dieser Arbeit war es, die Rolle der nicht-sprachlichen kognitiven Mechanismen innerhalb der situativen Sprachverarbeitung zu untersuchen und zu spezifizieren. Hierbei lag ein besonderer Schwerpunkt auf der Frage, ob sich die Erkenntnisse, die mithilfe des Visual World Paradigmas gewonnen wurden, auf Situationen generalisieren lassen, in denen auf interne Gedächtnisrepräsentationen von visuellen Objekten zugegriffen werden muss. Außerdem wurde untersucht, wie Sprache verdeckte visuelle Aufmerksamkeit, also Aufmerksamkeit jenseits der aktuellen Fixation, beeinflusst.

Eine Variante des Visual World Paradigmas, die bereits erste Ergebnisse über die Nutzung mentaler Repräsentationen geliefert hat, ist das Blank Screen Paradigma, bei dem der visuelle Kontext zunächst allein gezeigt wird und wieder verschwindet, bevor der

sprachliche Stimulus abgespielt wird (Altmann, 2004). Hierbei wurde beobachtet, dass Probanden zum Teil die Stellen, an denen zuvor ein Referent gezeigt wurde, fixieren, sobald der linguistische Stimulus darauf verweist. Dieses Paradigma wurde von uns modifiziert, um genauer zu untersuchen, ob und auf welche Gegenstände für sprach-gesteuerte Augenbewegungen zugegriffen werden kann, wenn der erinnerte visuelle Kontext die angenommene Kapazität des visuellen Kurzzeitgedächtnisses (fünf Objekte) übersteigt. Ausgangspunkt für die Entwicklung des Paradigmas, war die Annahme von seriellen Positionseffekten. Werden einem Probanden eine Reihe von Wörtern hintereinander präsentiert, werden die Wörter am Anfang der Abfolge und am Ende der Abfolge besser erinnert als die in der Mitte. Traditionell werden diese zwei Effekte entweder den verschiedenen Speichern, oder verschiedenen Verarbeitungsstufen zugeordnet: Der Primäreffekt, also das bessere Erinnern von Wörtern am Anfang einer Abfolge, wird mit dem Langzeitgedächtnis assoziiert, das vielschichtige Repräsentationen von Konzepten enthält. Der Rezenzeffekt, das bessere Erinnern von Wörtern am Ende einer Abfolge, wird mit dem Kurzzeitgedächtnis oder Arbeitsgedächtnis assoziiert, in dem die Wörter nur oberflächlich abgebildet werden. Übertragen auf Abbildungen von Gegenständen, wie sie im Blank Screen Paradigma verwandt werden, würden wir bei der Repräsentation im Kurzzeitgedächtnis die Merkmale Form, Position und den Namen des Gegenstandes erwarten, während im Langzeitgedächtnis zusätzlich semantische Merkmale, wie die Zugehörigkeit zu einer Kategorie, oder die Eigenschaft sich für eine bestimmte Tätigkeit zu eignen (Affordanz) enthalten wären. Eine alternative Erklärung unter der Annahme, dass es nur einen einzigen Gedächtnisspeicher gibt, besagt, dass die Gegenstände am Anfang der Abfolge sowohl sensorisch, als auch semantisch verarbeitet werden, wo hingegen für die Gegenstände am Ende der Abfolge aus Mangel an Ressourcen nur noch oberflächliche, sensorische Verarbeitung möglich ist. Beide Erklärungsansätze gehen somit von vielschichtigen Repräsentationen für früh wahrgenommene und oberflächliche Repräsentationen für spät wahrgenommene Gegenstände aus.

In der von uns entwickelten Version des Blank Screen Paradigmas wurden zunächst Bilder von sechs Gegenständen und einer Person nacheinander an verschiedenen Stellen auf einem Computerbildschirm gezeigt. Anschließend hörten die Probanden einen Satz, der ein restriktives Verb beinhaltete, das nur eines der gezeigten Gegenstände als direktes Objekt zuließ: Zum Beispiel wurden Abbildungen von einem Mann, einer Pfeife, einem Messer, einem Mantel, einem Korkenzieher, einem Spazierstock und einem Hut gezeigt, bevor der Satz *Der Mann raucht vermutlich die Pfeife* über Lautsprecher gespielt wurde. Hierbei wurde die Position der Pfeife in der Abfolge der gezeigten Bilder variiert, so dass sie entweder am Anfang, in der Mitte oder am Ende der Abfolge gezeigt wurde. Die Augenbewegungen des Probanden wurden während der Wiedergabe des Satzes mit einer Kamera aufgezeichnet. Wir konnten beobachten, dass im Verlauf des Satzes verschiedene

serielle Positionseffekte auftreten. Während des Verbs ermittelten wir einen Primäreffekt: Gemessen an einem Vergleichsobjekt, traten mehr neue Fixationen der ursprünglichen Position des Zielobjektes auf, wenn dieses am Anfang der Präsentationsabfolge stand. Dagegen trat während des referenziernden Nomens ein Rezenzeffekt auf, die Position des Zielobjektes wurde also häufiger fixiert, wenn es am Ende der Präsentationsfolge auftrat. Während des dazwischenliegenden Adverbs wurde die Zielobjektposition unabhängig von der Präsentationsabfolge häufiger als die Vergleichsobjektposition fixiert. Diese Ergebnisse legen nahe, dass während des Verbs vor allem auf vielschichtige mentale Repräsentationen von visuellen Objekten zugegriffen werden kann, während auch oberflächliche Repräsentationen während des Nomens zur Verfügung stehen. Das Fehlen von seriellen Positionseffekten während des Adverbs deutet an, dass hier alle mentalen Repräsentationen in ähnlichem Umfang die Grundlage sprachgesteuerter Blickbewegungen sein können.

Für das im Visual World Paradigma beobachtete Zusammenspiel von Sprachverarbeitung, visueller Aufmerksamkeit und mentalen Repräsentationen gibt es mehrere Erklärungsansätze. Altmann & Kamide (2007) beschreiben ein Modell, in dem die Überschneidung einzelner Merkmale des linguistischen Stimulus auf der einen, und der mentalen Repräsentation des visuellen Objektes auf der anderen Seite die Wahrscheinlichkeit bestimmt, mit der die (ehemalige) Position des visuellen Objektes fixiert wird. Dahinter steht die Vorstellung, dass die Neuaktivierung der übereinstimmenden Merkmale durch den linguistischen Stimulus dazu führt, dass sich diese Aktivierung in der ursprünglichen Repräsentation des visuellen Objektes ausbreitet. Erreicht nun durch diese Aktivierungsausbreitung das Merkmal der ursprünglichen Lage des visuellen Objektes eine gewisse Aktivierung, schwenkt die visuelle Aufmerksamkeit automatisch zu dieser Position. Je mehr Merkmale übereinstimmen, desto wahrscheinlicher wird eine Augenbewegung. Beispielsweise kann auch die Übereinstimmung des Kategoriemerkmals “Musikinstrument” dazu führen, dass eine Trompete fixiert wird, wenn in einem Satz ein Klavier vorkommt und kein Klavier im visuellen Umfeld vorhanden ist. Ebenso können antizipatorische Augenbewegungen, also Augenbewegungen in Erwartung eines passenden Objektes, damit erklärt werden, dass die semantischen Einschränkungen eines Verbes mit der Affordanz eines visuellen Objektes übereinstimmt. Dabei gehen Altmann & Kamide (2007) davon aus, dass tatsächlich die Erwartung die Augenbewegung bedingt und nicht eine bloße Assoziation des Verbs mit dem Objekt. Eine zentrale Voraussage dieses Models ist es, dass die Verlagerung der visuellen Aufmerksamkeit, die der Augenbewegung vorausgeht, das Ergebnis eines automatischen Prozesses ist. Einen alternativen, wenngleich verwandten Ansatz bieten Knoeferle & Crocker (2006, 2007). In ihrem Modell wird die visuelle Umgebung nach möglichen Referenten für bereits verarbeitete Nominalphrasen oder Ereignisse, sowie nach erwarteten Referenten abgesucht. Die Information, die so aus der Umgebung gewonnen wird, kann zur Disambigua-

tion linguistischer Strukturen und zur Bildung weiterer linguistischer Erwartungen benutzt werden. Während auch in diesem Modell die Ausrichtung der Aufmerksamkeit automatisch ablaufen könnte, ist es mit einem Prozess kompatibel, der durch Willenskraft oder interne Ziele als Nebenprodukt der aufmerksamen linguistischen Verarbeitung gesteuert wird.

Um den von uns beobachteten zeitlichen Verlauf der Augenbewegungen während eines Satzes zu erklären, verknüpfen wir Aspekte beider Modelle. Der Primäreffekt während des Verbs lässt sich gut mit der Überschneidung von Merkmalen beschreiben: Das Verb selbst (z.B. rauchen) ist direkt mit dem Gegenstand (Pfeife) assoziiert und daher Teil der vielschichtigen Repräsentation des Gegenstandes, falls dieser am Anfang der Abfolge gesehen wurde und daher, je nach Lesart, ins Langzeitgedächtnis übergegangen ist oder auf semantischer Ebene verarbeitet wurde. Diese Überschneidung gibt es nur für entsprechend komplexe Repräsentationen, so dass deutlich weniger Augenbewegungen erwartet werden, wenn der Gegenstand intern nur oberflächlich repräsentiert wird. Die Verarbeitung des referenzierenden Nomens dagegen aktiviert das Namensmerkmal, das sowohl in oberflächlichen, als auch in komplexen mentalen Repräsentationen des Gegenstandes vorhanden ist. Obwohl hier also alle zur Verfügung stehenden Repräsentationen Grundlage einer Augenbewegung sein könnten, erwarten wir die meisten für den Fall, dass der Gegenstand spät gesehen wurde: Erstens ist das allgemeine Aktivierungslevel für die spät gesehenen Objekte hoch im Vergleich zu denen, die in der Mitte gesehen wurden. Zweitens ist eben besonders das Namensmerkmal, das als quasi sensorisches, oberflächliches Merkmal zu sehen ist, in besonderem Maße aktiviert, wenn wir die Repräsentation mit der komplexen Repräsentation für früh gesehene Objekte vergleichen.

Die Augenbewegungen während des Adverbs stufen wir als die eigentlich antizipatorischen Augenbewegungen ein. Damit unterscheiden wir uns von bisherigen Interpretationen, nach denen bereits während des Verbs die Antizipation eines direkten Objektes zu beobachten ist. Anlass für diesen Unterschied bietet uns das Fehlen eines seriellen Positionseffektes. Ursprünglich bestanden zwei konkurrierende Hypothesen bezüglich der seriellen Positionseffekte: Wenn die Antizipation dem Wesen nach konzeptuell wäre, würden wir unter Berücksichtigung des oben beschriebenen merkmalbasierten Erklärungsansatzes hier ebenfalls einen Primäreffekt erwarten, da die konzeptuellen Merkmale wie die Eigenschaft, sich für eine bestimmte Tätigkeit zu eignen, nur Teil der vielschichtigen Repräsentation ist, die für früh gesehene Gegenstände gebildet wird. Alternativ könnte Antizipation auf der lexikalen Ebene wirken, das heißt bestimmte Wörter würden vorhergesagt. In diesem Fall würden wir hier die gleichen Muster erwarten, wie während des darauffolgenden Nomens, also einen Rezenzeffekt. Da beide Effekte nicht beobachtet wurden, sondern nur der allgemeine Effekt, dass das Zielobjekt häufiger als das Vergleichsobjekt fixiert

wurde, gehen wir davon aus, dass hier ein anderer Prozess zugrunde liegt. Ähnlich dem in Knoeferle & Crocker (2007) beschriebenen Mechanismus, sehen wir die Erklärung in einer gesteuerten Verlagerung der visuellen Aufmerksamkeit hin zu dem Objekt, das am ehesten den linguistischen Erwartungen entspricht. Da dieser Prozess die vorherige Bildung linguistischer Erwartungen voraussetzt, läuft der Prozess möglicherweise nur ab, wenn der Zuhörer aufmerksam zuhört, also das Ziel hat, den Satz zu verstehen.

Einen weiteren, unabhängigen Beleg für die Koexistenz von automatischer und intern gesteuerter Ausrichtung der visuellen Aufmerksamkeit durch sprachliche Stimuli liefern die in Kapitel 3 dargestellten Experimente. Angelehnt sind unsere Experimente an das Posner Paradigma (Posner, 1980), bei dem die Ausrichtung der verdeckten visuellen Aufmerksamkeit ermittelt werden soll. Hierfür wird an einer Stelle, auf die zuvor ein meist visueller Reiz verwiesen hat, ein neuer Reiz gezeigt, auf den der Proband reagieren muss, wobei der Proband die Augen nicht von einem zentralen Fixationspunkt bewegen darf. In unserer Variante wurden zuerst zwei fotografische Abbildungen einfacher Gegenstände links und rechts des Fixationspunktes gezeigt. Diese verschwanden wieder und ein einzelnes einsilbiges Wort wurde über Lautsprecher vorgespielt. Kurz darauf mussten die Probanden mit einem Tastendruck auf einen Punkt reagieren, der entweder an der gleichen Stelle, wie das genannte Objekt oder an einer anderen Stelle erschien. Um die Automatizität des Prozesses zu untersuchen, variierten wir die Aussagekraft des Hinweisreizes aus Wort und Bild: für die erste Gruppe war die bezeichnete Stelle nur in 50 % aller Fälle, also mit Zufallswahrscheinlichkeit, die Stelle, an der der Punkt erschien. Hiervon unterrichteten wir auch die Probanden und wiesen darauf hin, dass der sprachliche Reiz ignoriert werden könne. Für die zweiten Gruppe, war der Hinweisreiz in 75% aller Fälle hilfreich, hier erwarteten wir also, dass die Probanden ihre Aufmerksamkeit bewusst der bezeichneten Stelle zuwenden würden. Für eine dritte Gruppe wies der Hinweisreiz nur in 25 % der Fälle auf die Stelle, an der danach der Punkt erschien, wir wiesen die Probanden also darauf hin, dass sie am besten ihre Aufmerksamkeit auf die dem genannten Gegenstand gegenüberliegende Seite richten sollten, um schnell reagieren zu können. Ob die visuelle Aufmerksamkeit tatsächlich auf der vorherigen Position des genannten Bildes gerichtet war, ermittelten wir indem wir die Reaktionszeiten verglichen, mit der die Probanden auf den Punkt reagierten. Bereits für die erste Gruppe, bei der das Wort ignoriert werden durfte, ermittelten wir signifikant kürzere Reaktionszeiten für kongruente Proben, das heißt für den Fall, dass der Hinweisreiz auf die später getestete Stelle wies, unabhängig wie schnell nach dem Beginn des gesprochenen Wortes (200, 500 oder 800 ms) der Punkt erschien. Hieraus leiten wir ab, dass ein Hinweisreiz aus Wort und Bild automatisch die Aufmerksamkeit auf die betreffende Stelle lenken kann. Für die zweite Gruppe, für die der Hinweisreiz in der Mehrzahl der Fälle informativ war, beobachteten wir ebenfalls einen erleichternden Effekt

für kongruente Proben. Darüber hinaus war der Unterschied zwischen kongruenten und inkongruenten Proben hier größer als in der ersten Gruppe. Wir konnten also beobachten, dass willentliche Einflussnahme die Wirkung des sprachlichen Stimulus auf die Ausrichtung der visuellen Aufmerksamkeit steigern kann. Im letzten Fall, in dem der Hinweisreiz in der Mehrzahl der Fälle irreführend war, beobachteten wir zwei unterschiedliche Effekte, je nachdem ob der Teststimulus 300 oder 1200 ms nach dem Wortanfang erschien. Nach 1200 ms ermittelten wir kürzere Reaktionszeiten für inkongruente Proben als für neutrale Proben, in denen das gesprochene Wort auf keines der gezeigten Gegenstände referierte. Da wir ja explizit eine Orientierung auf die gegenüberliegende Seite empfohlen hatten, entsprach dies unseren Erwartungen und zeigt, dass die Probanden in der Lage waren sich willentlich vom Hinweisreiz weg zu orientieren. Im Gegensatz dazu waren nach 300 ms kürzere Reaktionszeiten für kongruente Proben zu beobachten. Die aufmerksame Verarbeitung des Wortes führte also auch zu einer Verlagerung der Aufmerksamkeit zu der vom Hinweisreiz bezeichneten Stelle, obwohl das willentliche Ziel war, die Aufmerksamkeit gerade auf die gegenüberliegende Seite zu richten. Zusammengenommen zeigen diese Experimente, dass es einen automatischen Einfluss von Sprache auf die visuelle Aufmerksamkeit gibt, die willentlich verstärkt werden kann.

Unsere Ergebnisse verdeutlichen die Notwendigkeit nicht-linguistische kognitive Prozesse in ein aussagekräftiges Modell der situativen Sprachverarbeitung mit einzubeziehen, zu dem wir einen ersten Entwurf vorstellen. Für den Umgang mit dem Visual World Paradigma lassen sich sowohl Vorbehalte als auch Bestätigung ableiten. So ergibt sich für uns die Konsequenz bei Entwicklung und Interpretation betreffender Experimente die unterschiedlichen Mechanismen, die unserer Analyse nach antizipatorischen sowie referentiellen und assoziativen Blickbewegungen zugrunde liegen, zu beachten, um eine Vermischung oder Verwechslung möglichst zu vermeiden. Dass wir neben dem automatischen Einfluss von Sprache auf die visuelle Aufmerksamkeit auch die Wirkung von internen Zielen und willentlicher Einflussnahme beobachten konnten, führt uns zu der Annahme, dass die Aufgabe, die den Probanden gestellt wird, die Ergebnisse beeinflussen kann. Auf der anderen Seite zeigen unsere Experimente, dass sich die Resultate des Visual World Paradigmas durchaus auf komplexere, Sprecher oder Hörer umschließende Situationen generalisieren lassen, bei dem auf interene Gedächtnisrepräsentationen zurückgegriffen werden muss.

Acknowledgments

This thesis is the result of a long process. Although I have experienced ups and downs all along the way, it falls apart into two phases: an enthusiastic first part, where every step seemed to lead to new exciting ideas, and the write-up. Different people have accompanied this process at different times and contributed in different ways. I am thankful for all these small or more important contributions and I hope that this thesis is able to do them justice.

First of all, I would like to thank my supervisor, Matthew Crocker. He supplied me with guidance and support while at the same time leaving enough space for me to develop and pursue my own ideas. In addition to the supervision itself, I am also thankful for the chance to work in a well-equipped and well organized environment. I also would like to thank the members (and former members) of the psycholinguistics group of Saarbrücken for creating a good atmosphere and especially Berry Claus, Maria Staudte and Juliane Steinberg for interesting discussions, feedback, and encouragement facing problematic data and other common obstacles in academia.

I am very thankful to John Henderson and Fernanda Ferreira who welcomed me as a visiting researcher in Edinburgh. This period formed me importantly as a researcher.

I am grateful to the DFG for providing financial support and to all members of the IRTG for providing a friendly and stimulating environment. Special thanks go to Judith Köhne, Antske Fokkens, Mark Buckley, Bart Cramer, Arnab Goshal, and Rebecca Dridan as well as to Silke Theison, Jessica Gauer, and Cathrin Bautz who enriched my non-academic time in Saarbrücken greatly.

As to the second phase mentioned above, it was the moral support that kept me going, provided by my family and friends. In this context, I would like to express my deep gratitude to Judith Köhne and Barbara Hemforth for their invaluable support and feedback, without which I probably would not have finished.

Finally, I want to thank Robert Künnemann for his support, patience and love throughout the last four years.

Contents

List of Figures

Chapter 1.

Introduction

In everyday life, we process language in a variety of contexts. In some of them, the *situation* in which language is processed is inherently related to the linguistic content. This includes situations in which interlocutors are conversing about objects, people, or events in their immediate surrounding or are collaborating on a task. Consider a situation in which two people prepare a meal together. Their communication while instructing each other and coordinating the individual steps will regularly include references to objects in their environment, such as individual ingredients or kitchen equipment. In order to understand what interlocutor A is saying and to act accordingly, interlocutor B has to identify the object(s) mentioned by A. If, for instance, A asks B "*Could you hand me the carrot, please?*", B will look for the carrot, fixate it, and form a referential connection between the word *carrot* and the object carrot in order to plan her next step. If A asks B instead "*Could you peel the carrot?*" B might even look for the carrot before A has mentioned it since the verb *peel* already indicates what kind of object A will mention next (i.e., something that can be peeled). The study of *situated language processing* investigates the connection of language processing on the one hand and the processing of the surrounding situation on the other hand. As illustrated by the example, language can influence the perception of the situation, in that attention is drawn quickly to objects which are being mentioned or become relevant for the linguistic content. Information acquired from the scene, on the other hand, can influence linguistic processing in restricting linguistic predictions or disambiguating between possible interpretations. This interplay of different cognitive processes will often prove to be beneficial for the goal of efficient communication.

In the situation described above, the two interlocutors are required to process their visual surroundings and direct their attention to relevant objects to facilitate communication. The scenario becomes more complex, if objects are mentioned that are not currently in the field of view. The carrot in our example might be situated on the shelf behind B or stored away in the fridge. In this case, one possibility for B would be to search the environment

for the carrot. Alternatively, she might remember where she saw it before and be able to locate it without a search. In order to establish reference between the word and the object out of view in a similar fashion as to an object right in front of her, however, the linguistic stimulus would first have to trigger memory access and subsequently a shift of attention. Whether and, more interestingly, how exactly this happens is an open question.

The rapid integration of language and scene information is often important in order to achieve a specific goal, as preparing a meal together in the example given above. Going one step further, it is possible and has been argued that looking at objects relevant for the current linguistic input is independent of a concurrent task. Consider a third person C sitting next to the two cooks at the table enjoying a cup of tea and passively following the conversation of A and B. Most likely, C forming referential links between objects and words will not contribute to the success of the conversation or the outcome of the cooking activity. Nevertheless, C might find himself looking at the mentioned objects from time to time. Here, the question remains whether this looking behavior emanates from C's intention to pay attention to the cooking activity, or from an automatic process linking the linguistic stimulus to a real world object.

This dissertation aims to advance the knowledge about situated language processing in a broad manner. One objective is to generalize existing findings step by step to more complex and hence more natural situations. This goal requires a systematic investigation of the influence of non-linguistic processes that play a role in scene processing. An important aspect that is not considered by most existing studies is that scene processing in natural situations is dependent on memory access and retrieval as illustrated in our example situation above. For this reason, this work explores the activation of memory representations by language, and the connection between language, memory, and the direction of visual attention. Another important factor that has not been systematically addressed in this context is the influence of the presence or absence of a concurrent task that requires a fast integration of language and situation, as the cooking activity above. In fact, language processing often occurs without such a task and might even compete or interfere with other goals or activities. The influence of different tasks that either benefit or suffer from language-mediated attention is therefore put to test.

The second aim which complements this rather pragmatic approach is to specify, evaluate, and extend theoretical assumptions and conceptions prominent in accounts of situated language processing. In order to achieve this, existing theories are first examined with regard to the role of non-linguistic components and their interaction, before addressing issues that remain unclear experimentally. One point in question here is the potential automaticity of integrating linguistic material with scene information and the influence of

internal goals in this context, which has important implications on language-mediated gaze in the absence of a task. A further issue concerns the use of internal memory representations of visual objects for linguistic processing. Although it has been established that such representations can be used at least in highly restricted contexts, little is known about their nature. One prominent question we will pursue here is what kind of memory representations of visual objects are accessibly for language-mediated attention and ultimatly to inform linguistic processing. From the extensive literature on memory we single out at least two candidates: shallow and short-lived representations traditionally associated with short-term memory, and rich, conceptual representations that are, in some models, part of long-term memory.

The remainder of this dissertation is structured as follows. Chapter 2 first lays out the methodological and theoretical basis for the study of situated language processing as evidenced by language-mediated eye movements and points out particular aspects that are in need of further specification and validation. Next, it gives an overview of the non-linguistic cognitive mechanisms involved in scene processing focusing on the aspects proving relevant for language processing. Chapter 3 reports two experiments testing for automaticity and task-relatedness of language-mediated attention. Chapter 4 describes three experiments addressing the accessibility of memory representations for language-mediated eye movements, the nature of these representations, and the use of different memory representations for different linguistic processes. This last point is validated with another experiment reported in Chapter 5. Chapter 6 provides a general discussion of the results and their implications for theories of situated language processing.

Chapter 2.

Language-mediated Eye Movements

The study of situated language processing investigates linguistic processing in context, that is the situation in which it takes place. A growing body of research in this field suggests that language processing is tightly interlinked with other cognitive processes such as the orienting of visual attention in a concurrent scene (Cooper, 1974; Tanenhaus et al., 1995; Griffin & Bock, 2000). Moreover, the visual information available within the situation can rapidly influence linguistic processing as evidenced by the early (visual) resolution of syntactic ambiguities (Tanenhaus et al., 1995; Knoeferle et al., 2005). The apparent link between overt visual attention (i.e., eye movements) and language processing has been exploited in the *Visual World Paradigm* to study psycholinguistic phenomena. While eye movements in this paradigm have often been used as a mere index of underlying linguistic processes, the tight connection between the two supports a generally interactive cognitive architecture, where one phenomenon can only be fully understood if the other is taken into account. Consequently, the mechanism controlling visual attention as well as other cognitive components involved in scene processing should be included in a cognitive model of situated language processing. Existing accounts (e.g. Tanenhaus, Magnuson, Dahan & Chambers, 2000; Knoeferle & Crocker, 2007; Altmann & Kamide, 2007) mostly focus on the influence of language on the probability to direct visual attention to a specific target leaving other components, in particular the use of memory representations of scene objects and the assignment of *visual indices*, that is, internal pointers to scene objects which are followed when these scene objects are refixated, largely unspecified. This chapter will firstly give an overview of experimental research and theoretical accounts in the context of the Visual World Paradigm highlighting those aspects that deserve further investigation. Secondly, the non-linguistic cognitive capacities relevant for situated language processing are introduced discussing their potential function and identifying specific problems to be addressed experimentally.

2.1. Eye Movements in Psycholinguistic Research: The Visual World Paradigm

The Visual World Paradigm has proven to be a powerful tool to study online language processing in a closely time-locked manner. In a typical visual world experiment, the participant is presented with a visual context and a linguistic stimulus simultaneously. The participant's eye movements in response to the linguistic stimulus are then recorded with an eye tracker. The participant's gaze patterns in the co-present scene are used as an index of the current linguistic processing stage: if, for example, the participant fixates a candy shortly after hearing *the candy*, this can be interpreted as evidence that the participant processed the word *candy* and established reference to the visual object candy. If, in another case, the participant fixates a cake among a number of non-edible objects after hearing the fragment *The boy will eat*, this could indicate that the participant has processed *eat*, evaluated the selectional restrictions of the verb and identified a referent for the yet missing complement of the verb. These two examples, taken from Tanenhaus et al. (1995) and Altmann & Kamide (1999), demonstrate two kinds of eye movements: *Referential eye movements* that occur in response to a referring noun phrase and *anticipatory eye movements* that convey linguistic expectations in response to an unfinished utterance. In this section, relevant findings of word-level and sentence-level studies are presented as well as methodological details and possible linking hypotheses for this paradigm.

2.1.1. Referential Eye Movements

The first study that connected spoken language to eye movements in a co-present visual context (Cooper, 1974) found a temporal coordination between the utterance of referring noun phrases and the fixations of visual referents. In this study, participants were listening to short passages of prose while inspecting an array of black-and-white line drawings. Participants tended to fixate objects when they were mentioned (e.g., a lion on hearing the word *lion*) and when a semantically related concept was mentioned (e.g., a lion on hearing the word *Africa*). While Cooper noted that listener's fixations happened often while the word was pronounced, a more detailed investigation of the time course of referential eye movements was conducted by Allopenna, Magnuson & Tanenhaus (1998). They presented participants with visual arrays of objects such as the one in Figure 2.1, each containing a *target* object, (e.g., a beaker), a cohort competitor (beetle), a rhyme competitor (speaker), and an unrelated competitor (carriage). The two competitor types differed in the occurrence of *phonological overlap*: The cohort competitor started with the same phoneme sequence as the target with the second syllable disambiguating between

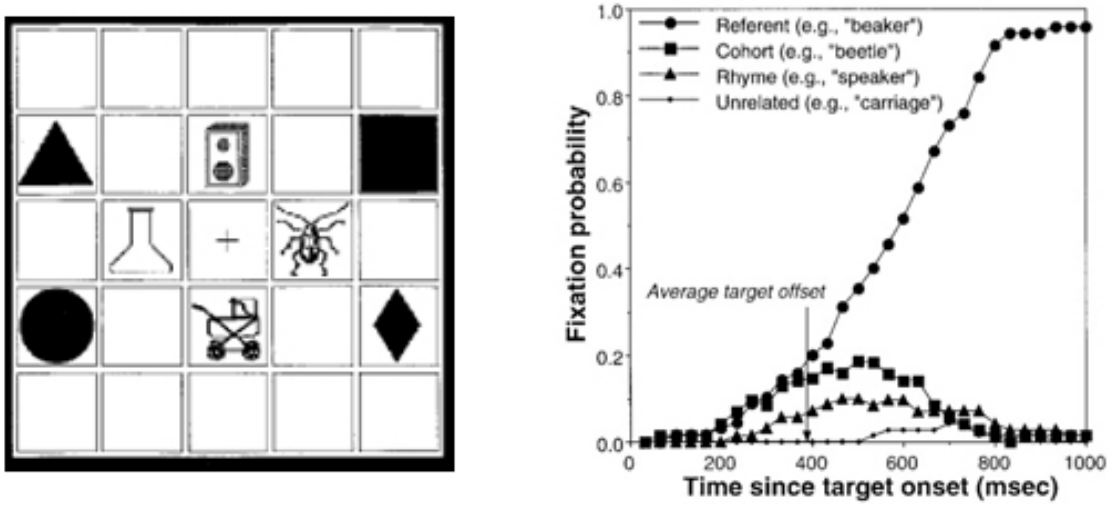


Figure 2.1.: Example display and graph of fixation proportions taken from Allopenna et al. (1998)

the two. The rhyme competitor, on the other hand, differed only by the first segment. They analyzed the fixation proportions to these objects during the unfolding of the word *beaker* within the instruction "*Pick up the beaker; now put it below the diamond*". The time graph in Figure 2.1 shows that at the onset of *beaker*, when information of which object to click on was not yet revealed, participants were equally likely to fixate either object. Starting already 180 ms later, however, fixations on the beaker increased showing a rapid influence of the spoken word on eye movements even before the word was completed. Furthermore, the beetle was fixated equally often as the beaker in this early stage. This implies that those visual objects whose name is consistent with the portion of the speech stream already processed function as candidates for the establishing of reference. Only towards the end of processing the spoken word, the beaker was fixated more often than all other objects. Towards the end of the word *beaker* there were also slightly more fixations on the speaker than the unrelated competitor, but less than to the beaker. The authors attribute this to the phonological overlap of *beaker* and *speaker* during the rhyme. Unlike the beetle at the beginning of the word, however, the speaker is not a plausible candidate at this point, since it is incompatible with the beginning of the word. This suggests that an overlap of specific features between the spoken word and a concurrent object might drive eye movements irrespective of whether or not it is a plausible referent of the noun.

In addition to phonological overlap of the names, visual similarities between a hypothetical referent of a noun and another visual object can drive eye movements (Huettig & Altmann, 2005; Yee & Sedivy, 2006; Huettig & Altmann, 2007; Dahan & Tanenhaus, 2005). Huettig & Altmann (2005) tested whether semantic competitors also attract eye

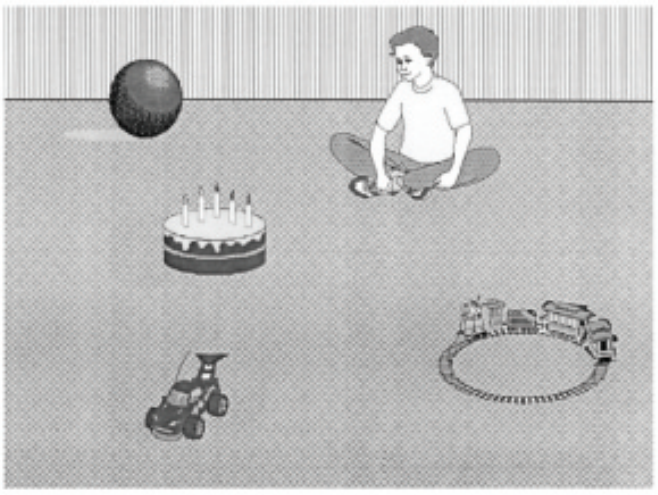


Figure 2.2.: Example picture taken from Altmann & Kamide (1999)

movements. In their experiment, the display contained e.g. a piano, a trumpet, and two unrelated objects. Upon hearing a sentence mentioning the piano, participants were more likely to fixate the piano than any other object. The trumpet, however, was more likely to be fixated than the unrelated competitors. Also, in a display which did not contain a piano, but only the trumpet and three unrelated competitors, the trumpet was most likely to be fixated. Further, Huettig & Altmann (2007) showed that competition of noun-driven eye movements is not restricted to linguistic or semantic similarity. In their experiment, participants listened to a sentence e.g. containing the word snake. The display again contained four objects, all of them linguistically unrelated to snake, but one of them, a cable, had a similar shape as a hypothetical referent of the word snake. The authors found more fixations on the cable than the other objects on display. In summary, the processing of a noun rapidly influences gaze behavior in that referents and objects related phonologically, semantically or with respect to shape are inspected more often than other objects.

2.1.2. Anticipatory Eye Movements

Not only do people quickly react to the mentioning of a specific entity, Altmann & Kamide (1999) demonstrate that even *before* the referring noun is uttered, eye movements can be driven by the expectation of what is going to be mentioned. In their study, participants were presented a clip-art scene containing, e.g., a boy, a ball, a toy car, a toy train and a cake (see Figure 2.2 for this example) and listened to one of the sentences "The boy will move the cake" or "The boy will eat the cake". In the first case, participants fixated the

cake more often than the other objects after the onset of *cake*. In the second case, however, they often fixated the cake already when hearing the verb *eat*. Since the cake was the only edible object on the screen, the authors argued that the selectional restrictions of the verb together with the expectation of a direct object and the visually present cake allowed participants to correctly anticipate the cake to be mentioned next.

Since semantic overlap can also evoke eye movements as described in the last section, an alternative explanation for this experiment could be the association or semantical overlap between *eat* and *cake*. Stronger evidence that eye movements can reflect linguistic anticipation comes from a study by Kamide, Scheepers & Altmann (2003). Exploiting the flexible word order and distinct case marking for objects and subjects in German, this study suggests that world knowledge, case marking, and selectional restrictions are used compositionally to enable anticipatory eye movements. The authors describe an experiment similar to Altmann & Kamide (1999), but with two plausible agent-patient pairs with respect to the verb. A display contained, for example, a hare, a fox, and a cabbage for the verb *eat*. World knowledge informs us that either the hare could eat the cabbage, or the fox could eat the hare. Participants were then presented a spoken sentence that either started with "The hare-nom eats shortly" or "the hare-acc eats shortly" where the nominative case marking on *hare* indicates that the hare is the agent of the eating action, whereas the accusative case marking designates the hare as the patient. In the first case, the continuation contained the cabbage as the patient, and participants were more likely to fixate on the cabbage during both, verb and adverb. In the second case, the fox as the agent of the action followed the adverb. Although this is not the canonical word order in German main clauses, the accusative case marking influenced gaze behavior: While the cabbage was still fixated more often during the verb, both objects were looked at equally often during the post-verbal adverb. Relative to the nominative condition, the fox, being the appropriate scene object was thus fixated more often, indicating a combined use of case marking, selectional restrictions, and world knowledge to anticipate the next linguistic entity.

2.1.3. Scene Information Influencing Syntactic Processing and Anticipation

Drawing on the apparent link between linguistic processing and the direction of eye movements as discussed above, several studies show that the information provided by the scene can influence syntactic processing. Tanenhaus et al. (1995) instructed participants to manipulate real-world objects with sentences like "Put the apple on the towel in the box", where *on the towel* was temporarily ambiguous between modifying *the apple* and specifying the goal for *put*, where in general the goal interpretation is preferred. The display for this

study contained a box, an apple lying on a towel, an empty towel, and either another apple which was lying on a napkin (Two-referent context), or an unrelated object (One-referent context). In the one-referent context, participant initially interpreted *on the towel* as the goal, evidenced by eye movements towards the empty towel. In the two-referent context, however, *on the towel* was correctly interpreted as a modification of *apple*, which enabled them to decide which of the two apples to pick, as reflected by eye movements between the two apples and fewer eye movements towards the empty towel.

In addition to the referential context, information provided by scene events were found to influence syntactic disambiguation and anticipation processes. Knoeferle et al. (2005) presented scenes with three characters participating in two different agent-action-patient events with the middle character being at the same time the agent of one event and the patient of the other, combined with German sentences that were initially ambiguous between SVO and OVS word order. This was achieved by choosing feminine noun phrases whose case marking for nominative and accusative is identical and which referred to the middle character which could thus be the subject of a sentence describing the first event, or the object of a sentence describing the second one. The verb, however, identified the event relevant to the sentence and thereby enabled participants to choose the correct structure early on. Indeed, participants fixated the correct referent of the second noun phrase before hearing it, during the post-verbal adverb.

2.1.4. The Blank Screen Paradigm

In a variant of the Visual World Paradigm, language-mediated eye movements are studied in the case where the scene itself is not co-present with the language. Inspired by the finding of Richardson & Spivey (2000) and Spivey & Geng (2001) that eye movements may be directed systematically to empty regions of a screen if information connected to formerly present objects in these regions has to be retrieved (see section 2.2.2 for more discussion), Altmann (2004) conducted an anticipation study similar to Altmann & Kamide (1999) but exploiting a blank screen. In the experiment, participants first inspected a scene containing two characters, a target object and another object, before the scene was removed again leaving a blank screen. One second later, a sentence containing a restrictive verb selecting for the target object was played back to them. Interestingly, the pattern of eye movements was comparable to those on a co-present scene: the region formerly occupied by the target object was inspected more frequently than the region formerly occupied by the other object. Although the overall proportion of trials with relevant

fixations was considerably smaller than in the original study,[1] this experiment suggests that language's grip on attention is not only effective when referents are co-present, but also when they were experienced earlier. Following Richardson & Spivey (2000), Altmann (2004) attributes the eye movements on a blank screen to the use of *spatial indices*, and describes two possible explanations for why the eye might follow these indices. The first one rests on the idea of the world as an external memory (O'Regan, 1992): Instead of storing visually rich information internally, only an index, or pointer, is stored and any information needed about this object is retrieved directly by looking back at it. In this view, the visual system is blind to the fact that the scene has already disappeared. An alternative explanation favored by Altmann (2004) is the connection of a spatial index to an internal memory trace of the object, where the eyes follow the pointer, if information of this trace is retrieved. Irrespective of what explanation is adopted, the findings from the blank screen paradigm stress the connection between language, visual attention, and scene *memory*. In order to gain a deeper understanding of situated language processing, it is therefore necessary to understand the influence and interaction of memory and visual attention.

2.1.5. Linking Eye Movements to Linguistic Processes

The above presented research clearly shows that eye movements on a co-present scene as well as on a blank screen formerly occupied by a scene are closely time-locked to the processing of linguistic material. In order to draw inferences on linguistic processes as well as to understand the mutual influence of processing scene and linguistic stimulus, it is important to specify *how* they are linked. An early account (Allopenna et al., 1998) linked the proportion of eye movements towards an object to the lexical activation of its name – dependent on the visual context (objects on display) and the amount of processed linguistic input. While this linking hypothesis, which they implement using the computational model TRACE (McClelland, Elman & Diego, 1986), is able to account for the data of their experiment, it cannot easily be transferred to experiments using whole sentences, as well as to experiments that found eye movements towards objects that were not named, but semantically or visually related to the spoken word (Yee & Sedivy, 2006; Huettig & Altmann, 2007). In this section, two accounts will be presented, one focusing on the mental activation necessary to trigger an eye movement and the other on the role of the reference establishing process on subsequent language understanding.

[1] In the original study, 54% of trials contained anticipatory eye movements, on the blank screen, only in 4% of the trials an anticipatory eye movement was launched towards the exact location formerly occupied by the target object, and in 20% of trials towards the quadrant formerly containing the target.

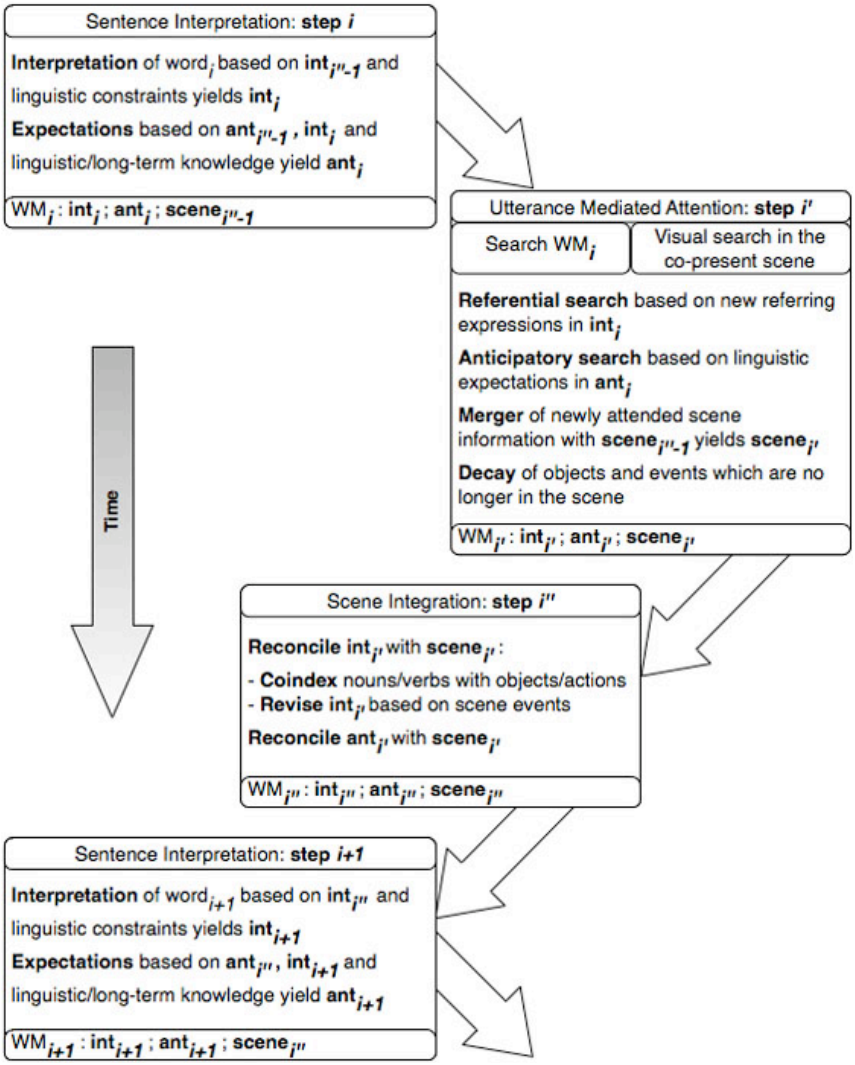


Figure 2.3.: A processing step illustrating the functioning of the CIA , taken from Knoeferle & Crocker (2007)

The Coordinated Interplay Account

To model the interplay of scene, utterance and world knowledge in situated language processing, Knoeferle & Crocker (2006, 2007) propose their coordinated interplay account (CIA). This model is especially well-suited to account for the influence of visual information from the scene on subsequent prediction and processing steps. They outline a step-wise algorithm as illustrated in Figure 2.3: At a given point during the processing of an utterance the listener holds an interpretation of the utterance fragment already processed, expectations of up-coming words, and an internal representation of the scene in working memory. Once a new word is encountered, the sentence interpretation and the linguistic expectations are

updated. Next, the listener searches the co-present scene and the representation of the scene in working memory for referents of the new interpretation and anticipated entities, derived from linguistic expectations. Based on the newly identified entities and attended proximal scene information, the scene representation in working memory is updated, whereas entities that are no longer present in the scene experience decay. The newly formed representation of the scene is now integrated with the interpretation: reference is established by co-indexation of verbs with events and noun phrases with objects, and the interpretation and expectations are updated according to the information provided by the attended portion of the scene.

The strength of the CIA is providing a motivation for language-driven eye movements as part of situated language understanding. In order to be able to use the information provided by the scene, visual attention is directed towards relevant objects or events, as soon as the phonological input and/or linguistic expectations allow for this. Also, within the CIA, working memory is considered in addition to linguistic input and visual attention. One feature that remains somewhat schematic is the kind of internal representations used by the described mechanism, in particular whether linguistic expectations are necessarily lexical, or whether they could also comprise a semantic class and whether representations in working memory are similar to those of co-present objects.

Mayberry, Crocker & Knoeferle (2009) provide a connectionist model of an earlier version of their account (not including working memory representations) and their experimental data on co-present scenes. In their CIAnet – a modified simple recurrent network (Elman, 1990) – scene events and linguistic input are input to the network separately to produce an interpretation, consisting of a verb and two noun phrases with their respective thematic roles. Temporary interpretations of yet incomplete input sentences are understood to denote anticipation. Attention is instantiated as a gate, which selects one of two scene events, based on its consistency with the current interpretation. Their model is able to predict the correct noun phrase with its thematic role during the post-verbal adverb, when attention had been shifted to the relevant scene event.

While CIAnet spells out some of the details, it can not model language-mediated attention in a more general sense, as it is deliberately limited to produce interpretations in the case where event information or stereotypical information can be used to make predictions. Lexical items were realized as random feature vectors, so that anticipation could also be lexical only, and not affordance-driven. This makes it difficult to account for the findings from Altmann & Kamide (1999), where the semantic class of edible objects was predicted by the verb. Also, their attention mechanism is not a sufficient model of eye movements in the visual world for two reasons. Firstly, it can only be directed towards

events, and not individual scene objects. To be able to model anticipatory and referential eye movements, this mechanism would have to be refined. Secondly, as the scene entities were also realized as random feature vectors, no bottom-up effects on attention can be captured.

Activation of Features Result in Eye Movements

Altmann & Kamide (2007) account for language-mediated eye movements by supposing an automatic reactivation of internal representations of scene objects driven by featural overlap, which we will term here the featural overlap account (FOA). In their view, both, linguistic expectations, and referring expressions activate an internal, multi-dimensional representation: for the word *piano* this representation entails presumably the phonological form of its name, the form of a piano-object, the sound it can produce, the affordance of being used for playing music, associations, and so on. Crucially, this representation can interact with further internal representations by boosting their activation if there exists conceptual overlap. If, for instance, a trumpet was previously encountered in the scene, an internal trumpet representation has been formed that overlaps in semantic features with the one of piano: both are musical instrument, both can be used to play music, both might be associated with concerts, conductors etc. When these features become activated as part of the representation of *piano*, they also become reactivated in the trumpet representation. In consequence, the renewed activation of trumpet features spreads activation to the other features of the trumpet representation, which results in an overall higher activation of the trumpet representation. Other representations built up on previously inspected scene objects might also experience a boost in activation depending on their featural overlap with the piano representation. If, e.g., the scene contained a piano and a hammer in addition to the trumpet, we would expect the representation of the visual piano object to receive the highest boost in activation, followed by the trumpet, and then possibly the hammer, if any features are shared between hammer and piano.

To understand how this reactivation can induce a saccadic eye movement towards an object (or its former location) it is first necessary to assume that a shift of attention takes place: attention could be allocated to the object which currently enjoys the highest activation. Alternatively, the activation could in itself constitute the shift in attention. Secondly, the attentional system has to orient towards the *location* of the attended object. Altmann & Kamide (2007) argue that the location information is to be found as part of the internal representation, that consists not only of features in semantic memory common to all objects of this type, but also of an *episodic record* of experiencing this object in the current context. Within this episodic record, the spatial location of the object is

represented as well as other situation-specific properties. The attentional system can thus orient towards the current or former location of the object which increases the probability of a saccade towards this location.

One important aspect of this account is that the change in the attentional system occurs automatically. By building up a representation of the linguistic unit, the representation of the visual object undergoes reactivation independent of a voluntary search for referents in the visual scene. At the same time, the conditions under which a saccadic eye movement is launched are not further specified. As it is only the probability that rises, the eye movement itself can not be automatic. This is in line with the fact that in the experimental findings within the Visual World Paradigm it is only a fraction of trials in which participants indeed fixate the predicted object. The predictions derived from this account are thus on the one hand the relative probability with which certain objects or locations are looked at with regard to other regions. On the other hand, this account predicts an automatic shift of attention towards objects or locations associated with the current linguistic input or interpretation.

CIA and FOA

The two described models both integrate non-linguistic components of cognition to account for situated language processing, in particular the concepts of visual attention, spatial indexing, and memory. While CIA and FOA are not mutually exclusive, they do specify and stress different sources of influence from general cognitive abilities. Both models agree that language processing has an important effect on visual attention. The CIA describes a search of the visual and memory context that results in attending to relevant scene entities, which suggests that the language comprehender *actively pursues* the goal to find a referent in a top-down manner. The FOA, on the other hand, proposes that the language comprehender is *automatically* orienting towards the scene entity or region compatible with the utterance. With regard to memory the two accounts take on separate routes. Within the FOA, present and absent objects are not differentiated. Attention is always directed on the basis of representations build up perceiving the object in question prior to language comprehension. While memory representations are thus a central ingredient to the FOA, there is no notion of decay of these representations. The CIA, on the contrary, explicitly integrates temporal decay of objects in working memory, treating co-present entities qualitatively different. For both accounts, the notion of visual indexing forms the bridge between internal (memory) representations and individual locations in the visual context.

In summary, the most notable differences between the FOA and the CIA lie in the conception and implementation of memory as well as the process of directing visual attention. While ultimately it seems possible to combine the two accounts, further experimental investigation on exactly how memory, visual attention, and the assignment of spatial indices influences situated language processing seems to be necessary. The next section will point out the potential importance of these concepts by comparing the setting of the Visual World Paradigm to natural language comprehension situations.

2.1.6. Beyond the Visual World Paradigm: Situated Language Processing and Non-linguistic Cognition

In the experiments described above, we have seen that given a visual display and an utterance, participants were quickly able to integrate the different sources of information and attended to relevant regions as well as to use visual information for language processing purposes. It is, however, not clear whether the current form of the Visual World Paradigm is suitable to investigate situated language processing in all aspects. In particular, to determine whether the non-linguistic aspects of cognition and possible limitations they impose on the integration of language and situation are accounted for appropriately, it is important to characterize similarities and differences between the Visual World Paradigm and naturally occurring situations in which language is processed. Firstly, many natural language processing situations include two or more interlocutors and the linguistic material consists of dialogue, rather than isolated sentences. We will leave this aspect aside, however, and focus on the differences with regard to the situation. A typical language comprehension situation would not be restricted to a computer screen, but rather include the whole physical surrounding of the language comprehender. Important differences between, e.g., the clip-art displays used in Altmann & Kamide (1999) and a natural situation include the physical nature, the complexity, the dynamics, the physical and the temporal extension of the situation. These aspects will now be discussed in turn to identify questions, which require further experimental investigation.

Perceiving a physical object as opposed to a stylized clip-art picture is likely to lead to a much richer internal representation as it usually exhibits more details and affords physical manipulation. This could, in principle, result in the language comprehender assigning it a higher significance over clip-art pictures. A number of visual world studies used real-world objects that had to be manipulated by the participants (e.g. Tanenhaus et al., 1995). To our knowledge, language-mediated eye movements were similar to those in studies using clip-art scenes, suggesting that participants do not treat physical objects in a privileged way during language processing.

The different degrees of complexity in natural situations as compared to arrays and Ersatzscenes have been pointed out by Henderson & Ferreira (2004), but received little attention otherwise in the visual world literature. The vast majority of experiments used displays with 3-5 objects – this is a number that can be held in working memory simultaneously and arguably might also be attended at the same time. Andersson, Ferreira & Henderson (2011) tackled this issue, by using photographs of highly cluttered scenes. While participants still fixated mentioned objects in this setting, the overall probability was lower and the latency longer than in most existing studies. This suggests that highly simplified scenes provide us with results qualitatively comparable to natural situations, but considerably amplified.

The dynamics of a natural situation compared to a static clip-art display comprise the unfolding nature of ongoing events as well as movements of objects including their appearance and disappearance. Knoeferle & Crocker (2007) and Ellsiepen, Knoeferle & Crocker (2008) approximated an unfolding event using a sequence of clip-art scenes, where the event was completed and the protagonists static at the time the sentence was played. Compared to a static scene that depicted the event as ongoing, the information provided by the event received less consideration, but was still exploited to a certain degree for anticipation and syntactic disambiguation. While this loss in impact could theoretically be related to the dynamic nature of presentation, it is more likely that it is due to the event being completed and having to be retrieved from working memory, while the characters in the scene were still present.

The physical extension of the scenes used in most experiments are small enough to be completely in the field of view and thus easily surveyed without moving the head. In the experiments in Altmann & Kamide (1999), e.g., the whole scene subtended approximately 33° of visual angle horizontally. In contrast to this, a surrounding visual context could only be fully exploited by the listener, if internal representations of the objects out of view were accessible to the language processing system. While no results on immersive environments within the Visual World Paradigm have been reported so far, the blank screen paradigm (Altmann 2004, see section 2.1.4) does address the usage of internal representations. However, in these experiments, the visual context is constrained to only four objects, a quantity that can easily be held in visual working memory simultaneously. A natural scene would almost always surpass this limit by a multitude. While the possibility to use internal representations suggests that in an immersive environment the language comprehender can draw on internal representations for objects out of view, it is problematic to estimate their influence in a visually rich surrounding.

The temporal extension of a scene in a visual world experiment is much shorter than

in a natural situation. In most experiments, the participant is presented with a new visual context with every new sentence, whereas in a natural situation, although there might be dynamic changes to the environment at every instant, a lot of features will stay the same across sentences, or even dialogues. On the one hand, participants in a visual world experiment have thus comparatively little time to build up a representation of the scene. On the other hand, they might be more attentive to their visual context, because it is completely new and it is presented to them in connection with a sentence. Importantly, the task can play a role in whether the participant tries to actively integrate scene and sentence or not. If the task is to manipulate scene objects in accordance to spoken instructions (Tanenhaus et al., 1995; Allopenna et al., 1998), the participant has to establish reference to perform the task. In this case, the influence of purely linguistic processing on the direction of visual attention cannot be distinguished from non-linguistic task-related internal goals. In the look and listen task (Altmann & Kamide, 1999; Knoeferle & Crocker, 2006), on the other hand, the presence of the visual context is not entirely motivated for the participants, giving rise to the possibility that participants actively or implicitly engage in looking for connections between sentence and scene. Of course, this is also possible to happen in natural situated language processing under certain circumstances. Still, the strong stance of language guiding attention automatically loses some power of persuasion with this option.

These differences show that based on the current findings in the Visual World Paradigm, it is difficult to estimate the potential impact of memory and visual attention processes or restrictions on situated language processing. Also, it is not obvious how predictions of the CIA and the FOA translate to more natural language processing situations as the notions of memory and attention remain to some extent unspecified. The next section will provide some background on the cognitive components that we identified as being relevant in order to formulate specific questions on how visual attention, spatial indexing and working memory representations influence situated language processing.

2.2. Cognitive Components Involved in Situated Language Processing

2.2.1. Eye Movements and Visual Attention

The prominent role of eye movements in recent psycholinguistic research as presented in the previous section motivates a closer look on what an eye movement is and why objects or empty regions should be fixated. Let us first look at the two major elements of gaze: the fixation and the saccadic eye movement. A fixation is a period in which the eyes rest

relatively still on one location. This is the time when visual perception takes place. Since acuity is only high in the fovea, the eyes move frequently from one place to another to gather high acuity information of an object or a scene. These eye movements are called saccades and are characterized by short duration (typically below 50 ms), high velocity (up to 500° per second), and the loss of sensitivity to the visual input. Alternatively, they can be described as an *overt* shift of visual attention. When visual attention is allocated to a point in space which is not fixated, we speak of *covert* visual attention. Overt and covert attention are commonly taken to be correlated in the following way: While covert attention can be shifted without eye movements, a saccade is always preceded by a covert shift of visual attention (Henderson, 1992). The question of what influences eye movements is thus closely related to the question of what causes shifts in attention. In general, there are a number of known factors that play a role in the allocation of attention that have been classified as either features of the stimulus (bottom-up influence), or goals or intentions of the observer (top-down influence).

Bottom-up influences include features such as color, luminance and orientation, where regions that diverge from the majority of the scene on these dimensions are more likely to draw attention (Itti & Koch, 2000). Other sources of bottom-up influence are sudden onsets, i.e. objects suddenly appearing in the field of view, or the beginning of a movement (Posner, 1980; Abrams & Christ, 2003). These bottom-up features can influence the orienting of attention automatically. For instance, Posner (1980) showed that after the presentation of a peripheral flash of light, participants were faster in detecting a target object in the same location the flash appeared in than in another location, even in a situation where eye movement were suppressed and when the target object was equally likely to appear in both places. Top-down influences, by contrast, are usually considered non-automatic as they involve the volition of the observer – he can choose to direct his attention to any object, he is currently interested in or search a scene for a specific item. A challenge to this simple distinction between bottom-up and top-down influences comes from the finding that *deictic* stimuli such as arrows or the gaze direction of another person can also direct attention automatically (Driver, Davis, Ricciardelli, Kidd, Maxwell & Baron-Cohen, 1999; Tipples, 2002). Different from, e.g., peripheral flashes they require a certain degree of processing to extract the location they are pointing to and their grip on attention can therefore not be regarded as purely bottom-up and stimulus driven. Nevertheless, they can guide visual attention without the presence of a task-induced internal goal.

It is not entirely clear what kind of an influence language has on the direction of attention. Bottom-up influences are thought to be automatic but they are associated with a visual stimulus that requires little processing. Language mediated attention, on the

other hand, is quite complex: Even if we consider the relatively simple case of referential eye movements, the actual stimulus consists of two parts, namely a referring word and an object that is referred to. An attentional shift in response to this compound stimulus requires the processing of the visual object, the processing of the word, and the forming of a connection between the two: neither the visual object itself nor the linguistic part of the stimulus would be sufficient on its own to direct attention. It is therefore problematic to conceive of referring language as a bottom-up influence. Deictic stimuli like gaze direction and arrows seem to be more closely related to language than pure bottom-up stimuli as they also require processing. In contrast to a referring word, however, they are highly over-learned and have the same denotation in every situation. A referring word, on the other hand, points to its referent which is likely to occupy different locations in different situations.

If, on the other hand, we want to assume that referring language behaved like top-down influences, we have to identify an internal goal of the observer that triggers the direction of attention to the linguistically cued entity. Crucially, two kinds of internal goals have to be distinguished here: On the one hand, attention can be directed consciously and voluntarily to a specific location, if the language comprehender tries to make connections between what he hears and what he sees. In this case, he might actively search for a referent of a noun phrase or the anticipated referent of a verbal argument. On the other hand, the process of language understanding might produce internal goals that give rise to attentional shifts which the language comprehender is not aware of. While the act of language understanding might be deliberate and the language comprehender might direct his attention voluntarily to the linguistic input, the sub-goals that arise from this process could direct visual attention without the influence of volition. Both conceptions of internal goals are compatible with the formulation in the CIA (Knoeferle & Crocker, 2007), where the language comprehender specifically gathers information from the visual domain to facilitate linguistic processing. Both variants of the view that internal goals are controlling language-mediated eye movements are challenged to some degree by the following three experimental findings already: the fixation of empty space in the blank screen paradigm (see section 2.1.4), fixations of semantically related objects (section 2.1.1) and effects of lexical frequency and lexical neighborhood on eye movements.

In the blank screen paradigm, participants tend to fixate locations where relevant objects have been encountered before. Obviously, there is no visual information left in this place that could facilitate language processing. If an internal goal of the language comprehender or some language processing component would cause the shift of attention, this implies that this entity would be blind to the fact that the object is gone. We will

come back to this possibility in the next section, but propose that it seems more likely that participants shifted their attention unconsciously, and possibly without any internal goal. The finding that semantically related objects tend to be fixated, forms another challenge to the internal goal hypothesis. If the goal were establishing of reference and the gathering of visual information, directing attention to an object that is only semantically related would not lead to an accomplishment. It would only be helpful when the speaker made a mistake: In a visual context containing a table with an apple, some flowers, and a cup, it could indeed be beneficial to direct attention to the apple after hearing "Could you hand me the orange, please?" because the apple is probably what the speaker meant. Whether this is enough motivation for directing attention to related objects, however, is to be doubted – especially because in the case of a shape competitor, situations in which one is uttered while meaning the other must be rather rare. The third finding that points towards an automatic, bottom-up like influence of referring language on visual attention is the influence of lexical frequency (the frequency with which this word appears in the language) and lexical neighborhood (the number of words with a phonological overlap with this word) on eye movements (Magnuson, Dixon, Tanenhaus & Aslin, 2007). Since these factors are not perceived by language comprehenders, they must influence attention on a subconscious level. However, it is possible that in this case a subconcious internal goal triggers attentional shifts, especially since the task used in Magnuson et al. (2007) required participants to search for the named object.

As this discussion shows, there are good arguments for both sides, and there is no apparent reason why visual attention should not be guided by automatic, volitional and internal goal-driven processes connected to language processing. If goal-driven factors prove to be dominant, the specific task used for an experiment might decidedly change eye movement patterns as it influences the internal goal in question. Isolated automatic effects, on the other hand, will be less influenced by a task, but might be blurred in eye movement patterns if internal goals or volition interfere.

2.2.2. Spatial Indexing

In the accounts of blank screen findings, spatial indexing takes the role of a mediator between internal memory representations and locations in the world. In this section, we will review the literature on spatial indices and compare different conceptions of what a pointer (a.k.a. spatial index, deictic pointer, visual index) is and how it is used.

The notion of visual indexing was introduced by Pylyshyn & Storm (1988) to account for the finding that people can track up to five moving objects simultaneously. Within

a number of experiments using this *multiple object tracking* task, participants tracked designated objects on a display containing a larger number of visually indistinguishable objects moving with variable velocities and in changing directions, even through occlusions (Scholl & Pylyshyn, 1999). One major difference between the allocation of visual attention and these visual indices is that indices can be assigned to a few objects in parallel instead of being focused on just one location. This number of possible indices is still highly limited, therefore not all objects in an environment can be indexed in this way, making it necessary to specify which objects will be assigned indices and on what basis. Pylyshyn (2001) described two schemes of indexing: in multiple object tracking experiments, participants assigned indices *intentionally* to designated objects. Alternatively, indices can be assigned in a stimulus-driven fashion. In a visual search experiment, e.g., participants were able to access properties of a small group of objects that suddenly appeared faster than the properties of similar objects that remained stable throughout the trial (Burkell & Pylyshyn, 1997). This suggests that the feature of sudden appearance favored objects in receiving an index in a stimulus-driven or bottom-up fashion. The visual indexing mechanism is described as being pre-attentive: Without attention being directed to their individual locations, objects can be assigned an index. An alternative reasoning suggests that visual indices are a form of covert visual attention that is split between locations (Cavanagh & Alvarez, 2005).

A slightly different, but related description of visual indices was put forward by Ballard, Hayhoe, Pook & Rao (1997). In their experiments, they recorded eye movements while participants were copying a pattern of colored blocks. Instead of memorizing the model and then building the copy, participants were frequently looking back and forth from the model and the space the building blocks were in – in many cases even within the placement of a single block. The authors account for this by introducing the notion of deictic pointers. These pointers have features similar to a variable in computer science. They are assigned functionally depending on the task. In their example, one pointer would be set to the block in the model, that is currently being copied. A second pointer would be set to the block in the block space and a third to the location in the copy, where the block should go. In performing the task of finding the appropriate block and putting it in the right location, the model can be easily checked for accuracy by following the pointer rather than conducting a visual search to find the position. This way, only a minimal amount of information has to be stored internally in working memory, while, at the same time, just using three deictic pointers at a time.

Taking the minimization of memory load one step further, it has been proposed that the world can serve as its own memory. Motivated by the theoretical computational

expense which emerges from integrating information between and within saccades into one coherent internal representation, O'Regan (1992) puts forward the hypothesis that in fact we do not rely on internal representations of the visual environment, but that we use this environment directly as a form of external memory. Whenever information about an aspect of this environment is needed, we do not consult internal memory, but acquire the information directly by directing our attention, i.e. gaze, towards the region in question. This hypothesis is supported by the research on *change blindness*. In experiments, participants have been found to fail to notice a substantial change in the scene, if the change was masked by a short distortion in the picture (e.g. Rensink, Regan & Clark, 1997). In a particularly impressive demonstration, some people did not even notice the exchange of their conversational partner after a short disruption (Simons & Levin, 1998). This experimental work suggests that we do not always build a complete representation of our environment. On the other hand, participants are much more likely to notice a change on an object they already fixated before (Hollingworth & Henderson, 2002). It is therefore not plausible that we rely entirely on visual pointers while not retaining any information internally. Still, visual pointers seem to be useful in a system that is incapable of storing all visual information but needs a way to access relevant information from the scene in a direct manner.

An indication that visual pointers might also be used in accessing non-visual properties, comes from an eye-tracking study by Richardson & Spivey (2000). They presented short video clips of different people reciting a random fact in the four quadrants of the screen. After they heard all four facts, they were asked a question concerning one of the facts. Participants frequently re-fixated the quadrant in which the video about that fact used to be, although it was not there any longer. In this experiment, participants were thus following spatial indices while accessing memory, not about the visual properties, but semantic material merely associated with an object appearing in that location. Furthermore, these objects were not present any longer. If this process was aimed at recovering information in the fixated location, it therefore has to be blind to the fact that the object disappeared. Also, it is not clear what kind of information should be recovered, as the semantic content of the spoken fact had never been at that location. This suggests that unlike in Ballard et al. (1997), the indexing process as well as the following of the indices was not task-oriented in this case. It is better explained by an argumentation similar to Altmann & Kamide (2007) discussed above: an automatic re-activation of the episodic trace of perceiving the fact together with an object in a particular location.

To conclude, the research discussed above converge in suggesting there exists a system which indexes specific locations or objects which as a consequence can be accessed directly

by the attentional system. These pointers can be used for tracking objects and manipulating objects but they are also followed in language processing and when recalling semantic information associated with the indexed object. The question of how these indices are established is answered in different ways by different authors: While Pylyshyn (2001) describes a bottom-up, stimulus driven assignment based on salient features, Ballard et al. (1997) assume a functional, task-oriented assignment. Richardson & Spivey (2000) and Altmann (2004) do not specify how and which objects are indexed. In their experiments, both assignment procedures are possible: A bottom-up procedure would index up to five objects with salient features – in the absence of alternatives, all objects would thus be indexed. A task-oriented procedure would also try to index all objects, as in the indexing phase it is not clear yet, which objects will become relevant for the tasks of understanding language or remembering a fact.

2.2.3. Memory

The study of human memory has produced numerous proposals regarding how memory is structured, what information can be retained for how long, and how information is retrieved from memory. Since it is beyond the scope of his thesis to give a comprehensive summary of existing findings and models, this section aims to focus on aspects of experimental work and theory that are relevant to the storing of scene information for retrieval that is triggered by language.

2.2.3.1. Short-term and Long-term Memory

Memory is often conceptualized as consisting of different stores of differing capacities, notably the short-term memory and the long-term memory. While short-term memory is highly limited in capacity and subject to rapid decay, long-term memory can store a seemingly limitless amount of information for periods of time ranging from minutes to a life time. In their multi-store model, Atkinson & Shiffrin (1968) propose that information in short-term memory is rapidly lost and replaced by new information, unless it is rehearsed, which prevents it to be lost. If information remains in short-term memory for a sufficient period, it can enter long-term memory. This view of two distinct systems is supported by the finding of serial position effects in free recall.

In the free recall task, participants study a list of items serially, e.g. spoken words at a fixed rate, and are subsequently asked to recall as many items from the list as possible in any order. In this task, accuracy depends on the position of the item in the study list:

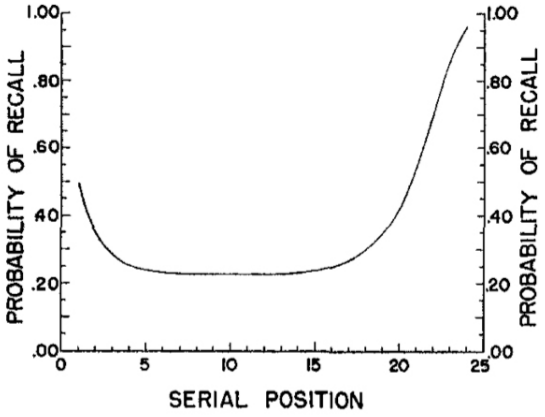


Figure 2.4.: Idealized serial position curve for 24-word list taken from Murdock (1962)

items in the beginning and towards the end of the study list are much more likely to be reproduced correctly than items appearing in the middle of the list. The primacy effect – the relative advantage of items in the beginning of the list – is usually limited to the first 1-3 positions of a list, while the recency effect – the relative advantage of items towards the end of the list – is stretched over more items and rises over the last positions. Murdock (1962) summarized his own findings as well as contemporary research in the idealized curve in 2.4 and describes the primacy part as "rather steep (possibly exponential)" while the recency effect resembles an "**S** shaped curve".

Atkinson & Shiffrin (1968) attribute the primacy effect to items having entered long-term memory while they consider the recency effect to be due to the items still residing in short-term memory. In this view, the reason for only very few items entering long-term memory is the need of rehearsal. During the presentation period, not all objects can be rehearsed sufficiently to be transferred to long-term memory, as new material is coming in and older items in short-term memory are replaced. The very first few items are privileged here because in the yet partly empty short-term store, they do not have to compete with other objects and can be rehearsed by themselves. Additional evidence for this idea comes from the finding of a more pronounced primacy effect with slower presentation rates (Glanzer & Cunitz, 1966): Here, this privileged situation lasts slightly longer, hence items are more likely to be transferred to long-term memory. Also, the recency effect can easily be disrupted by asking participants to count backwards for 30 seconds between the presentation period and the recall period (Postman & Phillips, 1965). The counting introduces new material to short-term memory which then replaces the former content. The primacy effect, on the other hand, is not disrupted in this experiment.

2.2.3.2. Working Memory Capacity

Working memory, which denotes a similar or even equal concept as short-term memory in the description above, is characterized by a limited capacity. There has been considerable debate in the literature on the exact size of this capacity, i.e. how many individual items can be stored in working memory simultaneously. Prominent suggestions were the influential magical number seven ±2, put forward by Miller (1956) which was later adjusted to a magical four by Cowan (2001).

In addition to the specific number of items than can be retained in working memory at a time, the nature of an item must be defined in order to characterize capacity. According to Miller, the kind of information that forms an item is not constant. On top of digits, letters or words, so called *chunks* can function as items, where a chunk is a cluster of several items that can be easily grouped together. Later conceptions of working memory have proposed different stores for information from different modalities. In their multi-component model of working memory, Baddeley & Hitch (1974) assume a specialized verbal buffer, the phonological loop, and a visual component, the visuo-spatial sketchpad. From this point of view, the capacity of working memory has to be established individually for the different components.

For visual working memory, Luck & Vogel (1997) established a limit of only four items – no matter of whether these items were individual features like orientation or color, or whether they were integrated feature conjunctions. This suggests that visual working memory does not store features, but rather integrated objects. Correspondingly, Zimmer (1998) found that the *location* of an object was automatically stored in the context of a comparison task where only the form of the object was relevant.

In addition to visual features such as the form and the location of an object, the name of an object belonging to the verbal modality is relevant for its representation in the context of the Visual World Paradigm. Firstly, participants are likely to name objects in the preview phase, as they expect some of them to be mentioned later. Secondly, the name triggers the re-fixation of the object or its location, in addition to conceptual properties such as being edible. Therefore visual features, verbal content, and conceptual information has to be kept in memory in order to allow language-mediated eye movements on a blank screen. The multi-component model of working memory is able to accommodate these requirements to a certain degree by means of the episodic buffer, introduced by Baddeley (2000). In this episodic buffer, information from different modalities can be combined to form integrated objects. Note, however, that only the perceptual features (i.e., phonological and visual information) originate from short-term memory proper, while the conceptual or

semantic features are integrated by allowing for an interface to long-term memory. Similar to verbal and visual subsystems, the episodic buffer is capacity limited.

2.2.3.3. Single Store Models

In contrast to the conceptions of working memory described above, which to some extent posit a separate repository where specific items are stored, more recent models describe working memory as the activated part of long-term memory (O'Reilly, Braver & Cohen, 1999; Cowan, 2001; Nairne, 2002; McElree, 2006; Oberauer, 2002). Although there are different view points on whether working memory is still to be considered a separate component which shows different characteristics than (non-activated) long-term memory, the neural substrate is conceptualized as being shared. An "item" or *trace* is rather a bundle of activated features, in this view. Recent evidence comes from the lack of a neural dissociation between short-term and long-term memory processes (Oztekin, Davachi & McElree, 2010). To account for forgetting, the activation is thought to decay with time (Cowan, 2001), or is overridden by new information from the same modality (Nairne, 2002).

An early account of serial position effects under the single store hypothesis was put forward by Craik & Lockhart (1972). They attribute the characteristics of primacy and recency to different levels of processing. Similar to the argumentation that early list items can be rehearsed by themselves, they propose that early list items can be processed more deeply, including the semantic level and the triggering of associations. Later items are only processed on a superficial, possibly phonemic level, which can be easily overridden by new material. According to Craik & Lockhart (1972), deeper levels of processing lead to more stable representations in memory. Taking into account the more recent single store models, that entail the activation of features in long-term memory, this stability can also be derived directly from the assumption that semantic processing of the stimulus lead to a broader activation pattern, in which individual features subsequently spread activation to other features within the same representation (McClelland et al., 1986). As each feature thus receives a constant activation boost, the whole representation will be more likely to survive than a shallow representation, where the individual features only experience a limited re-activation by other features.

2.3. Towards an Account of Situated Language Processing Considering Cognitive Limitations

As described in section 2.1.6, generalizing from visual world experiments to more ecological language processing situations requires examination of the cognitive components that are involved, namely visual attention, spatial indexing and working memory. In turn, these components, their significance in situated language processing, and experimentation needed to establish a more adequate integration into situated language processing models will be discussed.

Visual world experiments were able to establish a linkage between visual attention and language processing. The prominent question that has not yet been answered satisfyingly is whether visual attention is influenced automatically by language similar to bottom-up influences, whether it is necessary to voluntarily attend the linguistic input, or whether an actively pursued aim to connect picture and sentence was predominant in visual world experiments. To test this, it is necessary to disentangle the task from the processing of both visual and linguistic stimuli. Even the very simple look & listen task does not accomplish this separation. Although the two components are not causally connected, participants are asked to attend to visual and linguistic stimuli at the same time promoting the forming of connections between both modalities. In the experiments described in chapter 3, the potential automaticity of language mediated visual attention is addressed directly. This is achieved by examining the effect of language processing on covert visual attention while disentangling task, linguistic stimulus, and visual stimulus.

Although experiments from the blank screen paradigm are interpreted by means of spatial indexing, no specific theory of spatial index assignment and the nature of spatial indices has gained consensus. However, whether the assignment of spatial pointers happens in a top-down or bottom-up fashion has important implications for situated language understanding: If pointers are merely assigned top down, a language comprehender would be unable to integrate visual information which is currently out of view and has not been considered relevant to the task of language understanding so far. If, on the other hand, pointer assignment would be strictly bottom-up, only the most prominent objects in the situation would be accessible for language processing. In experiment S1 in chapter 4, bottom-up and top-down pointer assignments are contrasted.

The second unresolved issue regarding spatial pointers is their nature: following the logic of Pylyshyn and Ballard, the pointers are really only indications of locations not connected to any internal content in the mind. Altmann & Kamide (2007)'s linking theory, on the other hand, associates the spatial location directly with an "episodic trace", that

is, a representation in memory. If we reconsider the sparse conceptualization of a pointer that only consists of a label and an associated location, it becomes hard to account for anticipatory eye movements on the blank screen. For referential eye movements, the label might be directly accessed by the acoustically perceived name, but anticipatory eye movements likely rely on a more conceptual representation, e.g. the affordance of being edible in the example in Altmann & Kamide (1999). This sparse conceptualization therefore seems to accommodate referential eye movements much better than anticipatory eye movements, which leads to the prediction that on the blank screen anticipatory eye movements should occur less in comparison. If, on the other hand, we accept that pointers are not only these sparse conjunctions of location and label, it is important to establish which kinds of memory representations can be tied to them. From the discussion above, we saw that we can either distinguish between short-term and long-term memory representations, or, under the assumption of a unitary store, between activation patterns of different strength and scope. Without fully committing to one or the other, we will distinguish between shallow representations, that could be either residing in short-term memory, or consist of a limited activation pattern only including surface features, and conceptual representations, that are either part of long-term memory, or constitute rich activation patterns including semantic features and associations. An open question so far is whether both, shallow and conceptual representations of visual objects can be connected to visual pointers and are accessible for language-mediated eye movements. Experiments S2 and S3 in chapter 4 and Experiment W in chapter 5 investigate the extent to which these representations can be the basis for language-mediated eye movements.

Chapter 3.

Covert Visual Attention: The Automatic Influence of a Word-picture Pair

Psycholinguistic research conducted within the Visual World Paradigm has shown that language may guide eye movements in a visual scene (Cooper, 1974; Tanenhaus et al., 1995; Altmann & Kamide, 1999). In particular, referents of linguistic expressions are often fixated while processing their name in situated language understanding. In the last chapter, two different models of how linguistic input together with world knowledge and scene information might drive eye movements were described. Altmann & Kamide (2007) followed Tanenhaus et al. (2000) in assuming that language causes unconscious shifts of attention that may then result in an eye movement. They also attribute eye movements in the Blank Screen Paradigm (i.e., when the visual scene was removed before the start of the linguistic stimulus) to the automatic activation of the location of an object when it is referred to. They argue that the (former) location is part of the episodic trace of experiencing that object and that all aspects of this episodic trace get slightly activated if another aspect, e.g. the name, is experienced again. If the activation of the (former) location is high enough, an eye movement towards it is conducted despite the object not being there any longer. This account thus supposes language-mediated shifts in visual attention to be an automatic process. Knoeferle & Crocker (2007), on the other hand, identify the internal goal of establishing reference and gathering information from the scene as the source of language-mediated eye movements. This internal goal could arise either as a by-product of language processing, or it could be the deliberate effort to make sense of both sentence and scene.

While we know that visual attention can be directed by automatic bottom-up influences, as well as volitional control and top-down influences (see section 2.2.1), it is not entirely clear to which class eye movements in the visual world paradigm belong. Participants might simply choose to look at a particular object or gaze might be influenced by a concurrent

task in contrast or in addition to an automatic process. In the case of a spoken instruction to click on a particular object, for instance, they need to direct their gaze to that object in order to perform well on the task. Although with this particular task, it has been shown that even manipulations that participants are typically not aware of (e.g. lexical frequency or lexical neighborhood density, see Magnuson et al., 2007) affect gaze behavior early on, it remains unclear to what degree these findings generalize to settings in which the spoken word is not relevant for the task. Sentence level studies, on the other hand, often employ the "no-task" or "look & listen" task of just looking at the picture and listening to the sentence in order to understand it (Altmann & Kamide, 1999; Knoeferle et al., 2005). In this case, there is no explicit need for participants to synchronize their gaze with the spoken sentence, but since their visual attention is not required for anything else, it is possible that participants are doing this consciously. Neither the emergence of language-mediated eye movements in "look and listen" experiments nor the influence of subtle manipulations in combination with a specific click-on task are thus sufficient evidence to conclude an automatic influence of language on visual attention.

In order to differentiate automatic from volitional influences on visual attention, the experiments presented in this chapter examine shifts of *covert* visual attention (i.e. shifts of visual attention without eye movements) and disentangle the task given to the participants from the processing of both, linguistic and visual stimulus. We use covert attention, because within the spatial cueing paradigm (Posner, 1980), the orienting of covert attention can be tested straightforwardly for its degree of automaticity. This paradigm further enables us to separate the task from the processing of the stimulus by varying the degree to which the processing of the visual and linguistic components and their integration is encouraged or discouraged. The underlying assumption that makes the spatial cueing paradigm suitable here is that we expect language to affect covert visual attention in a similar way as it affects eye movements, since the planning of a saccade involves prior orienting of covert attention to its destination (Henderson, 1992). Nevertheless, it is thus far an open question whether covert attention is influenced by referential language, if eye movements are suppressed.

In a standard spatial cueing experiment, the participant sees a display with a central fixation cross and has to detect a particular target or make a binary decision about it (Posner, 1980). Before that target appears, one location in the display is cued. In a valid or *compatible* trial, the target appears in the cued location, whereas in an invalid or *incompatible* trial, it appears in a different location. We speak of a cueing effect, if reaction times are shorter in compatible trials than in incompatible ones. The paradigm is well suited to study two aspects of automaticity: speed and the influence of volition. Speed can be examined by varying the time interval between the cue and the target. Automatic

orienting towards peripheral flashes have been shown to occur around 100-200 ms after the cue, while volitional orienting can take up to 1200 ms (Friesen, Ristic & Kingstone, 2004). The influence of volition is tested by varying the predictiveness of the cue. If the cue is compatible in the majority of trials (predictive), participants are expected to voluntarily orient in order for them to optimise their response. If the validity of the cue is at chance level (unpredictive), on the other hand, participants are most likely to ignore it, because it cannot help them. In this case, a cueing effect indicates an automatic influence in the sense that participants do not engage in volitional control. If, in a third case, the cue is incompatible more often and thereby systematically points to the wrong location (counter-predictive), participants are expected to voluntarily orient away from it to the other, opposite location. An advantage for the cued location here indicates automaticity in that participants are not able to voluntarily orient away from it.

Prior research on covert attention has shown cueing effects for direct or peripheral cues, e.g. sudden onsets or flashes of light in the cued location (Posner, 1980), but also for symbolic cues which need some kind of interpretation like arrows, gaze, and directional words (Posner, 1980; Driver et al., 1999; Hommel, Pratt, Colzato & Godijn, 2001; Tipples, 2002, 2008). In contrast to an early study suggesting that orienting in accordance to arrow cues is only possible if volitional control is engaged (Jonides, 1981), more recent studies were able to detect an automatic (fast and involuntary) influence for all three types of symbolic cues, when the cue was either uninformative or counter-predictive (Hommel et al., 2001; Tipples, 2008). The present study deploys complex cues which are composed of an object photograph and a spoken word referring to it. Although such a cue is clearly symbolic, as it needs some interpretation, it is qualitatively different from the symbolic cues described so far. Arrows, gaze and directional words (left, right) are all deictic in nature in that they inherently point in one direction or the other. In our case of a compound picture-word cue, the word alone does not point to any location, nor does the picture itself. Only by connecting the word to its referent, i.e. the picture, it becomes a spatial cue. While the word *left*, for example, has the same orienting effect in every situation, orienting in response to a referring expression would depent on the (prior) location of the referent and could be targeting left or right equally likely. The experiments presented in this chapter test whether such referring language can guide covert attention similar to other, deictic symbolic cues.

In the context of situated language understanding, we can differentiate several hypotheses regarding the automaticity and the involvement of volition and internal goals in the mediation of attention by language. The *automaticity hypothesis* entails that a linguistic stimulus will under all circumstances direct attention to a relevant object. Since volition

is not affecting the orienting process, a concurrent task does not have an influence. The *volitional orienting hypothesis*, on the other side of the spectrum, states that a language comprehender can choose whether she directs her attention to language relevant objects or not. A conflicting task would therefore prevent the linguistic stimulus to have an effect on visual attention. In between these two extremes, we can formulate a third *internal goal hypothesis*. It states that attention is directed by an internal goal emerging from the process of understanding the linguistic stimulus without the language comprehender engaging volitional control over her attentional system. A concurrent task would not inhibit the direction of attention as long as the linguistic stimulus is processed. If the linguistic stimulus is ignored, however, no attentional shifts are predicted. If we can confirm the automaticity hypothesis, this supports the linking theory of Altmann & Kamide (2007). The intentional orienting hypothesis as well as the internal goal hypothesis, however, are compatible with the view of Knoeferle & Crocker (2007) and a challenge to the former.

In order to differentiate between these hypotheses, the experiments reported in this chapter use predictive, uninformative and counter-predictive cues. The automaticity hypothesis predicts a cueing effect for all three types, as the effect should not be contingent on the influence of volition or the relevance of the linguistic stimulus to the task. The intentional orienting hypothesis predicts a cueing effect to occur with predictive cues, but not with uninformative ones, because in the former case participants engage in volitional control to orient towards the cued object, whereas in the latter case they do not as this would interfere with their task. For counter-predictive cues, the intentional orienting hypothesis predicts participants to reliably orient *away* from the cue and not towards it. Finally, the internal goal hypothesis predicts a cueing effect for predictive and counter-predictive cues since in both cases the task requires participants to process the word and integrate it with the visual stimulus. Processing the linguistic portion of a counter-predictive cue will thus direct attention towards the referenced object although volition is engaged in orienting away from it. For the uninformative cues, no cueing effect is predicted, as the linguistic stimulus is ignored and therefore no language processing component can give rise to the internal goal of establishing reference. In the first experiment, predictive and uninformative cues are contrasted, the second experiment addresses counter-predictive cues.

3.1. Experiment A1: Predictive and Uninformative Word-Picture Cues

This experiment examined, whether covert attention is influenced by referring language and if so, whether that is dependent on volitional control or internal goals. Participants

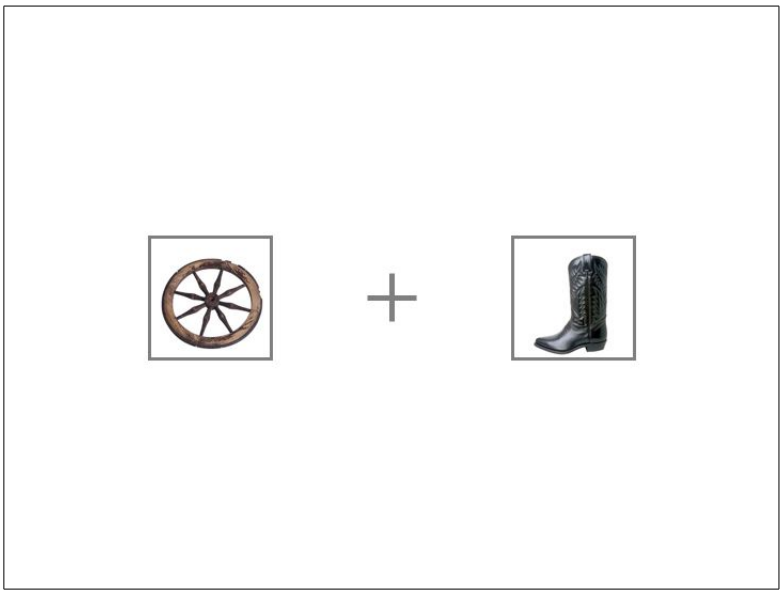


Figure 3.1.: Display of item 108, the words used are *wheel* and *boot*

had to detect a target object appearing in one of two boxes. Preceding the target, they first saw two object photographs in the boxes and subsequently heard a spoken word that referred to one of those objects. The cue thus consisted of the object photograph that was referred to, and the word that was referring. In a compatible cue trial, the target appeared in the box that was previously occupied by the object that was referred to, whereas in an incompatible cue trial, the target appeared in the opposite location. In case of a cueing effect, responses are expected to be faster in the compatible condition. In addition to trial type, the stimulus-onset-asynchrony (SOA) was varied between 200, 500 and 800 ms. On the one hand, this manipulation aimed to maximize the chance of detecting an effect, because it is unclear how the SOAs from previous research would translate to the use of spoken referring words. On the other hand, this manipulation enabled us to assess how quickly an effect emerges with automatic processes expected earlier then those under volitional control. To directly test whether an orienting effect is dependent on volitional control the predictiveness of the cue was manipulated between participants: In the predictive condition, the cue was compatible in 75% of the trials while in the uninformative condition the cue was equally likely to be compatible or incompatible. In the predictive condition, the optimal strategy in order to perform well on the task was thus to actively orient towards the cue, whereas in the uninformative condition, the best strategy was ignoring the spoken word. In addition, the temporal delay between the presentation of picture and spoken word further discouraged the volitional forming of a connection between the two.

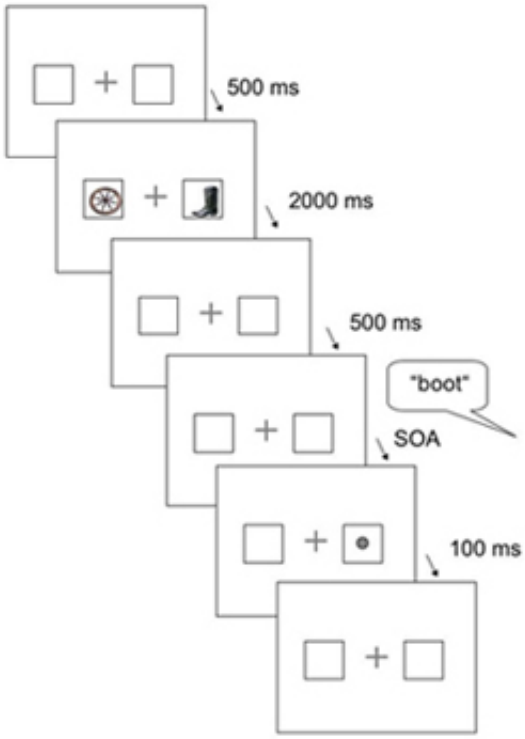


Figure 3.2.: Procedure of Experiment A1 & A2

3.1.1. Method

3.1.1.1. Participants

Twenty-eight students from the University of Edinburgh participated in the experiment for course credit. They were all native speakers of British English. Age ranged from 18 to 32 with a mean of 19.18. Seven participants were male.

3.1.1.2. Materials

192 experimental items, 48 catch items and 10 practice items were created (see A.1 for a list of all experimental items). An experimental item consisted of two object photographs taken from the commercial collection Hemera Photo Objects and two pre-recorded words referring to them. The photographs measured 120 x 120 px corresponding to 3.4° of visual angle. They were displayed 250 px to the left or right of fixation (7.1° of visual angle). The words were all British English monosyllabic picturable nouns and they were matched within items for frequency (Leech, Rayson & Wilson, 2001) and length. In a given trial, only one of the word recordings was used, there were thus two versions of each item naming either one

of the objects. Those two versions were used in separate lists for counterbalancing. In order to construct a sufficient number of items, all pictures and part of the words were repeated once during the experiment in new pairings. Practice items were similar to experimental items, except they were not matched for frequency and length. Each experimental item was randomly assigned to one of the six conditions resulting from crossing TrialType (`compatible`, `incompatible`) with SOA (`200`, `500`, `800` ms). In each of the six conditions, the named object was equally often on the left or right side of fixation and also the target appeared on both sides of fixation equally often. Predictiveness was manipulated between participants. In the lists used in the `predictive` condition, 75% of experimental trials were `compatible` and 25% were `incompatible`. In the `uninformative` condition, 50% of the experimental trials were `compatible` and 50% `incompatible`. In a catch trial, there was no target. Those were included to prevent participants from responding habitually. The lists were randomized individually for every participant.

3.1.1.3. Procedure

Participants were seated in front of an Eye-link 1000 remote eye-tracker with a viewing distance of 80 cm to a 20 inch monitor. As illustrated in Figure 1, each trial started with a fixation cross in the middle of the screen and two empty boxes on the left and right side of the fixation cross. After 500 ms the pictures appeared in the boxes and stayed there for 2000 ms. 500 ms later the word was played back over loudspeakers. Depending on the SOA condition, 200, 500, or 800 ms after word onset the target, a small grey circle, appeared for 100 ms in one of the boxes and participants were asked to press the space bar on the keyboard as fast as they could when they detected it. The trial terminated with the participant's response or after 1500 ms. After a delay of 1000 ms the next trial started automatically. After every 24 trials, 9 times during the experiment in total, participants were given the chance to have a short rest before going on.

Participants were instructed to keep their eyes fixed on the fixation cross throughout the trial. To make sure they attended to the pictures and tried to identify them, they were told there would be a memory test for the pictures after the experiment. They were informed about the probability of the target appearing in the cued location (50-50, 75-25 respectively) and suggested to use the optimal strategy: In the `uninformative` condition they were encouraged to ignore the word and only pay attention to the target whereas in the `predictive` condition they were encouraged to pay attention to the word and use it as a hint of where to expect the target.

At the beginning of the experiment there was a practice phase including 10 trials.

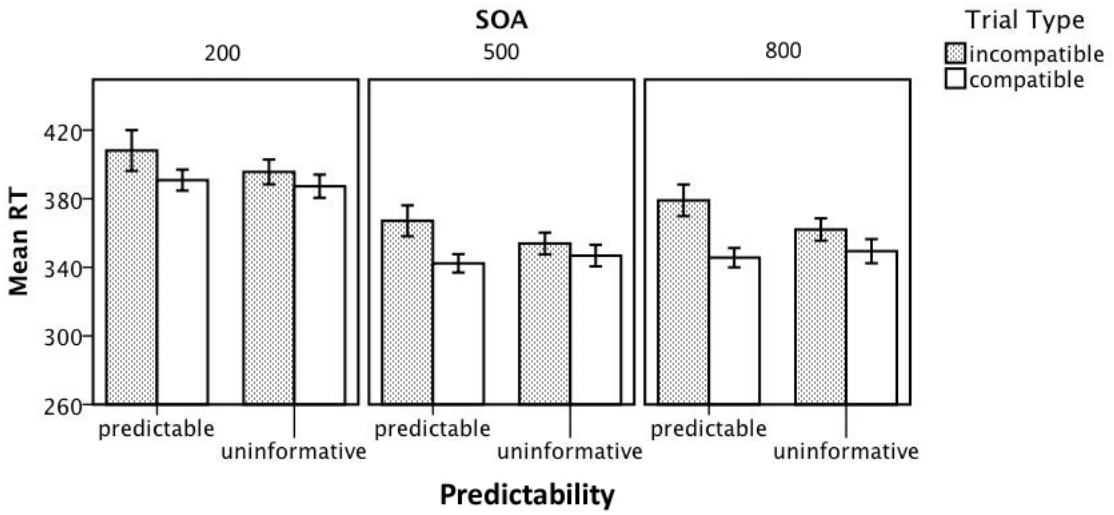


Figure 3.3.: Mean RTs of Experiment A1 by Predictability, SOA, and TrialType

During practice, subjects were provided with feed-back if they looked away from the fixation cross and if they made a response in a catch trial.

3.1.2. Results

Saccades away from the fixation cross were infrequent (15%) and did not affect RTs.[1] The overall error rate was very low (< 5%) and was not affected by the experimental manipulations. RTs further than 2 standard deviations from each individual participant's mean were removed (< 2%). We conducted repeated measures ANOVAs on the averaged RTs with the with-in participants factors TTYPE(`compatible, incompatible`) and SOA(`200, 500, 800`) and the between participants factor PRED(`predictive, uninformative`) (Figure 3.3). The compound cue of referring word and previously viewed object triggered an orientation to the cued location which was evidenced by shorter RTs in the `compatible` cue condition ($F(1,26) = 40.18$, $p < .001$). There was also a main effect of SOA ($F(2,52) = 45.79$, $p < .001$) due to shorter RTs with longer SOAs, but no interaction between TTYPE and SOA, indicating that the orienting effect was not dependent on any specific SOA. The main effect of PRED was not significant ($F < 1$), but there was an interaction between PRED and TTYPE ($F(1,26) = 11.21$, $p < .01$) due to a smaller difference between `compatible` and `incompatible` trials in the `uninformative` condition (8 ms) than in the `predictive` condition (24ms). A follow-up ANOVA on the subset in the `uninformative` condition confirmed the main effect of TTYPE ($F(1,13) = 12$, $p < .01$) for that group.

[1] We conducted a separate analysis which excluded those trials in which participants executed a saccade away from the fixation cross. It confirmed the analyses reported here.

Although the graph suggests a trend towards a more pronounced cueing effect with increasing SOA for the `predictive` group only, the three-way interaction of TTYPE, PRED and SOA was not significant ($F < 1$).

The results show that participants oriented their attention covertly in response to referring language in a situation where eye movements were suppressed. This orienting effect was neither entirely dependent on volitional control, nor on the intention to process the linguistic stimulus: Finding the effect in the uninformative group rather suggests that picture and word were integrated automatically although they were not co-present and the task was entirely independent of picture and word. In addition to this automatic effect, the results suggest a modulation of the orienting effect by volitional control or internal goals arising from the processing of the word. The interaction revealed that the cueing effect was stronger in the situation where it was relevant to the task and participants were encouraged to make use of it. In summary, the results suggest an automatic influence of a referring word on covert visual attention which is enhanced by either volitional control or the presence of internal goals arising from linguistic processing.

3.2. Experiment A2: Counter-predictive Word-Picture Cues

Experiment A1 demonstrated that a referring word directs covert visual attention. We identified an automatic component, that was effective even if the task did not encourage forming a connection between word and picture. On the other hand, the orienting effect was enhanced by the influence of either volitional control or the processing of the linguistic stimulus as evidenced in the condition where the task encouraged the use of the cue. This experiment contrasts volitional orienting with either automatic effects or effects emerging from linguistic processing by using counter-predictive cues. These cues are more likely to point to the wrong location which encourages participants to orient away from it to the opposite location. Existing studies using counter-predictive eye gaze and arrows have shown that at short SOAs an automatic attentional capture of the cued location was predominant, while at longer SOAs the participants successfully oriented their attention away from the cued location, resulting in a benefit at the location originally containing the incompatible cue (Driver et al., 1999; Tipples, 2008). In this experiment the internal goal hypothesis, i.e. the direction of attention by internal goals arising from linguistic processing, makes the same prediction as the automaticity hypothesis: automatic attentional capture and the goal of establishing reference both predict an orienting effect towards the cued position, whereas the volitional control hypothesis predicts attentional orienting towards the opposite location, where the target object is predicted to appear.

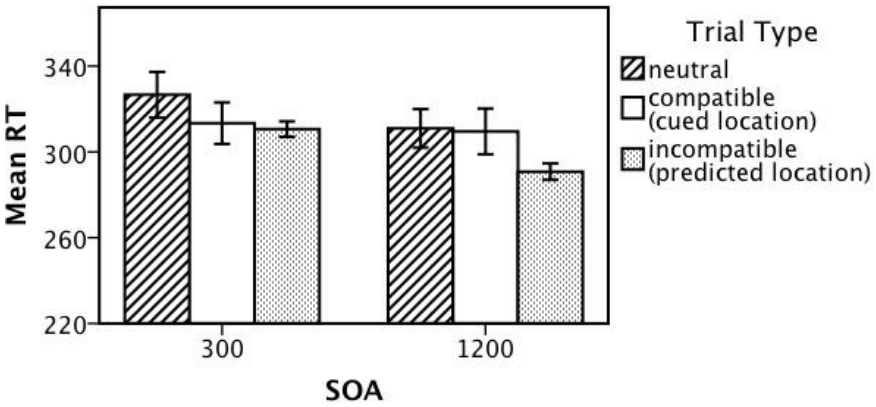


Figure 3.4.: Mean RTs of Experiment A2 by SOA and TrialType

3.2.1. Method

3.2.1.1. Participants

Twelve students from Edinburgh University took part for course credit. They were all native speakers of British English.

3.2.1.2. Materials and Procedure

Materials and Procedure were similar to Experiment A1 with the following exceptions: A `neutral` condition was included in which the word referred to neither of the pictures as a baseline against which to compare the compatible and the incompatible conditions. To increase the chance of detecting a possibly small effect, more items were created using the same pictures and recordings as in Experiment A1 three times each (see A.1 for the complete item list). Of the total number of 288 experimental trials, 75 % were `incompatible` (i.e., predicted), 12.5 % were `compatible` (i.e., cued and not predicted) and 12.5 % were `neutral` trials. Additionally, 36 catch trials were included. There were only two SOAs: 300 ms (`short`) and 1200 ms (`long`). The short SOA was expected to reflect mainly automatic orienting while the long SOA was expected to reflect only volitional orienting. Participants again were informed of the probability the target would appear in the cued location and their optimal strategy of orienting away from it. Responses were given using a button on a control pad because of higher precision in comparison to a keyboard.

3.2.2. Results

Similar to Experiment A1, the overall error rate was very low (< 5%) and not further analyzed. RTs further than 2 standard deviations from each individual participant's mean were removed. We conducted a repeated measures ANOVA on the averaged RTs (Figure 3.4) with the within participants factors TTYPE (`neutral, compatible, incompatible`) and SOA (`short, long`). Responses in the `incompatible` condition were faster, as evidenced by a main effect of TTYPE ($F(2,22) = 7.43$, $p < .01$), indicating a volitional orienting effect away from the cue. Also, RTs were shorter for the `long` SOA condition ($F(1,11) = 5.67$, $p < .05$). Most interestingly, there was an interaction between TTYPE and SOA ($F(2,22) = 5.20$, $p < .05$). Pairwise comparisons revealed that in the `short` SOA condition, RTs were shorter for both, the `compatible` and the `incompatible` condition compared to the `neutral` condition. This indicates an automatic attentional capture of the cue in addition to a volitional orienting effect away from the cue. In the `long` SOA condition, on the other hand, responses were equally slow in the `neutral` and in the `compatible` condition, leaving an advantage only for the predicted `incompatible` condition. We thus find an early advantage for the cued side in addition to a robust orienting effect towards the predicted side. The early cueing effect in this experiment cannot be due to volitional orienting, because volitional control was engaged in orienting away from the cue. These results are thus further evidence for an automatic influence of referring language on visual attention.

3.3. Discussion

The above presented experiments show a fast and involuntary influence of cues composed of a referring word and a picture of the referent on covert visual attention. In Experiment A1, participants were faster to detect a target in a position cued by a formerly present object and an auditorily presented word referring to it. The effect was present regardless of whether the cue was predictive of the location in which the cue would appear or not, but the difference between detection times in cued or uncued locations was greater in the predictive condition group, indicating a modulation of the orienting of covert visual attention by volitional control. Experiment A2 showed that even if participants engaged in orienting away from the word-picture cue, there was an early advantage in cued trials compared to neutral trials. These results are in line with existing studies on deictic symbolic cues and extend their findings to a qualitatively new class of cues that require the establishing of a link between two entities from different modalities: a referring word and its depicted referent.

In contrast to studies from the visual world paradigm, the experiments presented here disentangled the task from the integration of picture and word. For the uninformative group, the word was irrelevant for the task and was thus expected to be ignored. In addition, word and picture were not present simultaneously, which further discouraged an integration. In this condition, we therefore expected volitional control or language processing induced internal goals to have no effect, but were able to observe an automatic orienting effect towards the former location of the picture. For the predictive group, on the other hand, an integration of word and picture was beneficial for completing the task. The enhanced cueing effect here can thus be attributed to the engagement of volitional control or internal goals on both, integrating picture and word and orienting towards it. This enhancement suggests that within the visual world paradigm using a task that favors volitional shifts of attention in accordance to the utterance will quantitatively - and possibly qualitatively (i.e., resulting in different patterns) - change the observed eye movements. In Experiment A2, both, automatic and volitional orientation were examined at the same time: While the task encouraged integration of picture and word, volitional control engaged in orienting away from it. While at the long SOA, only this volitional orienting was observed, integration of picture and word succeeded in capturing part of the attention in the early SOA. Finding both, voluntary orienting and attentional capture at the same time can be explained in two ways. Either attention was split between both locations in this stage, or participants were in one of two consecutive states: an early orientation towards the referred object or the later volitional orientation away from it. In this line of reasoning, participants would only be faster in a certain portion of trials, of course, but in comparison to the neutral condition they would still be faster on average.

The fast and involuntary cueing effect observed supports the hypothesis underlying the linking hypothesis put forward by Altmann & Kamide (2007) that linguistic information and visual information are integrated automatically (i.e., fast and largely involuntary) and that reactivation of episodic traces automatically causes shifts of attention. The additional effect of volition, however, is not accounted for by this theory. The alternative linking hypothesis CIA put forward by Knoeferle & Crocker (2007), on the other hand, does not explicitly state whether language-mediated eye movements are due to automatic or volitional processes which is compatible with the results, although they are not directly predicted. The CIA, however, attributes language-mediated eye movements to the top-down goal of establishing reference. Arguably, no such top-down goal is pursued in the case of the uninformative group in Experiment A1. In summary, a complete linking hypothesis between language processing and gaze behavior should take into account the automatic integration of picture and word as well as volitional and language-related internal goals connected to the establishment of reference that may depend on the task.

Our results are consistent with the findings of Salverda & Altmann (2011) in that both studies demonstrate a fast and involuntary integration of visual and linguistic information. In contrast to their study, we uncoupled visual attention and eye movements completely and we removed the picture before we presented the word. The latter makes our results more suitable to account for the findings of the Blank Screen variant of the Visual World Paradigm and show that the fact that participants moved their eyes on a completely blank screen could indeed be due to the automatic attentional orienting triggered by language.

Chapter 4.

Sentence-level Studies: Anticipatory and Referential Eye movements

Existing research within the Blank Screen Paradigm shows that language-driven eye movements occur not only when the visual world is co-present, but also when it has been viewed before language is presented and is replaced by a blank screen (Altmann, 2004; Knoeferle & Crocker, 2007). There has been, however, little research testing these effects in a more diverse setting – most experiments featured displays containing only four objects. The inferences concerning the use of currently not available visual information that can be drawn from these studies are therefore limited. In particular, this poses a problem for the generalization of experimental results to more natural language comprehension situations as language is often processed in an environment containing a multitude of objects both within and out of view. Memory for visual objects in the short term has been reported to be accurate for four objects on average (e.g. Luck & Vogel, 1997). If a display contains more than four objects, which is true for most real life situations, it remains unclear, whether these objects are at all accessible to language processing and how limitations of memory have to be factored in. Conceivably, language-mediated eye movements in such a situation might turn out to be less accurate, to occur less often, or to favor objects which are still readily accessible. Alternatively, visual information might be ignored if its use and organization requires too much effort.

This chapter describes three experiments in which participants were presented with seven objects sequentially guided by the following objectives: First of all, we aimed to establish whether language-driven eye movements occur at all if the number of objects on display exceeds visual working memory capacity. Further, we aimed to test how the accessibility of object locations for language mediated gaze depends on the object's position in the sequence of presentation. The last point is important in two ways for the understanding of the underlying mechanisms of language-mediated gaze: Firstly, it allows

us to draw conclusions about the assignment procedure of spatial indices (see Section 2.2.2). Secondly, the potential emergence of serial position effects could inform us about the question whether shallow memory representations sometimes associated with short-term memory and more rich, conceptual memory representations are both accessible for language-mediated eye movements in situated language processing. Similar to Altmann & Kamide (1999) and Altmann (2004), participants were presented with a sentence containing a restrictive verb, where only one of the displayed objects was a plausible role-filler. This design allowed us to examine both, anticipatory eye movements (see 2.1.2) triggered by the verb, and referential eye movements (2.1.1) triggered by the noun phrase that names the object. While Experiment S1 targeted primarily at insights on the visual index assignment procedure, Experiments S2 and S3 test the idea that anticipatory and referential eye movements might rely on different memory representations.

4.1. Experiment S1

In this experiment, one character and six objects appeared one by one on a display, which went blank before a sentence containing a restrictive verb was played back. After the sentence finished, the pictures were shown again and the participant's task was to click on the object mentioned in the sentence. Language-mediated eye movements were assessed by comparing the number of trials containing looks to the location previously occupied by the target object, which was mentioned in the sentence, to the number of trials with looks to a location occupied by a comparison object (comparitor).

Existing blank screen studies favor the concept of spatial indices to account for the eye movements to empty regions (Richardson & Spivey, 2000; Altmann, 2004). If eye movements on the blank screen are indeed due to the eyes automatically following the index associated with the name of an object and if the number of indices is limited and smaller than the number of objects in our experiment (7), language-mediated eye movements would only be expected to occur if the target object is one of those objects that received an index. The question of which objects are indexed and which objects are not is assumed to depend on the underlying scheme of index assignment. In the next paragraphs, possible index assignment procedures and their consequences for language-mediated eye movements are discussed.

Bottom-up index assignment Pylyshyn (2001) assumes bottom-up effects to dominate the assignment of indices. He found that people can entertain 4 to 5 indices at a time and that entities that are visually salient will attract indices. One of the features that

makes an object visually salient is its abrupt onset or sudden appearance, as utilized in this experiment. Every new appearing object should thus be assigned an index, possibly overwriting existing assignments. If Pylyshyn's estimation of 4-5 indices is correct, we would therefore expect that the target will be indexed if it is one of the last five objects to appear. If it appears earlier, the index is most likely already overwritten. The described index assignment should then result in reliable looks to the target if it appears late, whereas in the case where it appears early, there should be no or only a small number of anticipatory and referential looks towards the target.

Top-down index assignment If indices are only assigned top-down (Ballard et al., 1997), that is, driven by the requirements of a concurrent task, manipulating the temporal order is not expected to have any effect. The click-on-the-object task we use here should not favor any particular index assignment, as all objects are equally likely to be mentioned and therefore target to the task. Since the number of objects exceeds the number of indices available, however, not all objects that will possibly turn out to be relevant can receive an index. We will assume that in the absence of any plausible top-down assignment strategy, a limited number of objects are picked at random to be indexed. With this procedure, the target will not receive an index on every trial. Therefore language-mediated eye movements will in general occur less frequently than in blank screen experiments with fewer objects on display, but they will not be influenced by the serial position of the target in the presentation sequence.

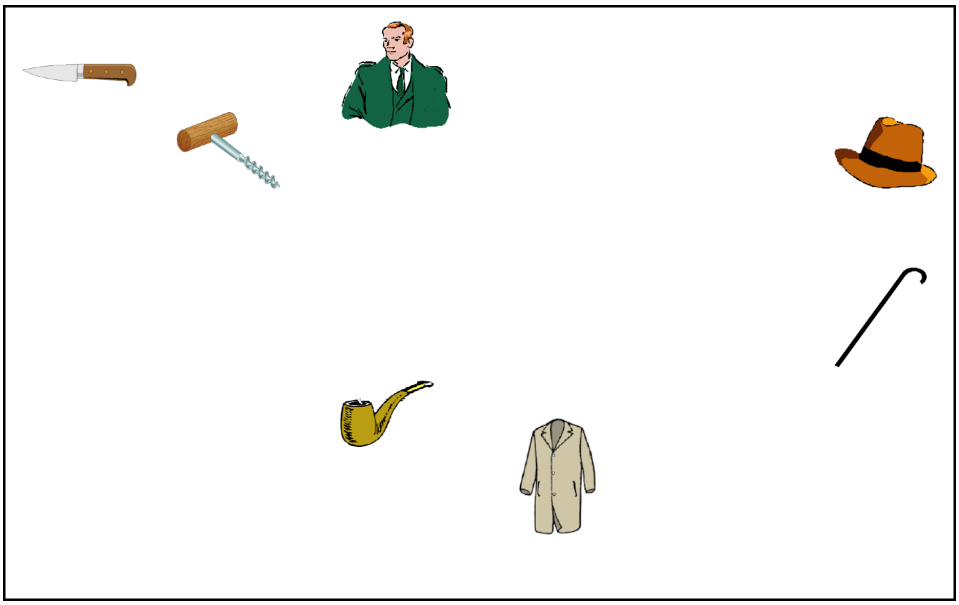
Indices as part of memory Instead of an independent capacity, spatial indices may be connected to or even form part of the representation of objects in memory. Research on the capacity of the visual short-time memory has determined the number of items that can be retained correctly to be approximately four (Luck & Vogel, 1997; Cowan, 2001), with performance decreasing when set size (i.e., items to be remembered) increased. Drawing on results regarding recall of verbal material presented in a serial fashion, we would expect to find serial position effects (Atkinson & Shiffrin, 1968, Section 2.2.3.1): If the target appears late in the trial, we should observe reliable language-mediated fixations increasing over the last 3-4 positions of presentation sequence (recency effect). If the target appears very early in the trial, usually the very first one or two positions, language-mediated eye movements should again be high (primacy effect).

4.1.1. Method

4.1.1.1. Participants

Thirty-six students from Saarland University were each paid 5 Euro for taking part in the experiment. They were all native speakers of German. Age ranged from 18 to 47 with a mean of 26.8. Fifteen participants were male, all were right-handed.

4.1.1.2. Materials and Design



	sentence	target object	comparison object
version 1	Der Mann raucht vermutlich die Pfeife *The man smokes probably the pipe*	pipe	knife
version 2	Der Mann schärft vermutlich das Messer *The man sharpens probably the knife*	knife	pipe

Figure 4.1.: Example item

A set of 30 experimental items was constructed (see Appendix A.2 for a list of all experimental items). An item consisted of a display and two sentences (Figure 4.1). In each display, there were six objects and a character. Those where randomly distributed in an invisible grid with 18 cells. The sentences contained two different restrictive verbs, each of which selected for only one of the objects on display as the relevant role filler. The two sentences were employed in two counter-balancing versions, such that the same visual object functioned as a *target* in one version and as a *comparitor* in the other version. The

comparitor object served as a baseline in the analysis. The other objects were distractors. Our main goal in choosing target, comparitor and distractor objects for an item was to construct a reasonably plausible scene. In particular, all objects were likely to be the object of an action performed by the character. As a trade-off, some objects exhibited an initial phonological overlap of one segment with each other or with one of the target objects.

The order in which objects appeared on the screen was manipulated. While the first entity to appear was always the character, the target object could appear in any of the other six temporal positions in the sequence of appearing objects, resulting in six conditions. The comparitor was in a fixed temporal position in every condition: for `pos1` (target was first object in sequence), the comparitor was in second position, for `pos2` in first, for `pos3` in fourth, for `pos4` in third, for `pos5` in sixth and for `pos6` in fifth. This way the exact same display with the same sequence could be used in `pos1` for one sentence and in `pos2` in the counterbalancing version.

Thirty filler items were constructed. Half of them also contained a restrictive verb, but in contrast to the experimental items, there were two to three objects in the display which could function as role fillers. The other half had unrestricted verbs, such that each visual object was a candidate for a role filler. In addition, three practice trials were constructed.

Twelve lists were created which contained each experimental item in only one condition and in one of the two counterbalancing versions using the latin square technique. Although verbs were always used for two different items, each participant was exposed to every verb only once, because the counterbalancing versions used a different verb. Lists were randomized individually with the constraint that there were at most two experimental items allowed in sequence.

4.1.1.3. Procedure

An SR Research EyeLink II head-mounted eye tracker with a sampling rate of 250 Hz monitored participants' eye movements. Pictures were presented on a 24" color monitor at a resolution of 1920x1200 pixels, sentences were played over loudspeakers. Participants' head movements were unrestricted and viewing was binocular, although only the dominant eye of each participant was tracked.

At the beginning of the experiment, the experimenter conducted the Miles test (Miles, 1930) to identify the participant's dominant eye. The participant then read the instructions on the screen with the experimenter answering comprehension questions. The experimenter adjusted the eye tracker and performed a 9-point calibration procedure. The experiment

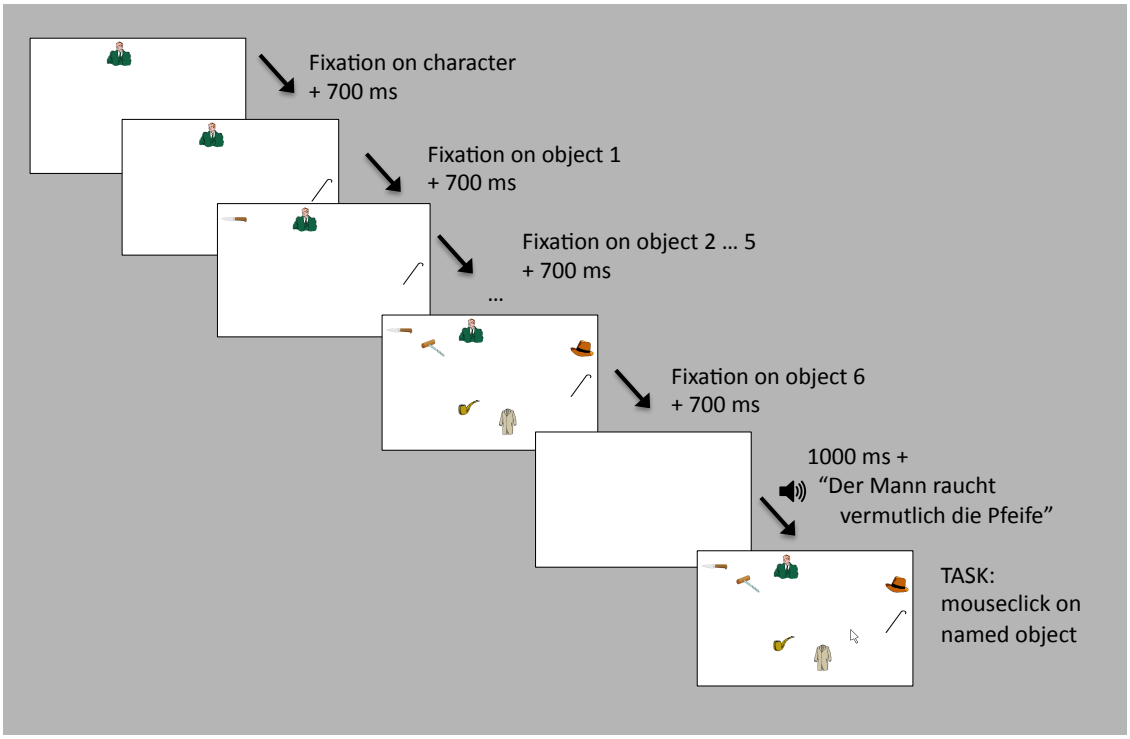


Figure 4.2.: Procedure of Experiment S1

started with the practice phase after which the participant was again encouraged to pose remaining questions. There was one break in the middle of the experiment and additional breaks if required by the participant or if the eye tracker needed recalibration.

The procedure of a single trial is sketched in Figure 4.2. Each trial started with a fixation marker on a random position in the inner part of the screen to validate the calibration of the eye tracker[1]. Next, the experimental display appeared only showing the character. Once the participant fixated it, the first object was triggered to become visible after 700 ms with the character remaining visible. As soon as the participant fixated that new object, the next one was triggered and so on, until all objects were visible on the screen. Another 700 ms later, all objects disappeared again, leaving the screen blank, and the sentence was played back after a delay of 1000 ms. After the end of the sentence all objects appeared again on the screen and participants had to click on the object mentioned in the sentence as fast as possible. One trial lasted approximately 12000 ms, depending on how quickly the participant fixated the new objects and how fast she clicked on the mentioned object. Participants were instructed to fixate newly appearing objects as fast

[1]The fixation marker was not in the center of the screen intentionally: In prior experiments, we found that in blank screen studies some participants tend to fixate the center of the screen and we suspect the centrally appearing fixation marker to encourage this behavior

as possible in the first phase of the trial and not to look back to the other objects. The experiment lasted approximately 30 min.

4.1.2. Predictions

The central prediction for this experiment is that participants inspect the region formerly occupied by the target object in more trials than the region formerly occupied by the comparitor object while processing the restrictive verb and the referential object noun phrase. This *target advantage* is the main diagnostic in this study and factors out effects that are purely induced by visual saliency or language-independent memory effects. More specifically, we expect a target advantage during verb and adverb, if participants are indeed able to anticipate the missing role filler in the situation where seven scene entities appeared on the screen. Correspondingly, a target advantage is expected during the processing of the second noun phrase, if participants are able to establish reference to the named object in this situation where seven scene entities appeared and disappeared again.

The hypotheses concerning pointer assignment described above provide us with more fine grained predictions with regard to the temporal position. The top-down assignment hypothesis predicts no effect of temporal position, but only a generally weak target advantage since in a considerable number of trials the target will not be indexed at all. The bottom-up assignment hypothesis, by contrast, predicts the following effect of temporal position: Object regions are only expected to be refixated if the object was in one of the last five temporal positions (`pos2-pos6`), but not if the object was in the first position (`pos1`), right after the character. A language induced target advantage is therefore predicted in these positions only, resulting in an interaction between position in sequence and type of object. According to the hypothesis that indices form part of memory, we predict a recency effect and possibly also a primacy effect. The primacy effect should be reflected by an enhanced target advantage for the first one or two positions in comparison with the "middle" position `pos3`[2]. The recency effect is expected to produce an increasing target advantage over the last three positions (`pos4-pos6`) in comparison to the middle position.

[2] Although `pos3` does not form the middle of the list, we refer to it as middle position, because it is the one not expected to be affected by primacy or recency, which have different scopes with recency extending over more positions in a sequence than primacy

4.1.3. Results

4.1.3.1. Method

Preprocessing The fixation data provided by the eye-tracking software includes spatial coordinates and information of when a fixation started and when it ended. The first step was thus to relate these information to the experimental stimulus. Spatial coordinates were automatically associated with one of the object regions *target*, *comparitor*, *character* and *background* using color-coded templates. The templates were created using the original item display and overlaying the object positions with colored squares that exceeded the original pictures slightly (300 x 300 pixels) to allow for some inaccuracies of tracker and fixations. Subsequent fixations on the same object were pooled into *inspections* (ins). In the next step, the inspections were temporally related to the speech stream. Since the individual words were of different lengths, each audio file was annotated for the exact onsets of NP1, verb, adverb and NP2 and additionally for the offset of NP2.

For the time course analysis, the inspection data was aligned to verb onset and NP2 onset respectively and associated with 250 ms lasting slots labeled by the end point of the slot – the time slot 250 thus contained the data from 0 to 250. An inspection was counted for a slot, if there was an overlap: it could either start within the slot, or continue from one of the previous slots.

For the inferential analyses, inspections were associated with one of the time windows VerbEnd, Verb and Np2. VerbEnd started 200 ms after verb onset and lasted until the end of the trial. Verb started 200 ms after verb onset and ended 200 ms after Np2 onset. Np2 started 200 ms after Np2 onset and ended 200 ms after Np2 offset. An inspection was counted for a time window only if it *started* within this window, not if it continued, as opposed to the time course analysis. This method was used in order to only analyze shifts of attention induced by the linguistic stimulus and to reduce the impact of random fixations independent of the spoken sentence. While VerbEnd included all inspections possibly due to the linguistic information concerning the object, Verb included only anticipatory looks, that is, all looks that follow the selective information of the verb until the next piece of information was processed. In the Np2 window, the information of the second NP was expected to drive referential looks in addition to a possible lasting influence of the verb information.

Ideally, we would further divide the Verb time window into verb and adverb (see e.g. Knoeferle et al., 2005; Kamide, Altmann & Haywood, 2003). As will become apparent in the next section, however, the overall low number of trials with inspections of target or

comparitor did not allow for such a fine-grained analysis. Also, more recent research within the Visual World Paradigm (Altmann, 2011) has suggested that the time to program and execute an eye movement might in fact fall below 200 ms, which contrasts with earlier findings (Allopenna et al., 1998). We decided for the 200 ms time lag, because we expected more noise in this early period which might obscure an effect[3].

Analyses Inferential analyses were conducted using multilevel logistic regression (mixed-effect models with a logit link function from the lme4 package in R; Bates, 2008) with the two fixed factors POS (`pos1, pos2,pos3, pos4, pos5, pos6`) for the position of the object in the presentation sequence and OBJ (`target, comparitor`) for the object being inspected. Random intercepts and slopes were included for participants and items. We report likelihood-ratio tests (Chi-Square test) that assess the contribution of fixed factors and interactions through model reduction in addition to model summaries. For the sake of conciseness, only the summaries of the full model including interactions are reported in the main text. For the reduced models assessing main effects, please refer the Appendix B. The dependent variable in the models, *ins*, is defined as 1 if a new inspection of the particular object was started in the relevant time window and 0 otherwise.

The full model described above, was of the following form in R syntax[4]:

- $ins \sim 1 + OBJ + POS + OBJ : POS + (1 + OBJ + POS \mid participant) + (1 + OBJ + POS \mid item)$

Due to an insufficient amount of data, this full model did not always converge. Whenever this was the case, the random slope terms were excluded one by one, until the model converged. If there were multiple options for the exclusion, the model with the best fit (i.e., with the highest log-likelihood) was used. If a random slope term showed full correlation with the intercept or another slope term, it was excluded as well. This procedure was performed individually for each time window where the full model did not converge. The actual model formula is reported along with the model summaries.

In all our mixed-effect models, the condition `pos3:comparitor` functions as the *reference condition*, so that all other conditions are compared to this one. Therefore all coefficients are to be interpreted relative to this condition. So, for example, if the coefficient of the interaction `pos1:target` is positive, this implies there were more fixations on the target in `pos1` compared to fixations to the comparitor in `pos3` after correcting for main

[3] In experiment S3, we diverge from this decision, because the experiment provided us with more data points and less noise

[4] In the following, we will abbreviate this form to $ins \sim OBJ * POS + (OBJ + POS \mid participant) + (OBJ + POS \mid item)$

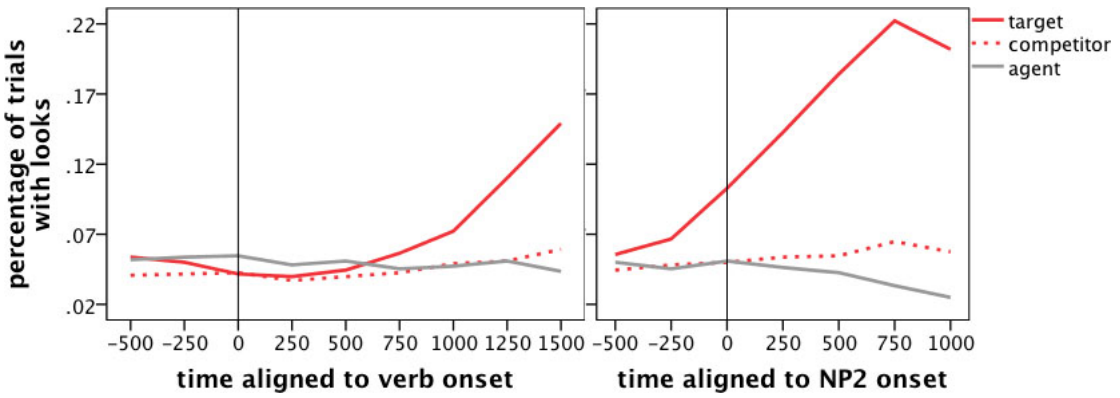


Figure 4.3.: Time course graph for character, target and comparitor objects aligned to verb onset and NP2 onset

effects. This specific condition was chosen, because it is expected to be different from other conditions according to the bottom-up index assignment hypothesis, as well as in the case of the emergence of serial position effects. According to bottom-up index assignment, where only the 5 objects that appeared last are expected to have an index, `pos3` should be favored in receiving an index in comparison to `pos1`. Serial position effects, on the other hand, disfavor `pos3`, since it is neither affected by the primacy effect nor by the recency effect.

4.1.3.2. Time Course Analysis

We first consider the time course graphs, shown in Fig 4.3. On the left-hand side, the inspection proportions are plotted relative to verb onset. We can see here that before verb onset, participants inspected character, target object and comparitor object nearly equally often. About 250 ms after verb onset, however, looks to the target and the comparitor object start to diverge with the target line rising slowly and then more rapidly in the following 1250 ms. On the right-hand side, inspection proportions are plotted relative to NP2 onset. We observe here that the target line starts to rise well before the NP2 onset and continues to do so in the following 750 ms.

These patterns suggest that participants were indeed fixating on the former location of the target object more than on the former location of the comparitor object. Since target inspections exceeded comparitor inspections already before the onset of NP2, we further note that participants were anticipating the target object based on verb restrictions. Somewhat surprisingly, we do not observe increased looks to the character before verb onset, but a rather flat line. We would expect referential looks to the character at this point, because it was just mentioned. Possibly, this is due to the task focusing on the

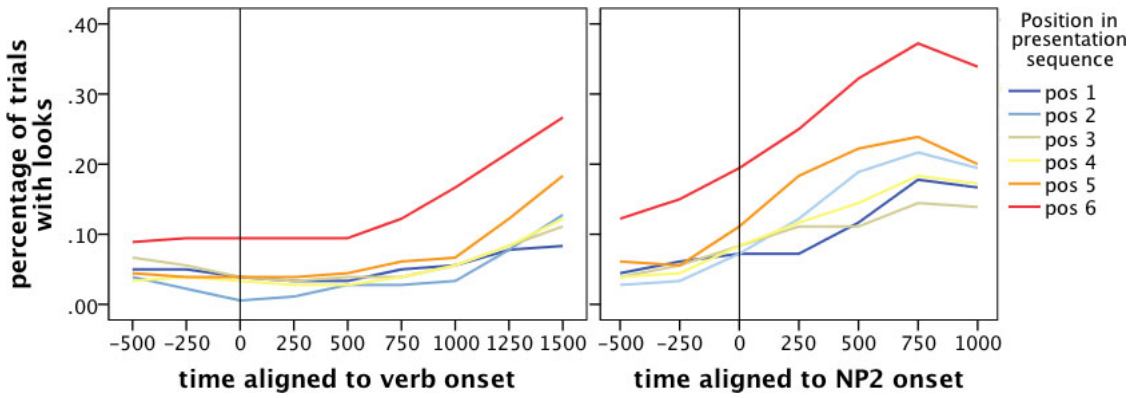


Figure 4.4.: Time course graph for target object only in all conditions aligned to verb onset and NP2 onset

object only and the character being highly predictable even before the sentence starts.

In Figure 4.4, proportions of trials with inspections of the target object only are plotted for the different POS levels (i.e., individual temporal positions in the presentation sequence), both aligned to verb onset and aligned to NP2 onset. Clearly, the target object was inspected in more trials, if it was in `pos6` than in any other position, even before the onset of the verb. This bias is thus not purely elicited by the linguistic input, but presumably by the persisting salience of this object having been fixated last. Still, by the onset of NP2 the gap between position 6 and the other positions has increased, indicating a slightly steeper rise. The other positions do not differ substantially, one exception being `pos5`, which rises steeper than the others after NP2 onset. There is, however, an interesting difference regarding the order of the lines between verb onset and the end of the trial: At verb onset `pos6` is highest, `pos2` lowest and the rest indistinguishable in between while at the end of the trial `pos6` is still highest, followed by `pos5` and `pos2`, then `pos4` and `pos1`, and `pos3` lowest. This ordering emerges around 500 ms after NP2 onset. We will come back to this ordering later.

4.1.3.3. Fixation Data Analysis

For the VerbEnd region, there was a significant effect of OBJ ($\chi(1)$=29.91, p< .001) indicating language driven eye movements on a blank screen even if the number of objects exceeds working memory capacity. There was also a marginal effect of POS ($\chi(5)$=10.92, p=.05), due to significantly more looks in `pos6` compared to the baseline (`pos3`) which indicates a general tendency to look back at objects that appeared last, irrespective of their relevance to the sentence (see table B.2 in the appendix for the model summary). Most interestingly, however, there was a significant interaction between OBJ and POS

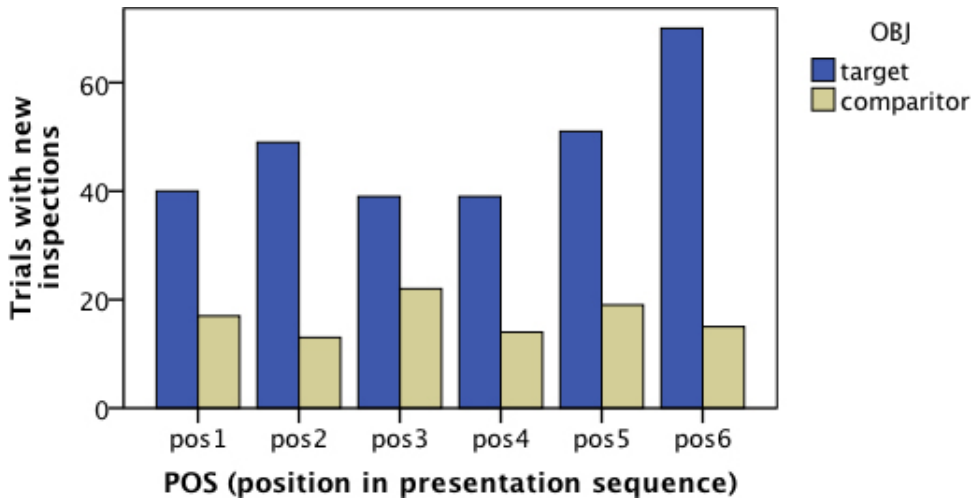


Figure 4.5.: Number of trials with newly started inspections in the time window VerbEnd

Predictor	Coefficiant	Std. Error	z value	p-value			
(Intercept)	-2.46	0.34	-7.27	3.66e-13	***	}	*baseline condition*
`target`	1.01	0.38	2.67	0.01	**	⎫	
`pos1`	-0.26	0.37	-0.70	0.49		⎪	*simple*
`pos2`	-0.59	0.39	-1.50	0.13		⎪	*effect*
`pos4`	-0.51	0.39	-1.33	0.18		⎪	*terms*
`pos5`	-0.30	0.37	-0.81	0.42		⎪	
`pos6`	-0.55	0.40	-1.39	0.16		⎭	
`target:pos1`	0.27	0.46	0.60	0.55		⎫	
`target:pos2`	0.92	0.47	1.95	0.05	.	⎪	
`target:pos4`	0.50	0.47	1.06	0.29		⎪	*interaction*
`target:pos5`	0.70	0.45	1.57	0.12		⎪	*terms*
`target:pos6`	1.50	0.46	3.25	0.00	**	⎭	

Table 4.1.: Model summary for VerbEnd time region (N = 2110; log-likelihood = -876.4)
Model: $ins \sim OBJ + POS + OBJ * POS + (1|subj) + (OBJ + POS|item)$

($\chi(5) = 11.33, p < .05$), indicating that language driven eye movements were influenced by the accessibility of the object in working memory and differed according to when the object was seen. The model summary (Table 4.1) indicates that the advantage of the target over the comparitor was significantly higher in `pos6` and marginally higher in `pos2` compared to the baseline. This pattern was not predicted by any of our hypotheses. The picture gets a little clearer, however, if we look at the total occurrence of new inspections in Figure 4.5. The target advantage is here observable in the difference between target and comparitor bars. If we consider only `pos2` - `pos6`, there is a tendency towards a U-shape pattern usually observed in free recall of verbal material, indicating a primacy effect for the second position, a recency effect increasing over the last two positions and a dip in the middle, at `pos3`, where the difference between target and comparitor bars becomes comparatively

Predictor	Coefficiant	Std. Error	z value	p-value			
(Intercept)	-3.80	0.48	-7.85	4.15e-15	***	}	*baseline condition*
`target`	0.64	0.52	1.24	0.22		⎫	
`pos1`	0.09	0.51	0.18	0.86		⎪	*simple*
`pos2`	-1.99	0.77	-2.58	0.01	**	⎪	*effect*
`pos4`	-0.89	0.71	-1.26	0.21		⎪	*terms*
`pos5`	-0.32	0.57	-0.57	0.57		⎪	
`pos6`	-0.44	0.54	-0.82	0.41		⎭	
`target:pos1`	-0.17	0.68	-0.25	0.80		⎫	
`target:pos2`	1.93	0.88	2.19	0.03	*	⎪	
`target:pos4`	0.97	0.83	1.17	0.24		⎪	*interaction*
`target:pos5`	0.76	0.70	1.09	0.28		⎪	*terms*
`target:pos6`	1.16	0.67	1.72	0.09	.	⎭	

Table 4.2.: Model summary for VERB time region (N = 2110; log-likelihood = -419.5)
Model: $ins \sim OBJ + POS + OBJ * POS + (1|subj) + (OBJ + POS|item)$

small. Surprisingly, the primacy effect seems not to emerge in `pos1`, where it would be expected. We will return to this point in the evaluation of the paradigm (Section 4.1.5).

For the VERB time window, we found a significant effect of OBJ indicating an anticipatory advantage of the target object over the comparitor object ($\chi(1)= 16.21$, $p< .001$). The main effect of POS was not significant($\chi(5)=4.24$). Although the interaction between POS and OBJ did not reach significance($\chi(5)=8.36$, $p=.13$), the model summary shows that the target advantage was significantly stronger for `pos2` compared to the baseline (Table 4.2) and marginally stronger for `pos6`. The higher coefficient of `pos2` also indicates a stronger influence of primacy than recency, which contrasts to the result in the bigger time window VERBEND.

For the NP2 time window, we also found a significant main effect of OBJ($\chi(1)=31.80$, $p< .001$) and an effect of POS ($\chi(5)=12.33$, $p< .05$) due to more looks in `pos6`. There was no significant interaction in this time window ($\chi(5)=6{,}09$, $p=.29$), the model summary, however, shows a significantly enhanced target advantage for `pos6` (Table 4.3), suggesting a recency effect for referential eye movements, and no primacy effect.

4.1.4. Discussion

This experiment set out to answer two questions: Firstly, we aimed to establish whether language can guide visual attention on a blank screen even if the number of previously inspected objects exceed the capacity of visual working memory. Secondly, we investigated whether the position of the target object within the presentation sequence influences the accessibility of this object for language driven eye movements. Regarding the first question,

Predictor	Coefficiant	Std. Error	z value	p-value			
(Intercept)	-3.23	0.39	-8.19	2.7e-16	***	}	*baseline condition*
`target`	1.22	0.45	2.68	0.01	**	⎫	
`pos1`	-0.52	0.55	-0.95	0.34		⎪	*simple*
`pos2`	0.03	0.49	0.06	0.96		⎪	*effect*
`pos4`	-0.15	0.50	-0.30	0.77		⎪	*terms*
`pos5`	-0.15	0.50	-0.30	0.76		⎪	
`pos6`	-0.34	0.53	-0.65	0.52		⎭	
`target:pos1`	0.29	0.63	0.46	0.64		⎫	
`target:pos2`	0.48	0.57	0.85	0.40		⎪	
`target:pos4`	0.13	0.58	0.23	0.82		⎪	*interaction*
`target:pos5`	0.69	0.58	1.19	0.23		⎪	*terms*
`target:pos6`	1.22	0.59	2.06	0.04	*	⎭	

Table 4.3.: Model summary for NP2 time region (N = 2110; log-likelihood = -674.9)
Model: $ins \sim OBJ + POS + OBJ * POS + (POS|subj) + (OBJ|item)$

we found confirming evidence: Time-course graphs as well as inferential analyses converge on the finding that the target object is inspected in more trials than the comparitor starting after verb onset. The second question is also answered positively: The interaction we found between OBJ and POS for the full time window VERBEND indicates that the degree to which the target object is inspected more than the comparitor depends on when the object was seen.

These results provide us with no support for the top-down index assignment hypothesis since it predicted no difference in the looking behavior depending on the temporal position of the object. The bottom-up index assignment hypothesis can also be rejected: It predicted no difference between `pos3` and `pos6`, but less target fixations in `pos1`. Instead, we found no significant difference between `pos3` and `pos1`, but significantly more target inspections in `pos6`. The pattern of our results is mainly consistent with the hypothesis of indices forming part of memory. In the full time window VERBEND, we find an indication of a serial position effect: The target advantage is significantly greater for `pos6` (recency) and marginally greater for `pos2` (primacy). The increasing target advantage over the last positions that we observed in the time course graph (Figure 4.4) and in the proportions of trials with fixations (Figure 4.5), however, was not confirmed by the inferential statistics.[5]

Interestingly, the pattern of eye movements with regard to the temporal position differs between referential eye movements in NP2 and anticipatory eye movements in VERB. The model summary for VERB suggests a primacy effect for `pos2`, while the magnitude of the `target:pos6` coefficient indicates that the recency effect is less effective on anticipatory

[5] In fact, we can see this trend in the model summaries by looking at the coefficients: Although not significantly different from zero, the coefficient of the target advantage in `pos5` is higher than the one for `pos4` and smaller than the one for `pos6`.

eye movements. In the time window NP2, on the other hand, the greater target advantage for `pos6` in the model summary can be interpreted as the indication of a recency effect, while no primacy effect can be found. Although this pattern could not be confirmed by a significant interaction through model reduction, it encourages an explanation which is based on the nature of the *representation* of the visual object, which is being used for language processing. Following Atkinson & Shiffrin (1968), the primacy effect is associated with the activation of a representation in long-term memory, while the recency effect is attributed to a representation in short-term memory (see section 2.2.3.1). Alternatively, under the hypothesis of a unitary memory store, the level of processing producing different kinds of representations can be the source of these effects (Craik & Lockhart, 1972): If the stimulus was processed on a surface level only, a shallow representation is built, which decays quickly. Deeper, semantic processing, which is only possible for the first one or two objects due to processing constraints, leads to a rich, conceptual representation, which is more stable than the former. If we adopt this interpretation, the observed pattern leads to the supposition that anticipatory eye movements rely on conceptual representations, while referential eye movements are due to shallow representations in memory. On the theoretical side, this idea is backed up by the conjecture that the restrictive verb addresses the affordance of an object as a possible argument – the question here is which of the objects on display can be smoked? This aspect requires deep semantic processing of the visual object *pipe* and is part of the conceptual representation of *pipe*. For referential eye movements, on the other hand, no deep semantic processing or activation of the conceptual properties seem to be necessary: The name of the object is activated the moment the visual object is perceived (Zelinsky & Murphy, 2000; Navarrete & Costa, 2005) and is part of the shallow representation. This shallow representation which may only consist of the name and the position of the object could then drive referential eye movements.

As a preliminary conclusion, we suggest that our results are best explained by an influence of primacy and recency suggesting a connection between visual pointers and representations in memory. Interestingly, the influence of primacy and recency effect seem to vary depending on the nature of eye movements. This effect, however, is only observable in the model summaries and could not be confirmed by a significant interaction using model reduction, and therefore demands further experimental investigation.

4.1.5. Evaluation of the Paradigm

Importantly, the results demonstrate that the paradigm developed for this experiment is able to show that language drives eye movements in a blank screen context, even when the number of objects exceeds the established visual working memory capacity. For the effects

concerning serial position and accessibility from memory, the results are less straightforward: The primacy effect here is atypical, since it is not observable in the very first position, but only in the second position. In fact, `pos1` is not even the first position, since the first scene entity to appear was the character. It is not entirely obvious, why the effect should then turn up at second position, but we will assume that characters and objects are treated separately and that the first object did not get the same amount of attention as the following ones, because the preceding character was more complex to process. If this were indeed true, a more canonical primacy effect is expected, if presentation started with the objects. The recency effect, on the other hand, is only detected on the very last position, while we would have expected it for the last three positions. Also, the very last position seems not to be an ideal testing position, since after this last object disappears participants could in principal keep their gaze on the same location as no other object appears. Both issues are addressed with a different ordering in experiment S2.

An additional concern with this experiment is the fact that objects appearing early also stayed longer on display. Although participants were instructed not to look back to old objects, it is possible that they still attracted part of their attention and were thus more accessible than expected – the recency effect would thus trade-off with the time objects were available for encoding. Another reason for the difficulty in finding canonical serial position effects might be the task: serial position effects have typically been found in a free recall task. In contrast to that task, where the next object recalled can be chosen spontaneously, in this experiment a fixation of the object is mandatory, since the participant has to perform a mouse click. To approximate the free recall task, a task that does not enforce eye movements might therefore to be preferred.

4.2. Experiment S2

Experiment S1 demonstrated that language can guide eye movements on a blank screen even when the number of objects in the preceding visual context clearly exceeded visual short term memory capacity as well as the number of visual indices available. Importantly, it also showed that the serial position of an object in the sequence of presentation influenced the availability of this object for language-mediated eye movements. While the general pattern suggested the emergence of primacy and recency effects, the evidence remained inconclusive. We observe several aspects of the experimental method that might have contributed to this which are addressed by the present experiment S2 in order to maximize the chance of detecting canonical primacy and recency effects. Specifically, the task in this experiment was changed to a picture-sentence verification task, which is more similar to the free recall task which typically elicits serial position effects. The expectation was that participants would spontaneously fixate the regions formerly occupied by the relevant picture if they remembered it or, more precisely, if the representation of the object in memory was sufficiently activated. In contrast to the task used in S1, this task was less restrictive and a fixation was not mandatory, because participants were not asked to click on the object. Also, the presentation mode was strictly serial in S2: the pictures appeared for a fixed time period and disappeared before the next picture appeared in order to prevent the early appearing objects of receiving more attention. Finally, the position of the character in the presentation sequence was shifted from the beginning of the sequence to the end. This modification aimed at increasing the chance to detect a primacy effect, which is usually restricted to the very beginning of a sequence.

4.2.1. Representation Structure in Memory

The results from experiment S1 suggest a more pronounced primacy effect for anticipatory eye movements and a recency effect for referential eye movements. This pattern encourages an account which is based on the nature of the *representation* of the visual object, which is being used for language processing. Traditionally, the primacy effect is associated with the activation of a representation in long-term memory, while the recency effect is commonly attributed to the more shallow representation in short-term memory (Atkinson & Shiffrin, 1968, see section 2.2.3.1). Alternatively, under a single store hypothesis, serial position effects can be explained with different levels of processing (Craik & Lockhart, 1972, see section 2.2.3.3). Both accounts suppose a rich representation containing information about affordances, associations and other semantic properties in addition to surface level information to be built up for items seen early: either because the representation is

part of long-term memory which holds semantic information, or because the item was processed semantically. Items seen late, on the other hand, only evoke a relatively shallow representation based on surface features, because they are part of short term memory or because they only received shallow processing. Integrating these explanations with a conceptualization of memory as activation patterns, we arrive at conceiving different memory representations for the object *pipe* depending on its serial position which we schematically depict in Figure 4.6: If the pipe object was seen early (top half), its representation includes not only the phonological code "pipe", its former location on the screen and its visual form, but also affordances (*smokable*), semantically related features (*smoke*, *smells*), and associations (e.g., *Sherlock Holmes*). All of these features are interconnected within the representation of "pipe", resulting in a relatively stable representation by the constant spread of activation among them. When, on the other hand, the same object was seen late (lower half), the representation built up consists only of perceptual surface features like location, and visual form and the phonological form, which is activated automatically the moment the visual object is perceived. Presumably, these features exhibit a higher level of activation, as there was not much interfering material in the same modalities (Nairne, 2002), thus making the representation similarly accessible as the richer one built for early items. However, this representation is expected to be less stable as there are fewer features activated which may be overwritten easily by new, incoming material.

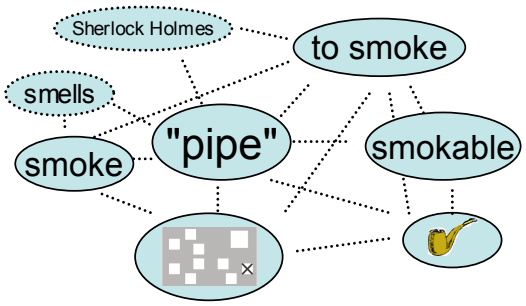
Assuming these shallow or rich representations in memory, let us now consider what feature patterns seem minimally necessary to accommodate anticipatory and referential eye movements. For referential eye movements, we suppose the name and the location to be the important features: The perception of the spoken word should reactivate the name feature which then spreads activation to the former location encouraging an eye movement to this location. Those features form part of both the rich conceptual representation, and the more shallow representation. In the shallow representation, however, these features are expected to exhibit a higher level of activation, therefore shallow representations are better candidates for referential eye movements. For anticipatory eye movements, there are two possibilities: The restrictive verb, e.g. "smoke", could address the affordance of an object as a possible argument of the verb. In this view, the critical question is which of the objects on display has the property of being smokable. This feature requires deep semantic processing of the visual object *pipe* and is part of the rich representation, only. Alternatively, anticipatory eye movements could be driven mainly by *lexical* expectations. This hypothesis is supported by findings from the reading literature, where a specific word was read faster or even skipped if it was predictable in the context, or if the transitional probability between the prior word and the current word was high (Frisson, Rayner & Pickering, 2005; McDonald & Shillcock, 2003). In this view, the processing of the verb

Early position

stimulus was processed on multiple levels

rich representation

activated features include: phonological form, location, visual form, associations, affordances...



Late position

stimulus was processed on visual and phonemic level (naming)

shallow representation

activated features include: phonological form, location, visual form

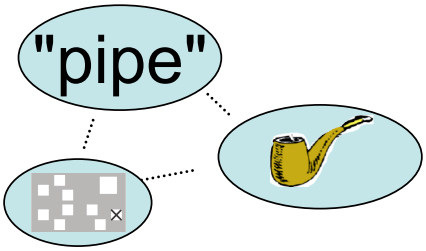


Figure 4.6.: Schematic depiction of different memory representations for *pipe* object at the end of the presentation sequence depending on its serial position during presentation. Size and intensity of features correspond to level of activation.

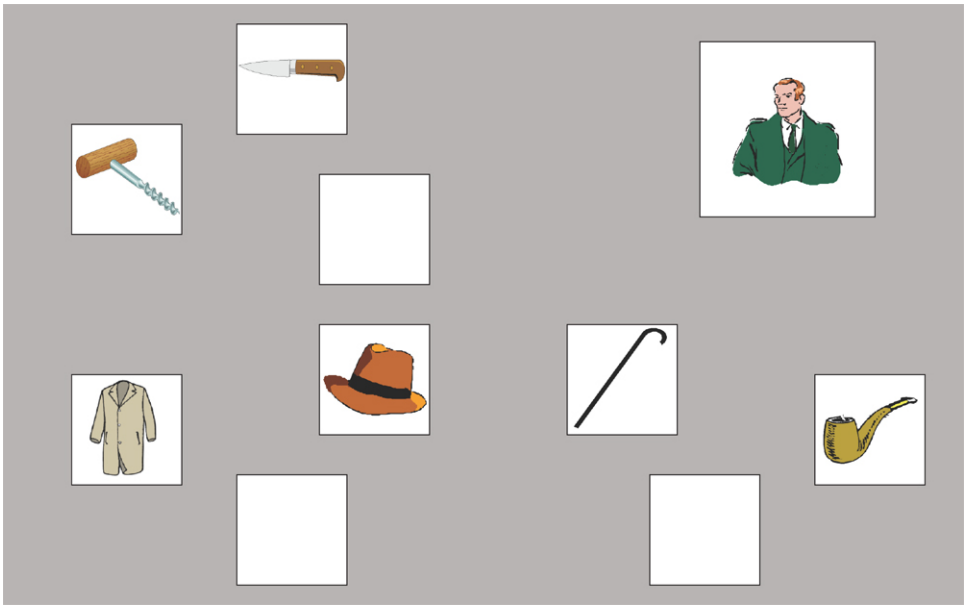
would lead to a prediction of the next word. This prediction could then be matched to the phonological code, which is part of both shallow, and rich representations. While the first account predicts a primacy effect on anticipatory eye movements, the second account is compatible with no such effect. Detecting the exclusive emergence of a primacy effect in anticipatory eye movements and an additional recency effect in referential eye movements would therefore indicate that indeed these two kinds of language-driven eye movements are distinct processes and rely on different representations: Anticipatory eye movements rest on the *conceptual* expectation of what is to be mentioned next, while referential eye movements are driven by the direct match of the processed noun phrase with the name of an object.

4.2.2. Method

4.2.2.1. Participants

Thirty-seven students from Saarland University were each paid 7,50 Euro for taking part in the experiment. They were all native speakers of German. Age ranged from 18 to 40 with a mean of 23.9. Three participants were male, five were left-handed.

4.2.2.2. Materials



	sentence	target object	comparison object
version 1	Der Mann raucht vermutlich die Pfeife *The man smokes probably the pipe*	pipe	knife
version 2	Der Mann schärft vermutlich das Messer *The man sharpens probably the knife*	knife	pipe

Figure 4.7.: Example item

The materials were the same as in S1 with the following exceptions: The background of the display was not blank, but was colored grey with 10 white boxes measuring 250 x 250 pixels on it to encourage accurate fixations (Figure 4.7). The objects were distributed in the display such that in one quadrant there was only the person, in one quadrant there were the target and one distractor, in one quadrant there were the comparitor and one distractor and finally in the last quadrant there were the two remaining distractors. This was done in order to allow for a quadrant based analysis of eye movements, where person, target and comparitor were always in different quadrants.[6]

The order in which objects appeared on the screen was manipulated similar to S1 except that the target object could appear in any of the first six positions in the sequence of appearing objects while the last image to appear was always the character. This modification was done in order to elicit a more canonical primacy effect for the first position and to have a cleaner recency effect, since participants could not keep their gaze

[6]We do not report this analysis.

on the last object to appear.

The existing filler items were slightly changed and an additional 20 filler items were constructed resulting in 50 fillers in total. The objective for these additional fillers was to balance answers between yes and no for the picture-sentence verification task. In a no-filler, one of the noun phrases would refer to an object or a person which was not present in the display before. As all experimental items were positive examples, there had to be at least as many negative examples. To avoid the occurance of a restrictive verb to be a clear predictor of a positive answer, ten of the new no-filler items contained restrictive verbs with exactly one possible role filler object on the screen, similar to the experimental items. This object, however, was not mentioned in the sentence. Also, ten of the fillers had the object mentioned in the sentence also present on the display - half of them with restrictive verbs but more than one possible role filler and the other half with nonrestrictive verbs. Twenty fillers had a simple yes/no question associated with them to make sure participants payed attention to the whole sentence and not only the noun phrases.

Twelve lists were created which contained each experimental item in only one condition and in one of the two counterbalancing versions using a latin square technique. The lists were randomized individually for each participant with the restriction that there had to be at least one filler between two experimental items.

4.2.2.3. Procedure

The procedure was similar to S1 except for the following differences: A trial started with the background grid. After 400 ms the first object appeared in one of the squares and remained for 1200 ms. Then the object disappeared and after 400 ms the next object appeared in another square and so on. The last object to appear was the character. 400 ms after it disappeared, the sentence was played back to the participants. Their task was to decide, whether both, person and object, that were mentioned in the sentence had also been present as pictures on the display before. They then had to indicate their answer as rapidly as possible by pressing one of two buttons for “yes” and “no” on a button box. Participants used the index finger of their dominant hand for a “yes” response. Reaction times were measured. After the first button response, in 25% of the trials there was a comprehension question appearing on the screen which participants also answered using the button box. The trial ended automatically after the response. The experiment lasted approximately 45 min.

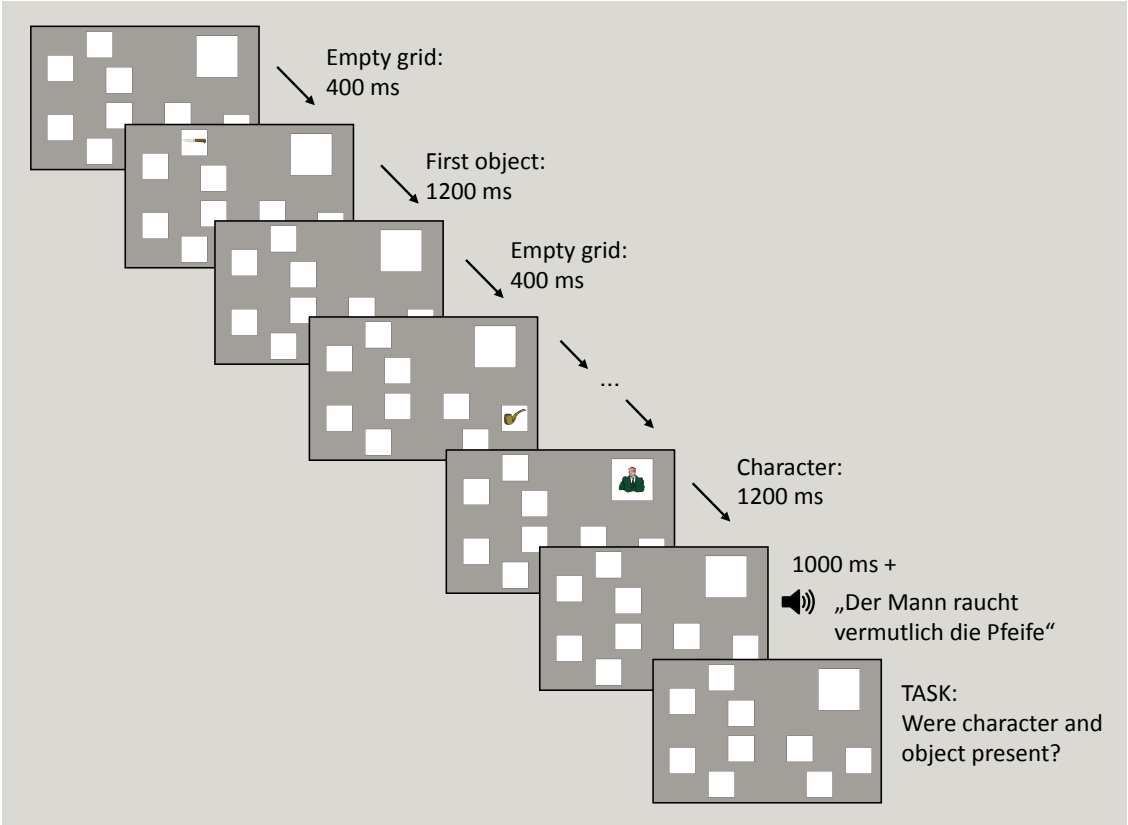


Figure 4.8.: Procedure of Experiment S2

4.2.3. Predictions

Based on our hypothesis that anticipatory eye movements exploit rich, conceptual memory representations while referential eye movements may rely on more shallow lexical representations (the *2-representations hypothesis*), we derive the following predictions: We expect to find an interaction of position (POS) and object type (OBJ) in each time window. While in the purely anticipatory time window VERB this interaction is expected to originate from a primacy effect and thus an enhanced target advantage in `pos1-2`, in the referential time window NP2, we expect a rising target advantage over `pos4-6`. In the combined time window VERBEND, both effects should be visible.

If, on the other hand, anticipatory eye movements are based on lexical expectations and therefore rely on the same representations as referential eye movements, the *lexical expectation hypothesis*, we should find the same pattern in all three time regions: a recency effect as evidenced by an increasing target advantage over the last three positions and possibly also a primacy effect.

4.2.4. Results

One subject was excluded due to a high error rate. Additionally, there was data loss due to equipment failure for two participants. The correctly recorded data for these participants was kept in the analysis.

4.2.4.1. Time Course Analysis

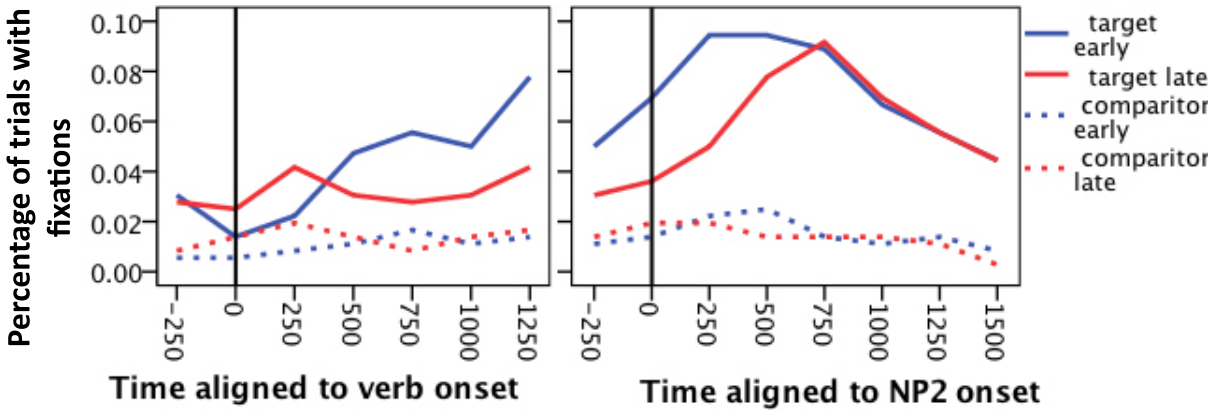


Figure 4.9.: Time course graph for target and comparitor in early(`pos1, pos2`) vs late(`pos5, pos6`) conditions aligned to verb onset and NP2 onset

In Figure 4.9 proportion of looks to target and comparitor object are plotted relative

to verb and NP2 onset for early (pos1,pos2) vs late (pos5,pos6) position of object in the presentation sequence. We can see here that after the onset of the verb, looks to the target object increase rapidly, if the target appeared early, while looks to the comparitor stay relatively stable. If it appeared late in the sequence, on the other hand, target and comparitor fixations stay nearly parallel up to about 1000 ms. On the right-hand side, the influence of the second NP becomes apparent: for the target appearing late, inspections of the target increase after NP2 onset untill 750 ms later. For the early targets, the peak is already attained shortly after NP2 onset and looks start to decline again. While we might expect this decline to start earlier, it can be explained by sustained inspections. Overall, this pattern is at least compatible with the prediction that a primacy effect should be observed after verb onset and a recency effect after NP2 onset.

4.2.4.2. Fixation Data Analysis

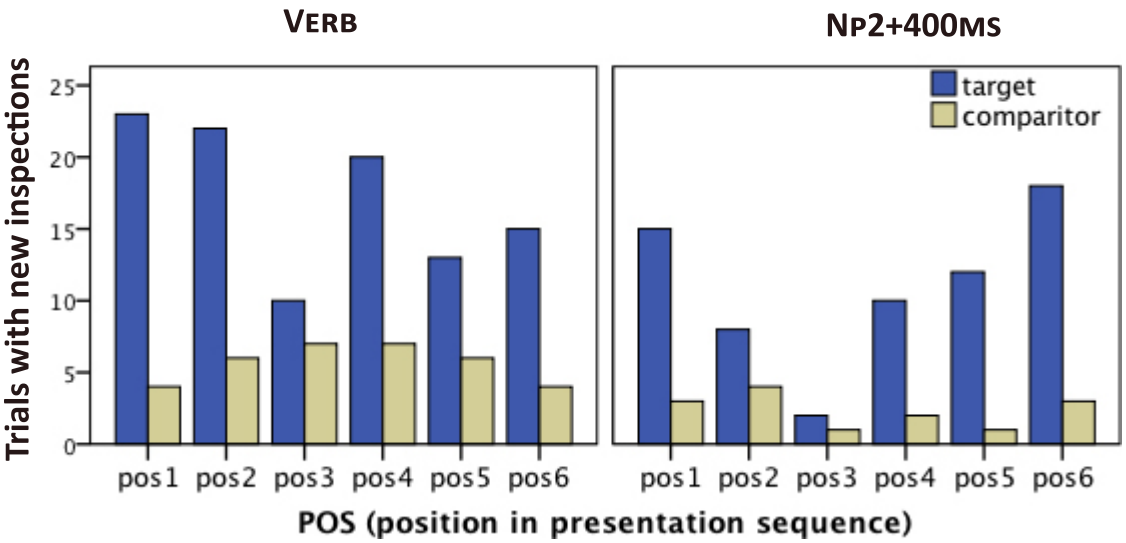


Figure 4.10.: Number of trials with newly started inspections to target and comparitor object in the time regions VERB and NP2+400MS

We used the same time regions as in experiment S1 (see 4.1.3.1 for the details). For the VERBEND region, there was a significant effect of OBJ ($\chi(1)$=21.32, p< .001) indicating language driven eye movements while the interaction between OBJ and POS was not significant.[7] The model summary (Table 4.4), however, indicates that the advantage of the target over the comparitor was significantly higher in pos1, suggesting a primacy effect, as well as in pos6 suggesting a recency effect.

[7]The interaction was marginal ($\chi(2)$=5.37, p= .06)for a reduced data set including just the data for pos1,pos6 and the baseline pos3 due to a stronger target advantage in pos1 and pos6.

Predictor	Coefficiant	Std. Error	z value	p-value			
(Intercept)	-3.52	0.42	-8.48	<2e-16	***	}	*baseline condition*
`target`	0.35	0.49	0.71	0.48		⎫	
`pos1`	-0.32	0.55	-0.58	0.56		⎪	*simple*
`pos2`	-0.16	0.53	-0.30	0.77		⎪	*effect*
`pos4`	-0.15	0.54	-0.28	0.78		⎪	*terms*
`pos5`	-0.10	0.55	-0.19	0.85		⎪	
`pos6`	-0.68	0.58	-1.16	0.25		⎭	
`target:pos1`	1.45	0.66	2.20	0.03	*	⎫	
`target:pos2`	1.00	0.66	1.52	0.13		⎪	
`target:pos4`	1.05	0.66	1.59	0.11		⎪	*interaction*
`target:pos5`	0.95	0.68	1.41	0.16		⎪	*terms*
`target:pos6`	1.60	0.70	2.30	0.02	*	⎭	

Table 4.4.: Model summary for VerbEnd time region (N = 2136; log-likelihood = -552.6)
Model: $ins \sim OBJ + POS + OBJ * POS + (OBJ + POS|subj) + (1|item)$

Predictor	Coefficiant	Std. Error	z value	p-value			
(Intercept)	-3.59	0.43	-8.30	<2e-16	***	}	*baseline condition*
`target`	0.11	0.55	0.19	0.85		⎫	
`pos1`	-1.04	0.70	-1.49	0.14		⎪	*simple*
`pos2`	-1.22	0.72	-1.68	0.09	.	⎪	*effect*
`pos4`	-0.76	0.66	-1.16	0.25		⎪	*terms*
`pos5`	-0.07	0.60	-0.12	0.91		⎪	
`pos6`	-1.30	0.79	-1.65	0.10	.	⎭	
`target:pos1`	1.93	0.81	2.37	0.02	*	⎫	
`target:pos2`	2.00	0.81	2.48	0.01	*	⎪	
`target:pos4`	1.42	0.76	1.87	0.06	.	⎪	*interaction*
`target:pos5`	0.48	0.76	0.64	0.53		⎪	*terms*
`target:pos6`	1.54	0.88	1.75	0.08	.	⎭	

Table 4.5.: Model summary for Verb time region (N = 2136; log-likelihood = -427)
Model: $ins \sim OBJ + POS + OBJ * POS + (POS|subj) + (OBJ + POS|item)$

Predictor	Coefficiant	Std. Error	z value	p-value			
(Intercept)	-5.74	1.03	-5.55	2.86e-08	***	}	*baseline condition*
`target`	-0.45	1.40	-0.32	0.75		⎫	
`pos1`	0.13	1.26	0.10	0.92		⎪	*simple*
`pos2`	1.53	1.16	1.32	0.19		⎪	*effect*
`pos4`	0.60	1.37	0.44	0.66		⎪	*terms*
`pos5`	-0.32	1.58	-0.20	0.84		⎪	
`pos6`	0.76	1.21	0.63	0.53		⎭	
`target:pos1`	2.34	1.57	1.49	0.14		⎫	
`target:pos2`	0.93	1.52	0.61	0.54		⎪	
`target:pos4`	2.73	1.68	1.63	0.10		⎪	*interaction*
`target:pos5`	3.47	1.85	1.87	0.06	.	⎪	*terms*
`target:pos6`	2.16	1.54	1.40	0.16		⎭	

Table 4.6.: Model summary for NP2400 time region (N = 2136; log-likelihood = -275.5)
Model: $ins \sim OBJ + POS + OBJ * POS + (OBJ + POS|subj) + (1|item)$

In the VERB time window, there was a significant effect of OBJ ($\chi(1) = 23.45, p < .001$) indicating anticipatory eye movements in addition to a marginal interaction ($\chi(5) = 9.74, p = 0.08$). In the model summary in Table 4.5 we see that the interaction is caused by a significantly enhanced target advantage in `pos1` and `pos2` indicating a primacy effect in addition to marginally enhanced advantages for `pos4` and `pos6`.

For the NP2 region, there was a significant effect of OBJ ($\chi(1) = 5.68, p < .05$) indicating referential eye movements, but no interaction between POS and OBJ. Since the NP2 region was on average 400 ms shorter than the VERB region, an additional analysis was conducted for a prolonged time region NP2+400 that began 200 ms after the onset of the second NP and lasted until 600 ms after the offset. For this prolonged region, the interaction was also not significant, but the model summary in Table 4.6 shows again a tendency towards a recency effect: The coefficients of the target advantage are positive and comparatively high for `pos4`, `pos5` and `pos6`. Due to the high error terms, there is only one marginally enhanced target advantage for `pos5`.

4.2.4.3. RTs and accuracy

This experiment allowed us to measure RTs on the decision task, starting at the onset of the second NP until button press. RTs more than 2 standard deviations away from the individual participants mean were removed as outliers. The mixed effect analysis revealed a marginal effect of POS ($\chi(5)$=10,43, p= .06) due to faster responses in `pos6` compared to `pos3` (Table 4.7). Surprisingly, the other conditions did not differ from the baseline - in contrast to the fixation data, we thus found only a recency effect.

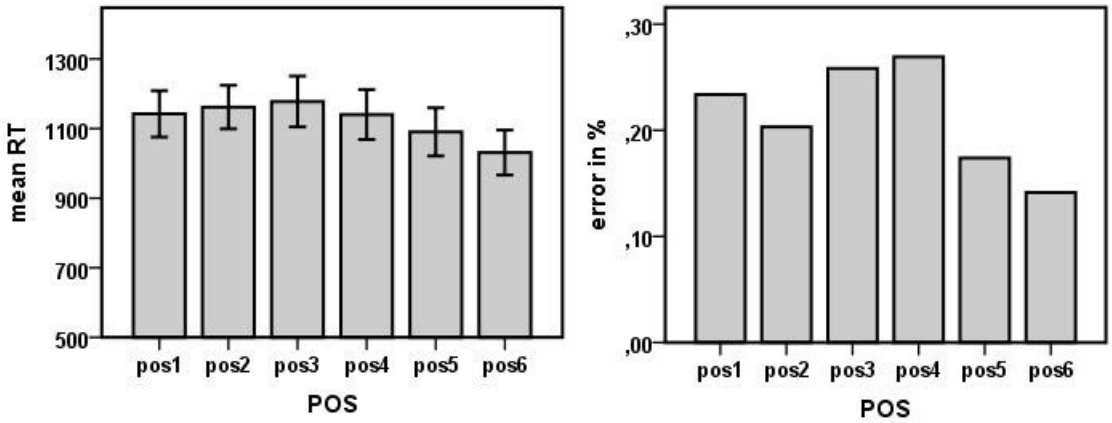


Figure 4.11.: Averaged RTs and error rates

Predictor	Coefficient	Std. Error	t value	MCMCmean	pMCMC	p-value	
(Intercept)	1191.41	61.39	19.41	1191.84	0.00	0.00	***
`pos1`	-4.34	45.63	-0.10	-5.17	0.92	0.92	
`pos2`	-24.24	45.34	-0.54	-25.11	0.59	0.59	
`pos4`	14.13	45.42	0.31	13.27	0.77	0.75	
`pos5`	-5.73	45.69	-0.13	-6.64	0.88	0.90	
`pos6`	-111.51	45.14	-2.47	-112.14	0.01	0.01	*

Table 4.7.: Model summary for Reaction Times (N = 1018, log-likelihood = -7613)
Model: $RT \sim POS + (1|subj) + (1|item)$

For the accuracy data, the pattern was very similar: there was a significant effect of POS ($\chi(5)$=16,17, p< .01) due to significantly fewer errors in `pos6` and marginally fewer errors in `pos5` indicating a recency effect, only (Table 4.8). This pattern supports the two-representation hypothesis: consider that the task was to verify whether character and object *named* in the sentence were also present in the scene, the emphasis was thus on the phonological level and not on the deeper semantic representation.

4.2.5. Discussion

This experiment set out to confirm the trend towards serial position effects, as suggested by experiment S1 and to investigate whether primacy and recency have a different impact on anticipatory and referential eye movements. The presence of serial position effects is confirmed by the above described results: While the accuracy data show clear evidence for a recency effect, the interaction in VERB is driven by a significant primacy effect in addition to a marginal recency effect.

Predictor	Coefficient	Std. Error	z value	p-value	
(Intercept)	-1.38	0.26	-5.38	7.31e-08	***
pos1	-0.07	0.26	-0.27	0.78	
pos2	-0.45	0.28	-1.63	0.10	
pos4	0.08	0.259	0.299	0.76	
pos5	-0.51	0.279	-1.82	0.07	.
pos6	-0.89	0.299	-2.979	0.00	**

Table 4.8.: Model summary for Accuracy (N = 1068, log-likelihood = -502.4)
Model: $error \sim POS + (1|subj) + (1|item)$

The second question, however, is still not answered conclusively. If anticipatory eye movements rely exclusively on rich representations containing conceptual information, we expected to see a primacy effect for the VERB region and no recency effect. While we did find a significant primacy effect for the first two positions, there was also a marginal recency effect in this time window. On the other hand, we did not find any significant serial position effect for referential eye movements in the time window NP2+400, where we predicted a recency effect. Only the model summary shows a greater target advantage for the last three positions, which does not reach significance. The strongest piece of evidence these results contain is the disappearance of the primacy effect in the referential time window which suggests a qualitative difference of the two kinds of eye movements. The strong version of the two-representations hypothesis, assuming that anticipatory eye movements show *only* a primacy effect and referential eye movements show *only* a recency effect is not supported. Indeed, the pattern suggests that we do find both effects in both time windows, but to a different degree. This is compatible with a *dominance* of the primacy effect in the earlier window and the recency effect in the latter. On the other hand, this data does not allow us yet to reject the lexical expectation hypothesis. Although this hypothesis did predict an interaction also for the referential time window, this lack can be easily explained by data sparseness.

If we reconsider the strong version of the 2-representation hypothesis, it rests on two assumptions. Firstly, only the first objects seen are to be processed in depth with the consequence that the conceptual affordances of the object are activated in addition to its visual aspects and its name etc. The objects seen later, however, are to be stored in the most shallow way. While this might be the general trend, we do not expect this to be the case in every trial: There are certainly many sources of random noise, like visual saliency or individual familiarity with specific objects. These might influence the amount of attention participants pay to the objects during the viewing phase which will then determine the complexity and strength of the internal representation in addition to the

serial position. Secondly, anticipatory eye movements are to rest solely on the affordance of the object, while referential eye movements rely only on the name. Existing research in the visual world paradigm, however, shows that referential eye movements (i.e. eye movements during the perception of a noun) can also rest on other features than the name (Huettig & Altmann, 2005; Yee & Sedivy, 2006; Huettig & Altmann, 2007). Since both assumptions do not seem to hold, we need to revise our original 2-representation hypothesis. Anticipatory eye movements are expected to be based mainly on conceptual representations, which should be more readily accessible if the object appeared early in the sequence. For the referential eye movements, there could in principal be primacy and recency effects: The name should activate both, the deep conceptual representation as well as the more shallow representation. On the other hand, the conceptual representation has often already resulted in an anticipatory eye movement, therefore a *new* inspection during the noun phrase is expected in fewer trials. For this reason, the recency effect is expected to dominate in this time window. Since the present experiment failed to conclusively show this pattern, we address this hypothesis again with Experiment S3.

4.2.6. Evaluation of the Paradigm

Compared to experiment S1, we were indeed able to observe a more canonical primacy effect due to our alterations. The recency effect, however, failed to reach significance for the referential time window. It is of course possible, that there is no significant serial position effect in referential eye movements. On the other hand, it is also possible that our experiment was not able to detect such an effect due to an insufficient amount of positive data points in combination with a 2x6 factorial analysis.

The first problem, the rather low proportion of trials containing *any* eye movements, is partly due to the different tasks: In experiment S1, there were fixations on the target object after the onset of the verb in 28% of all experimental trials. Due to the alteration of the task, this dropped to 15% of all trials in experiment S2. A closer analysis of the data also revealed that the location of the target object in the display influenced the probability of fixating it. For this analysis we split the screen by two diagonal lines into the four regions top, right, bottom, and left. The analysis revealed on the one hand that fixating an object was most probable if the object appeared in the top region, followed by left, right and bottom and, on the other hand, that there was considerably more noise in the left region: here the target object was not fixated significantly more than the comparitor object. While these effects should in principal be handled by the careful counterbalancing of the materials, the resulting data sparsity is likely to hinder the detection of the more subtle effects this experiment is investigating.

The second problem concerns the number of levels of POS. We never expected a crossing interaction, that is, more looks to the comparitor object in some conditions, but rather a modulation of the target advantage. We suspect that the interaction between the levels of position with object type are masked by the strong main effect of object type. Since we are not primarily interested in *all* levels of position, it would therefore make sense to reduce the levels to only three: one as an indicator of primacy, one as an indicator of recency, and the middle position, where both effects have least influence.

4.3. Experiment S3

The previous study indicated a primacy effect on language-mediated eye movements during the verb region while failing to give clear evidence for a recency effect during the noun region. This result did not allow us to decide between the 2-representation hypothesis and the lexical prediction hypothesis. The former predicted the primacy effect, but also a recency effect during the noun region. The latter predicted exactly the same effects in both time windows. This experiment aims to resolve this issue by addressing the data sparsity problem we encountered in the analysis of experiment S2.

Firstly, this experiment was designed to test more directly for primacy and recency effects. For this reason the number of levels of serial position (POS) was reduced to only three, one to test for primacy, one to test for recency and one functioning as a baseline for both tests. Additionally, we altered the lay-out of the display and had the target and distractor only appearing in the upper right half of the screen (see Fig. 4.13). This was done to increase the total number of fixations, since in S2 participants were more likely to re-fixate object locations, if they were presented in this spatial region. Since both, target and distractor appear in this part of the screen, this change is not expected to result in a generally greater advantage for the target. We suspect that it is the eye movement itself which is facilitated in this region rather than a memory-related process. If, however, the target was unexpectedly treated preferentially in memory encoding due to its position, no serial position effects would be expected at all. As in experiment S2, the objective of this experiment was to investigate whether anticipatory and referential eye movements rely on different internal representations. If anticipation of a missing role filler occurs on the basis of the semantic restrictions, the underlying representation of the visual object has to include conceptual properties like the affordance to be smoked (pipe). If, on the other hand, anticipation is a realization of a purely lexical expectation, a representation only entailing the lexical level would be sufficient. As illustrated in Figure 4.6, we assume the memory representations of objects seen late in a sequence to be shallow, in that they only contain lexical information and other surface features. For objects seen early in the sequence, on the other hand, we assume deeper, conceptual representations to be built up, which contain affordances and associations. This experiment utilizes this emergence of qualitatively different representations for different serial positions to infer what features anticipatory and referential eye movements rely on. Following the 2-representation hypothesis, anticipatory eye movements are expected to be dependent on a conceptual representation, therefore we predict a primacy effect, which is characteristic for these representations. Referential eye movements, on the other hand, can also be based on a more shallow representation and are therefore expected to exhibit a recency effect, possibly in addition to a primacy

effect, as the rich, conceptual representation also contains the lexical level. Alternatively, anticipatory eye movements could depend on lexical expectations. In this case, both kinds of eye movements would exploit the same features and are thus expected to show the same patterns.

4.3.1. Method

4.3.1.1. Participants

Thirty native speakers of German, all students from Saarland University, were paid 6 Euro to take part in this experiment. Age ranged from 18 to 42 with a mean of 24.5. Seven participants were male, 4 participants were left-handed. There were no psychology students taking part.[8]

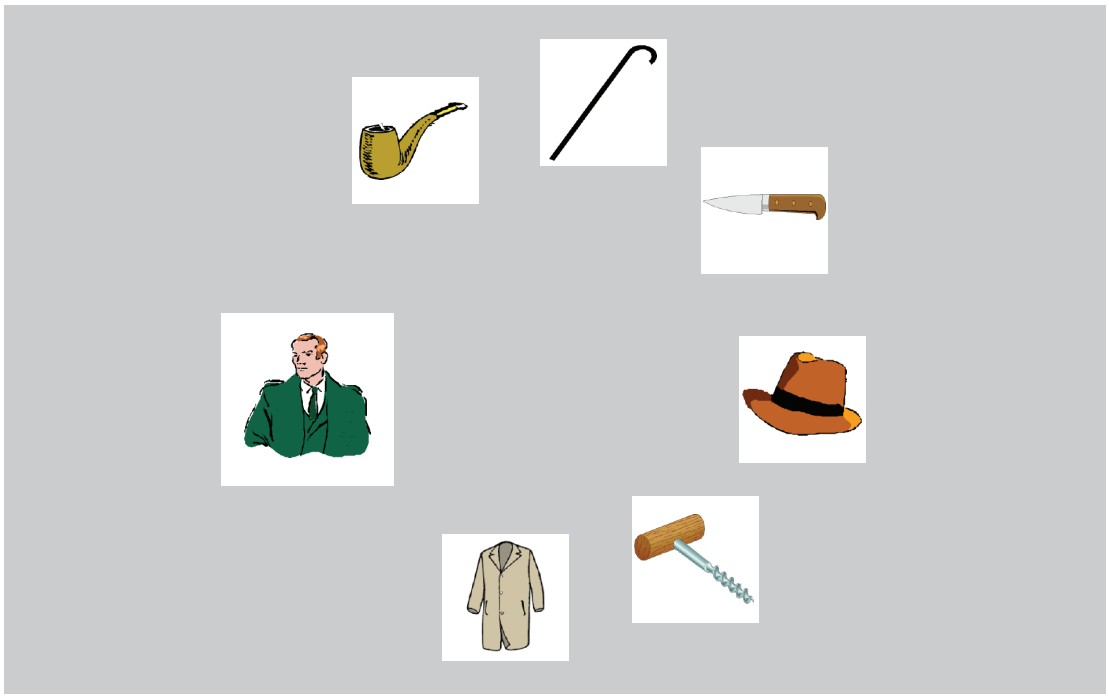
4.3.1.2. Materials

The materials were similar to S1 and S2 except for the following alterations: Out of the 30 experimental items of S2, 24 were picked. The display was changed to a clock-like lay-out (Figure 4.12) of white squares on a grey background with one larger square for the person (300 x 300 pixels) and 6 smaller squares for the objects (220 x 220 pixels). The person was always appearing in the lower left part of the screen and target and comparitor in one of the four positions on the upper right half of the screen. There was always one object between target and comparitor, resulting in four possible constellations, illustrated in Figure 4.13. A second lay-out was just like the first one, but rotated by 22.5°. Each item was assigned one constellation randomly and was used in two counter-balancing versions as in S1 and S2.

The order in which objects appeared on the screen was manipulated as in S1 and S2, except that target and comparitor object could only appear in `pos1`,`pos3` or `pos6` to test directly for primacy and recency effect against the baseline (`pos3`).

Forty-eight filler items were constructed. Half of them had restrictive verbs, too, the others had non-restrictive verbs. Of the restrictive ones, ten had only one possible role filler on the screen, similar to the experimental items. This object, however, was not mentioned in the sentence. The other restrictive verb fillers had between 2-3 possible objects in the display. In total, only twelve of the fillers had the object mentioned in the

[8] We were concerned that psychology students would be more likely to engage in mnemonics and be generally aware of serial position effects.



	sentence	target object	comparison object
version 1	Der Mann raucht vermutlich die Pfeife *The man smokes probably the pipe*	pipe	knife
version 2	Der Mann schärft vermutlich das Messer *The man sharpens probably the knife*	knife	pipe

Figure 4.12.: Example item

sentence also present on the display - half of them with restrictive verbs and the other half with nonrestrictive verbs. To counterbalance the bias of the experimental items to have the target always in the top-right region of the screen, these fillers had the named object always in one of the two bottom locations. In total, each location was thus equally likely to contain an object which was named in the sentence. Eighteen fillers were associated with a simple yes/no question to make sure participants payed attention to the whole sentence and not only the noun phrases. Six lists were created which contained each experimental item in only one condition and in one of the two counterbalancing versions using a latin square technique. The lists were randomized individually for each participant with the restriction that there had to be at least one filler between two experimental items.

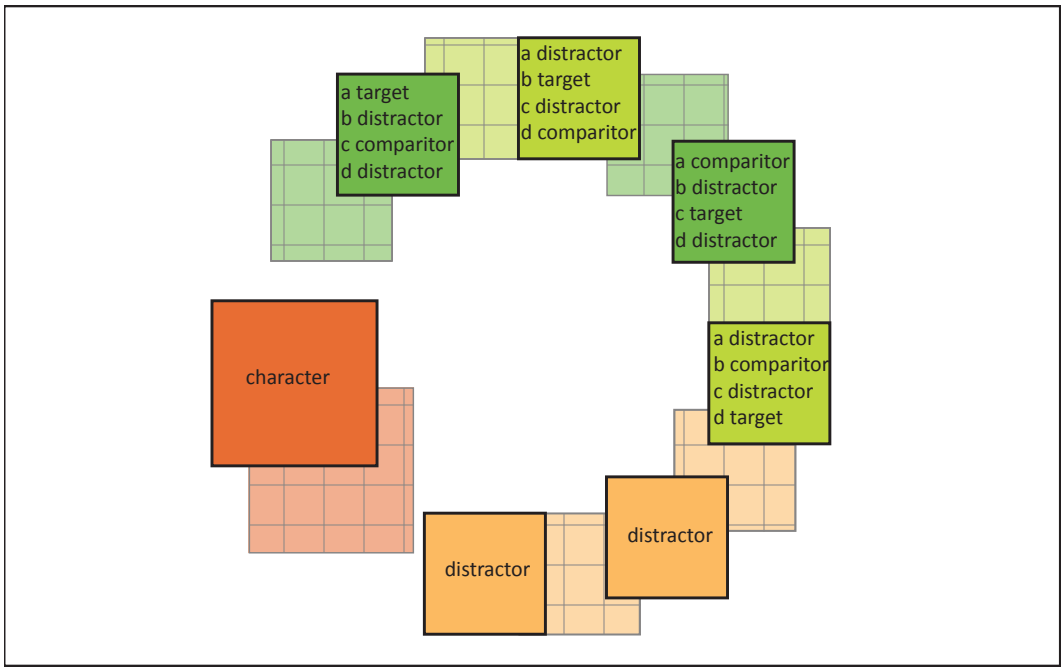


Figure 4.13.: The general lay-out of the screen, color coded for person (red), distractors (orange) and possible target/comparitor positions (green). The second, rotated lay-out is depicted faded in the background

4.3.1.3. Procedure

The procedure was the same as in S2 (see section 4.2.2.3). The experiment lasted approximately 35 min.

4.3.2. Predictions

Similar to experiment S2, the 2-representation hypothesis predicts a strong primacy effect on anticipatory eye movements as evidenced by an interaction between POS and OBJ in the primacy test, that is the analysis only containing levels `pos1` and `pos3` due to a stronger target advantage in `pos1`. For the recency test, that is the analysis only containing levels `pos3` and `pos6`, no effect or only a weak effect is expected. For referential eye movements, it predicts a strong recency effect and possibly a smaller primacy effect.

The lexical expectation hypothesis, on the other hand, predicts no difference between time windows: Either there should be primacy and recency effects for all time windows, or only recency effects, or none.

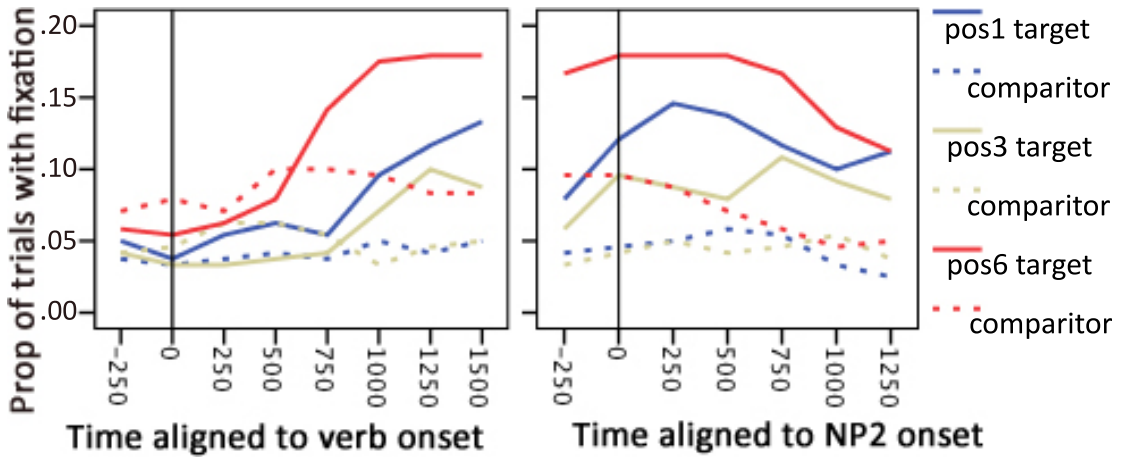


Figure 4.14.: Time course graph for target and comparitor objects aligned to verb onset and NP2 onset

4.3.3. Results

4.3.3.1. Time Course analysis

In Figure 4.14, the time course of the proportion of trials with fixation to target and comparitor object is depicted. It is important to keep in mind here, that only the *difference* between looks to target and comparitor object is indicating an influence of the linguistic stimulus. On the left-hand side, there is an advantage for target *and* comparitor in `pos6` right after verb onset, if we compare to the respective lines for `pos3`. Although looks to the target increase, they exceed looks to the comparitor object only after 750 ms. Looks to the target object in `pos1`, on the other hand, exceed those to the comparitor immediately after verb onset although only slightly rising at first. After the onset of the second NP (right-hand side), the difference between looks to target and comparitor object increase continuously for `pos6`, while for `pos1` the difference decreases again starting 250 ms after noun onset. This suggests again a different influence of the verb and noun on eye movements depending on whether they rest on a more conceptual representation in long-term memory or on a more shallow representation in short-term memory.

4.3.3.2. Fixation Data

In this experiment, the number of trials with fixations was considerably higher than in the previous two experiments, which indicates that the design alterations were successful in this respect. Therefore, it was possible to do an additional, more fine-grained analysis on more precise time windows than those described in section 4.1.3.1. The following time

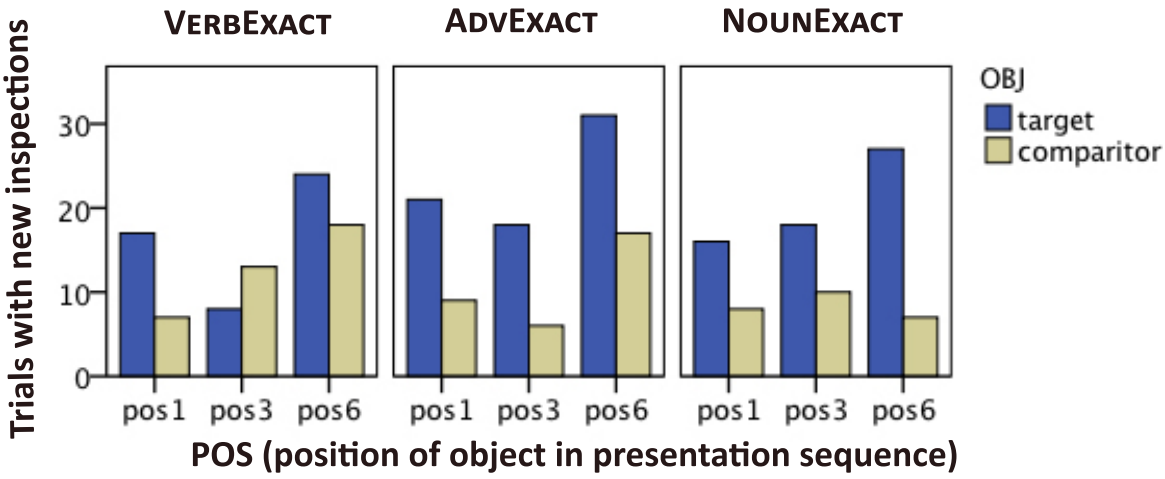


Figure 4.15.: Number of trials with newly started inspections to target and comparitor object in the time regions VERBEXACT, ADVEXACT and NOUNEXACT

windows were defined for this analysis: VERBEXACT started at verb onset and lasted until the onset of the post-verbal adverb. ADVEXACT started at adverb onset and lasted until the onset of the second noun.[9] NOUNEXACT started at noun onset and lasted until noun offset. The number of trials with newly started inspections in the new time windows are displayed in Figure 4.15. Also, we tested directly for primacy effects, comparing only levels `pos1` and `pos3` of POS and for recency effects, comparing only levels `pos3` and `pos6` of POS separately. The main effects were still assessed with the full model including all three levels of POS.

We first report the analysis based on the same time windows as in S1 and S2. For the VERBEND region, we found a significant effect of OBJ($\chi(1)$=15.07, p< .001), but no interaction between OBJ and POS. The same was true for the VERB region (effect of OBJ $\chi(1)$=4.89, p< .05) and the NP2 region (effect of OBJ$\chi(1)$=18.56, p< .001). No interactions and no main effects of POS were found. This shows a generally strong impact of the linguistic stimulus on eye movements in the anticipatory phase as well as the referential phase.

Let us now turn to the results of the more fine-grained analysis: For VERBEXACT there was no effect of OBJ, but a significant effect of POS ($\chi(2)$=6.70, p< .05) due to more looks in `pos6`. In the primacy test there was a significant interaction ($\chi(1)$=4.02, p< .05) caused by an enhanced target advantage for `pos1` compared to `pos3` (see Table 4.9). In the recency test the interaction was not significant ($\chi(1)$ =1.54). This indicates a target advantage only for the first position in this early time window, although more looks

[9] in S1 and S2 we used the onset of the NP, thus including the determiner in the region. In this experiment we decided for a more accurate coding

Predictor	Coefficiant	Std. Error	z value	p-value	
(Intercept)	-3.43	0.38	-8.93	< 2e-16	***
`target`	-0.52	0.54	-0.96	0.33	
`pos1`	-1.29	0.67	-1.93	0.05	.
`target:pos1`	2.00	0.77	2.59	0.01	**

Table 4.9.: Model summary for VerbExact time region (N = 960; log-likelihood = -169.3)
Model: $ins \sim OBJ + POS + OBJ * POS + (OBJ + POS|subj) + (1|item)$

Predictor	Coefficiant	Std. Error	z value	p-value	
(Intercept)	-3.79	0.42	-9.11	<2e-16	***
`target`	0.84	0.49	1.72	0.09	.
`pos6`	-1.19	0.66	-1.80	0.07	.
`target:pos6`	1.72	0.71	2.43	0.02	*

Table 4.10.: Model summary for NounExact time region (N = 960; log-likelihood = -212.6)
Model: $ins \sim OBJ + POS + OBJ * POS + (OBJ + POS|subj) + (OBJ + POS|item)$

were observed in `pos6` in general.

For AdvExact there was a significant main effect of OBJ ($\chi(1)$=5.22, p< .05) indicating a robust anticipation effect and a marginal main effect of POS ($\chi(2)$=5.35, p= .07) due to more looks in `pos6`. The primacy and recency tests showed no significant interactions($\chi(1)$ <1 in both cases). This indicates that the enhanced target advantage for the first object is very short-lived and that, apparently, even `pos3` elicited a target advantage here that was statistically indistinguishable from the ones in `pos1` and `pos6`.

For NounExact there was a significant main effect of OBJ ($\chi(1)$=10.85, p< .001) and no effect of POS. The primacy test showed no interaction ($\chi(1)$ =1.51). The recency test, however, showed a significant interaction ($\chi(1)$=5.47, p< .05) indicating an enhanced target advantage for `pos6` compared to `pos3` (see Table 4.10)

4.3.3.3. RTs and accuracy

For the RT analysis, RT was again defined as the time lag between the onset of the second NP and the button press indicating the decision whether both character and object mentioned in the sentence were present in the display. RTs further than two standard deviations away from the individual participant's mean were removed as outliers. There

Predictor	Coefficient	Std. Error	t value	MCMCmean	pMCMC	p-value	
(Intercept)	1522.91	60.01	25.38	1523.25	0.00	0.00	***
pos1	-74.64	40.69	-1.83	-75.43	0.06	0.07	.
pos6	-122.81	40.65	-3.02	-121.97	0.004	0.002	**

Table 4.11.: Model summary for Reaction Times (N = 685, log-likelihood = -5157)
Model: $RT \sim POS + (1|subj) + (1|item)$

was a significant effect of POS ($\chi(2)$=9.21, p< .05) due to significantly shorter RTs in pos6 and marginally shorter RTs in pos1. The coefficients in the model summary in Table 4.11 show that the recency effect had a bigger impact on RTs than the primacy effect.

For the accuracy data, there was also a significant effect of POS ($\chi(2)$=10.44, p< .01) due to significantly fewer errors in pos6 and pos1 compared to the baseline condition pos3 (Table 4.12). Both analyses show the presence of primacy and recency effect on the offline measures. This indicates that by using only the top-right half of the screen, we do not circumvent memory effects.

Predictor	Coefficient	Std. Error	z value	p-value	
(Intercept)	-1.32	0.21	-6.25	4.16e-10	***
pos1	-0.80	0.27	-2.94	0.003	**
pos6	-0.72	0.25	-2.84	0.005	**

Table 4.12.: Model summary for Accuracy (N = 720; log-likelihood = -317.4)
Model: $error \sim POS + (pos|subj) + (1|item)$

4.3.4. Discussion

In this experiment, we found a primacy effect on eye movements during the verb and a recency effect on eye movements during the referring noun. This supports the view that eye movements during the verb rely more strongly on the rich, conceptual representation of the object while referential eye movements during the noun may rely on more shallow representation. While this partly confirms our hypothesis, the results also show that these effects are subtle and short-lived: In the original anticipatory time window Verb, there was no interaction, the subsequent analyses show that this was due to eye movements during the adverb, which showed no significant serial position effects. The primacy effect is thus not dominating all eye movements before the onset of the noun, but emerges primarily during the verb.

The different patterns of eye movements in the three regions of interest lead us to put

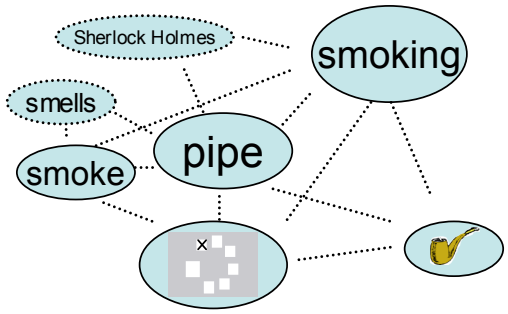
Position 1

stimulus was processed on multiple levels

(long-term memory)

rich representation

high level of activation



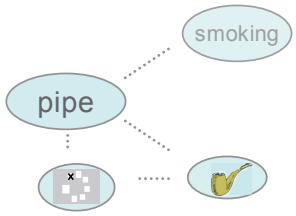
Position 3

stimulus was processed on visual and phonemic level (naming)

(short term memory)

shallow representation

low level of activation



Position 6

stimulus was processed on visual and phonemic level (naming)

(short term memory)

shallow representation

high level of activation

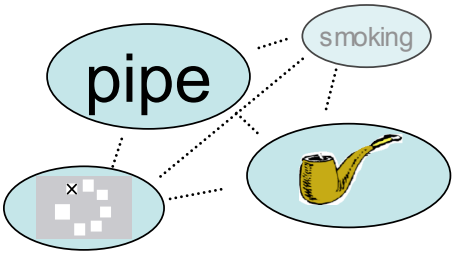


Figure 4.16.: Schematic depiction of different memory representations for *pipe* object in memory depending on its serial position: The rich representation for position 1 includes phonemic code, affordance, location, visual features, and associations; the shallow representations for position 3 and 6 include only part of the features with different degrees of activation.

forward a new interpretation. Our original division into anticipatory and referential eye movements proved incapable to capture the results of this study. Instead of assuming that all eye movements that occur before the NP reflect anticipation, we therefore propose that anticipation of a missing role filler can be observed primarily during the post-verbal time window adverb. Eye movements occuring during the verb itself are instead driven directly by the verb semantics in a quasi referential manner: The object *pipe* can be construed as an associate of the verb "smoke", which is reflected by an overlap between verb and associative or affordance-based features in the representation of the object. Allocating anticipation in the post-verbal time window is supported by eye movement patterns reported in studies where the verb semantics itself was not sufficient to anticipate the missing argument. As

discussed in section 2.1.2, in Kamide et al. (2003)'s experiments, the restrictions on possible role fillers introduced by the verb first had to be combined with case marking and world knowledge to enable the listener to anticipate the appropriate role filler. Eye movements reflecting this anticipation were only detected during the post-verbal adverb and not during the verb itself (for similar patterns see Knoeferle et al., 2005; Knoeferle & Crocker, 2006).

To understand the time course in which different memory representation become accessible during verb, adverb and noun phrase, reconsider the schematic representations of the target object "pipe" depending on its position in the presentation sequence as illustrated in Figure 4.16. When presented in the beginning of the list, the visual object "pipe" underwent deep processing resulting in a rich representation consisting of activated features for the phonemic code, the visual form, the location on the screen, but also semantic features such as the function of a pipe. The high number of active features will keep this representation's overall activation level high by means of spreading activation between the nodes. If "pipe" appeared later in the sequence, it was only processed on a superficial level, resulting in a more shallow representation presumably consisting only of features connected to the phonemic code, the visual form and the location of the object. In case it was at the end of the list, shown in the last row, this shallow representation would show a high level of activation, because no interfering phonemic and visuo-spatial information was able to override the activation pattern. If, however, "pipe" appeared in the middle of the list, the already shallow representation had suffered from decay or interference, leaving only a low level of activation for the small number of features.

Given these different kinds of representations, let us walk through the processing of the sentence "*The man smokes probably the pipe*" step by step. The first time the pipe representation becomes relevant is during the verb: *smokes* overlaps with the "smoking" feature describing the function of the pipe object in its rich representation. According to the featural overlap account discussed in section 2.1.5, the activation of the concept "smoking" by the verb *smoke* will spread to the other features which constitute the representation of "pipe", including its location. This, in turn, boosts the probability of executing an eye movement to the former location of the pipe during the verb. As the more shallow representations of pipe are less likely to contain a feature for its function, there will be no overlap between the two activation patterns and hence eye movements towards the former location of pipe are not likely to be affected by the processing of *smokes* in these cases.

The next word being processed is the adverb *probably* which is not expected to influence eye movements irrespective of the nature of the representation of pipe, because no one offers a basis for featural overlap. Instead, a prediction of the upcoming noun phrase is formed. It is possible that this prediction starts at the conceptual level based on verb-restrictions

and object affordances. In addition, possibly encouraged by the procedure and task, a lexical expectation is formed. Because the prediction now comprises conceptual and lexical features, there is an overlap between this prediction and all three representations sketched above. The probability of conducting an eye movement at this point mirrors the overall activation pattern of this representation, lowest for position 3.

Finally, the processing of the word *pipe* activates its phonological code which perfectly overlaps with the phonemic feature of all three memory representations. Again, the probability of conducting an eye movement depends on the level of activation, highest for the last position.

If we thus associate the verb region to lexical processing of the verb, the adverb region with anticipation of the next noun phrase and the noun region with lexical processing of the noun phrase, our results are mostly compatible with the lexical expectation hypothesis as even representations which presumably do not contain conceptual information were accessible for anticipatory eye movements. On the other hand, the lexical expectation hypothesis predicted the same pattern for anticipatory and referential time windows, whereas we did not find serial position effects during the adverb, but a significant recency effect during the noun phrase. This could be an indication that eye movements driven by expectations, although not relying solely on conceptual features, still exhibit different patterns than eye movements elicited by lexical processing.

One concern with the interpretation of the different patterns we found is the non-independence of the three time windows: If there was an inspection on the target object during an early time region, the probability of finding another inspection to it in a subsequent time window is low for two reasons: firstly our counting procedure is only sensitive to newly started inspections. It is thus only possible to find two inspections in the same trial, if the participant looks somewhere else in between. Secondly, even if the participant left the target object after fixating it during an early time region, she might not be willing to fixate the same region again right away. In the next chapter, we will describe an experiment that validates that the different patterns are not entirely due to the temporal relationship between the regions, but rather to the relationship between linguistic processing and different underlying memory representations. This is achieved by testing different types of reference in the same temporal position within a sentence.

Chapter 5.

Word-level Study: Contrasting Name and Category

Psycholinguistic studies in the visual world paradigm have utilized two kinds of eye movements as indicators of online language processing: referential and anticipatory eye movements (see section 2.1). The previous chapter developed a more fine-grained distinction between associative verb-induced eye movements, anticipatory eye movements, and noun-induced referential eye movements and provided evidence that those types of eye movements are affected differently by the accessibility of information in memory. We argued that the varying emergence of primacy and/or recency effects on these three types of eye movements indicate the reliance on qualitatively different memory representations. In this chapter, we provide independent evidence for the correspondence between serial position effects and the nature of underlying memory representations.

In the experiments in Chapter 4, participants were presented with a sequence of objects in different locations, one at a time, before the screen went blank and a sentence was played back. This sentence contained a restrictive verb that selected only for one of the previously depicted objects on the screen as a possible role filler, and a noun phrase referring to this object (‘*The man will smoke the pipe*‘). While processing the sentence, participants were in general more likely to refixate the prior location of the target object (*pipe*) than the location of a comparitor object. Crucially, this target advantage depended on the temporal position of the target object in the sequence of visual object presentation before the sentence was played back. Early verb-induced eye movements showed a primacy effect, that is the difference between inspections of target and comparitor object locations during the verb was greatest if the target had appeared early in the trial. Referential eye movements during the noun phrase, on the other hand, showed a recency effect: The advantage of the target over the comparitor during the noun was stronger if the target appeared late in the presentation sequence. For the interpretation, we adopted an account of serial position

effects that depends on the conceptual depth of representations in memory, where early items in a list are encoded with surface-level features as well as deep conceptual features while late items are represented by surface-level features alone (Atkinson & Shiffrin, 1968; Craik & Lockhart, 1972, see section 2.2.3). This then suggests that different types of eye movements rely on different kinds of stored memory representations: Verb-induced eye movements seem to rely on rich, conceptual representations while referential eye movements during the noun can also rely primarily on perceptual surface-level representations.

This inference relies on the difference in serial position effects in two time windows within the same sentence (verb and noun phrase). Observations in these two time windows, however, are not independent of each other. If a participant shifted her attention towards the target object location during the verb, she is less likely to do so again during the noun: First of all, she very recently fixated that location so she might prefer to inspect regions she has not yet visited. Secondly, she needs to shift her eyes to a different place in between, otherwise the look will only be counted as one long inspection which started during the verb. This dependence suggests that a primacy effect on the noun might have been underestimated. The temporal relationship between the two measuring regions also accommodates an alternative hypothesis: Possibly, objects that were seen early in the viewing phase are more likely to be revisited early during sentence comprehension coinciding with the verb, while objects that were seen late are more likely to be looked at late, that is, during the second noun phrase. Both lines of argument make it necessary to investigate whether the selective occurrence of recency and primacy effects for different kinds of words are also observed when measuring them independently from each other.

The present experiment attempts to verify the claim that different serial position effect patterns signalize the reliance on different memory representations. This is realized not by means of contrasting verbal with nominal material but with two different types of nominal reference to the same object: its name (basic level category) and its category (a hypernym). Similar to the influence of a verb, we expect eye movements triggered by a reference by category to rely more heavily on the full conceptual representation of an object. The name, on the other side, is expected to trigger eye movements based on both, the conceptual representation of an object, and the shallow perceptual representation. By using different types of reference, we can measure eye movements that are based on conceptual or shallow representations in the same position within the sentence eliminating possible confounds we were confronted with in the experiments in the previous chapter.

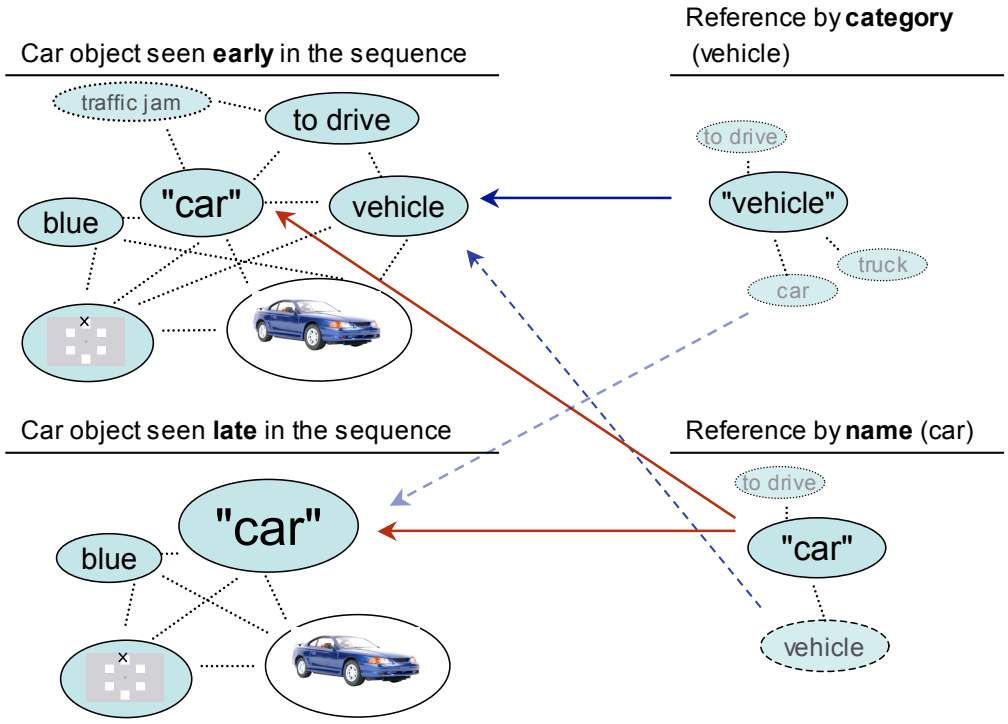


Figure 5.1.: Schematic activation patterns for visual stimuli in memory and linguistic stimuli during processing. Arrows indicate featural overlap between activations: Reference by name shows substantial overlap with both, shallow and rich representations in memory. Reference by category shows overlap primarily with rich representation.

5.1. Experiment

In this experiment, six objects appeared sequentially in different locations on a display, which went blank before a sentence containing two references to real objects was played back, such as '*Do you remember the car and the red object?*'. Participants' task was to respond to these sentences, deciding whether these two objects had been present in the previous display. The first reference was either the name of the object (*car*), or a hypernym (*vehicle*). As illustrated on the left-hand side of Figure 5.1, we hypothesize the rich, conceptual representation of the visual object seen early to contain information about the canonical name, as well as categorical information (vehicle) and other semantic and surface features. The more shallow representation build up if the object was seen late, on the other hand, is expected to contain information about the name and about other more perceptual features as, for instance, the color. The right-hand side shows part of the presumed activation pattern induced by the spoken word, which contains, in both

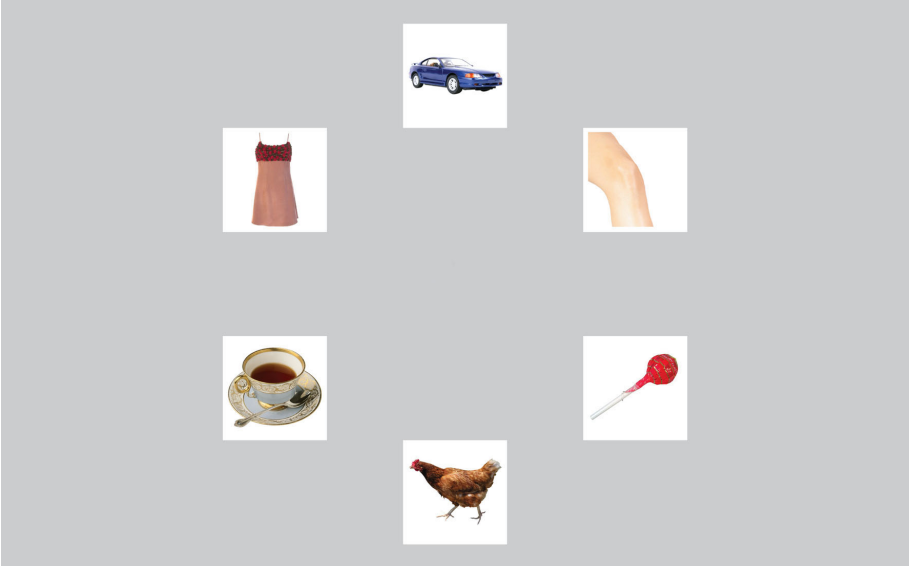
cases, the phonological form. For the reference by category, this coincides with a feature of the rich memory representation of the car object but not with any feature in the shallow representation. For the reference by the name of the object, seen below, this phonological form is part of both types of memory representations. Following the featural overlap account (Altmann & Kamide, 2007, see section 2.1.5), overlapping feature(s) are expected to re-activate the memory representation of the visual object, spreading activation also to its former location, which may then trigger an eye movement to this location. The apparent overlap therefore predicts eye movements when referring by name for both kinds of memory representation. For the reference by category, there should be eye movements primarily in the case of a rich conceptual memory representation. In addition, however, the processing of the spoken word will probably also activate features apart from the phonological code (Navarrete & Costa, 2005). As illustrated by the dashed arrows in Figure 5.1, these features could also exhibit some overlap with the internal memory representations. However, as the overall activation level of these features is presumably lower, this overlap is expected to have less influence than the phonological one.

The experimental hypothesis follows from the assumed activation pattern sketched above: If a primacy effect in language-mediated eye movements on a blank screen is indeed indicative of the reliance on a conceptual representation, we will see primacy for both types of reference, name and hypernym. If a recency effect in the same context is furthermore indicative of the reliance on a shallow representation, we expect a recency effect only for the reference by name, as the categorical information is not part of this representation. As an alternative to this new *2-representation hypothesis*, it is also possible that the partially consecutive emergence of primacy and recency effects in the previous experiments are not due to different levels of representation, but rather to the temporal delay between the two points of measurement. In this case, we would not expect a modulation of the serial position effects by the type of reference.

5.1.1. Method

5.1.1.1. Participants

Thirty-two native speakers of German, all students from Saarland University, were paid 5 Euro each to take part in this experiment. Age ranged from 20 to 46 with a mean of 25.4. Eight participants were male, 5 participants were left-handed.



Type	sentence
name	Erinnerst Du Dich an das Auto und an das rote Objekt? *Do you remember the car and the red object?*
category	Erinnerst Du Dich an das Fahrzeug und an das rote Objekt? *Do you remember the vehicle and the red object?*

Figure 5.2.: Example item

5.1.1.2. Materials

Thirty-two experimental items and 28 filler items were created (see Appendix A.3 for a full list of experimental items). Each one consisted of a display with six object photographs and a spoken sentence. The photographs were taken from the commercial collection Hemera Photo Objects. They were arranged in a circle around the center of the screen surrounded by white boxes on a grey background (Figure 5.2). The pictures were equidistant to the center of the screen as well as to their two immediate neighbors. There were two different layouts, where one was rotated by 30 degrees. The sentence always mentioned two objects. In half of the items, both objects were in the display, for the other half only one object was present. The objects could be referred to by their name, by a hypernym, or by a visual property (e.g., green, red, round, triangular). These types of reference appeared equally often and could be mixed in one trial.

For the experimental items, there were two versions of the sentence referring either by name or by a hypernym to the target object in the display. The two factors manipulated within participants were serial position (POS) and type of referring expression (TYPE).

POS had two levels (`pos1,pos5`) which corresponded to the serial position in the sequence of six presented objects. `Pos1` functioned as an indicator of a primacy effect, `pos5` of a recency effect. `Pos5` corresponds closely to `pos6` in the experiments S2 in 4.2 and S3 in 4.3, since there was exactly one picture following, the difference being that here it is an object while in the other experiments it was a person. TYPE had two levels: `name` and `cat` (category). This resulted in a total of four conditions.

The target object was always mentioned first thus allowing for enough time to elicit the relevant eye movements. There were five different carrier sentences, listed in Table 5.1. We conducted a naming norming study (N=12) where we instructed participants to write down the name of the object. Only objects that were assigned the same name by at least 10 participants as target objects were used, with two exceptions, where the intended name was embedded in a compound noun in the participants' responses. We also conducted a norming study that tested whether the target object was correctly identified using the hypernym for each experimental item display (N=10). All items for which this was not the case were excluded.

Results from a pilot study suggested, that not all locations on the screen are equally likely to be refixated. In trials where the target was in the top region, or the right region of the screen fixations were much more likely than in other trials where the target was in the bottom or left region. For this reason, we placed the target object in all experimental trials in the top region or the right region of the screen (see Figure 5.3 for the exact outline of the regions in both layouts). Since for each item this location was fixed and each item was presented in every condition, differences between conditions cannot be due to this decision. In order to prevent participants from expecting objects referred to in the sentence to be in this specific region, target objects in filler items and the second mentioned object in experimental items were placed in the remaining regions. This way, all regions were equally likely to contain an objects referred to in the sentence. An object mentioned in a sentence was furthermore equally likely to have been at the beginning (position 1 and 2), in the middle (position 3 and 4), or the end (position 5 and 6) of the presentation sequence.

1	Zu sehen war NP1 und NP2.	*There was NP1 and NP2.*
2	Erinnerst Du Dich an NP1 und an NP2?	*Do you remember NP1 and NP2?*
3	Hast Du NP1 und NP2 gesehen?	*Did you see NP1 and NP2?*
4	Du hast NP1 gesehen und NP2.	*You have seen NP1 and NP2.*
5	Du hast sicher NP1 und NP2 bemerkt.	*You have probably noticed NP1 and NP2.*

Table 5.1.: Carrier sentences with english translation

We created four lists. Each list contained all fillers and every item in only one

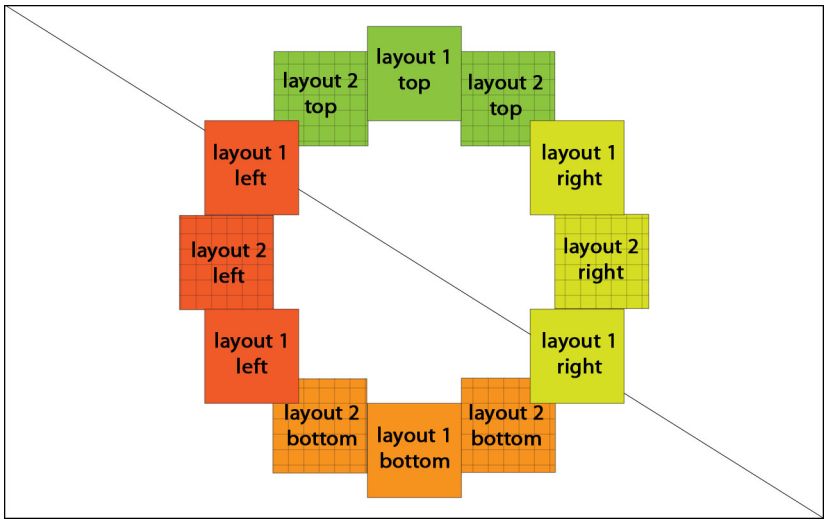


Figure 5.3.: Division of the screen in four regions: target objects were always presented in the top or the right region of the screen

condition. The lists were pseudo-randomized individually for each participant with the following restrictions. First, there could not be more than 2 fillers, experimental items in the name condition, or experimental items in the category condition in a row. Second, there could not be more than 3 trials in a row that required the same answer. Third, the two layouts always alternated in consecutive trials.

5.1.1.3. Procedure

The general procedure and equipment was similar to the experiments described in Chapter 4.

A trial started with the grey background template with the white boxes on it. After 1500 ms, the first object appeared in one of the boxes and remained for 1500 ms. Then, the object disappeared and after 200 ms the next object appeared in another box et cetera. 1000 ms after the last picture disappeared, the sentence was played back to the participants. Their task was to decide, whether both objects that were mentioned in the sentence had also been present as pictures on the display before. They then had to indicate their answer as fast as possible by pressing one of two buttons for “yes” and “no” on a button box. The experiment lasted approximately 30 min.

5.1.2. Predictions

The predictions we make concern only relative differences between conditions for the following reasons: We are agnostic to whether a name or a category should in general trigger more eye movements, this might also be dependent on the task to a certain degree. We also do not know whether an eye movement is executed with higher probability based on rich conceptual representations (when the target object was presented first, in position 1, within the series of object presentations that preceded the sentence) or on shallow but more recently build representations (when the object was presented in position 5). This depends presumably on the overall level of activation, about which we do not have sufficient information. For this reason, the two `name` conditions function as a baseline in the statistical models. As both kinds of representation allow for an eye movement to be triggered by the name of the object, differences between `name:pos1` and `name:pos5` reflect the general level of activation of these representations. Differences between `name:pos1` and `cat:pos1`, on the other hand, are expected to reflect which type of reference is more likely to trigger an eye movement.

Relative to the pattern for `name`, `cat` is expected to elicit a smaller amount of inspections of the target object's previous location in `pos5` compared to `pos1`, since only the rich representation build up for early shown objects can straightforwardly accommodate reference. We therefore expect an interaction between TYPE and POS in addition to possible main effects of either POS or TYPE. If both, name and category, trigger eye movements equally and if the overall level of activation for both representations is equal, we expect this interaction to be driven by less eye movements in condition `cat:pos5` compared to all other conditions.

5.1.3. Results

5.1.3.1. Method

Similar to the experiments in the last chapter, fixations were coded for target and non-target using color-coded templates. In these templates, the original squares containing the objects were enlarged from 220 px to 300 px in order to allow for measurement imprecision. Subsequent fixations on the same region were pooled into inspections and temporally related to the speech stream. The time window NOUN used for the analysis started 200 ms after the onset of the referential noun and lasted until 200 ms after the onset of the second noun. Trials were coded for containing a newly started inspection in this time window. Inferential analyses were again conducted using multilevel logistic regression with the

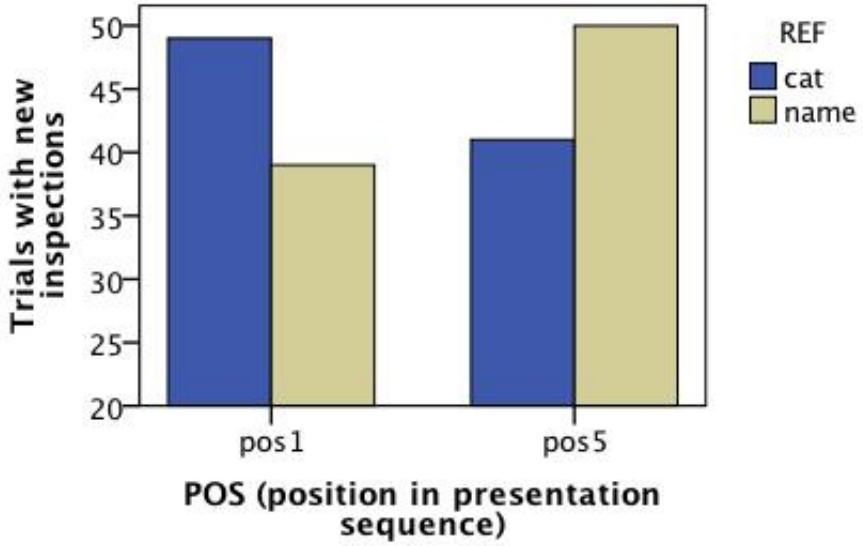


Figure 5.4.: Trials with new inspections of target

Predictor	Coefficiant	Std. Error	z value	p-value	
(Intercept)	-2.49	0.39	-6.38	1.81e-10	***
`cat`	0.59	0.28	2.12	0.03	*
`pos5`	0.76	0.31	2.43	0.02	*
`cat:pos5`	-0.87	0.38	-2.32	0.02	*

Table 5.2.: Model summary for NOUN time region (N = 960; log-likelihood = -398.7)
Model: $ins \sim TYPE * POS + (POS + TYPE|subj) + (1|item)$

two fixed factors POS (`pos1, pos5`) and TYPE (`name, cat`) and the baseline condition `pos1:name`. Random intercepts and slopes were included for participants and items. Two participants were excluded from analysis. One of them did not fixate all objects in the viewing phase, the other reported to have constantly pondered about were to move his eyes. In this case, we cannot expect eye movements to reflect linguistic processing and internal memory access.

5.1.3.2. Fixation data analysis

The two main effects of TYPE ($\chi(1) < 1$) and POS ($\chi(1) = 1.19$) were not significant. Importantly, there was a significant interaction between the two factors ($\chi(1) = 4.84, p < .05$). To be able to relate this interaction to our predictions, consider Figure 5.4 and the model summary in 5.2: Comparing to our baseline condition `name:pos1`, there was a boost

in inspections for `pos5` within the `name` level of TYPE. Furthermore, there were more inspections in `cat` than in `name` within the `pos1` level of POS. For these two conditions, we were not able to derive specific predictions based on the different structure of memory representations. Instead we take the difference between `name:pos1` and `name:pos5` to reflect the overall activation level of the two types of memory representations, where the representation of the more recently inspected object has an overall higher level of activation. The difference between `name:pos1` and `cat:pos1`, on the other hand, reflects a higher probability to launch an eye movement in response to a reference by category compared to the direct reference by name. Given these two results alone, we would expect the number of inspections in `cat:pos5` to exceed those in all other conditions. This is clearly not the case. Instead, the number of inspections in `cat:pos5` is almost equal to the number of inspections in the baseline condition `name:pos1`. This pattern suggests that a representation build for a recently inspected object is much less accessible for a reference by category than a representation build for the first object in the sequence.

5.2. Conclusion

The above described results support the hypothesis that primacy and recency effects in language-mediated eye movements are indicative of which kind of underlying representation is accessed. A rich, conceptual memory representation produces a primacy effect and accomodates eye movements in response to verbs as well as nouns referring either by category or name. More shallow representations, on the other hand, evoke a recency effect and primarily enable eye movements based on the referential noun itself, if it refers directly by name. This confirms the results and interpretation of the experiments in the previous chapter: As the different pattern of serial position effects in eye movements was also observed when measuring at the same position in the sentence, we can dismiss the alternative hypothesis that objects seen early are more likely to be revisited early in the subsequent sentence, while objects seen late are looked at with higher probability late in that sentence.

Chapter 6.

General Discussion

In order to attain a better understanding of situated language processing, the research presented in this thesis investigated and attempted to clarify the interplay of language processing on the one hand, and cognitive mechanisms involved in scene processing on the other hand. Part of the motivation was to determine the degree to which experimental results from the Visual World Paradigm scale up to the considerably more complex situations in which language processing usually takes place. In addition, we intended to acquire new insights regarding the representations and processes underlying situated language processing to complement existing accounts. The experiments on covert visual attention (Chapter 3) established that referring language can guide visual attention automatically, but volition can improve or partly suppress these effects. In Chapter 4, we demonstrated that language-mediated eye movements can rely on internal memory representations, even in the situation in which storage and access of these representations require some effort. Further, the results of Chapter 4 and Chapter 5 suggest that the complexity and general activation status of an internal memory representation determine its accessibility for language related processes. We will now shortly review these findings before we reconcile our data with the two accounts of language-mediated eye movements proposed by Altmann & Kamide (2007) and Knoeferle & Crocker (2006, 2007), and finally point out implications for situated language processing in natural situations.

6.1. Major Findings

In the preceding chapters, we reported the emergence of diverging memory related patterns of eye movements depending on the linguistic entity or process that triggered them. We identified two groups of language-mediated eye movements which a theory of situated language processing should be able to capture. The first group comprises referential eye movements in response to a noun phrase referring by name, hypernym, and eye movements

in direct response to a verb. Although these eye movements exhibit different memory related patterns, they have in common that they all show serial position effects. In contrast to this, eye movements that are driven by expectations, that is anticipatory eye movements that occur after processing the verb but before processing the noun phrase, do not show serial position effects. We will argue that this difference touches upon the automaticity of language-driven eye movements.

6.1.1. Referential and Verb-triggered Eye Movements

In our experiments, we captured three types of eye movements as a direct response to the processing of a linguistic unit: eye movements triggered by the name of an object, eye movements triggered by a category label of an object, and eye movements triggered by a restrictive verb. Our results are particularly conclusive about name-triggered eye movements, which are also generally the best studied type of language-related eye movements.

In Chapter 3, we investigated whether name-triggered eye movements happen automatically or whether they are under volitional control, using a modified version of Posner's paradigm on the orienting of covert visual attention. The measure for the allocation of covert visual attention was a speed-up in the detection of an unrelated target object in the critical location. We cued one of two possible locations by first displaying pictures of two objects in these locations, and then playing back a spoken word that referred to one of the objects. The short latency (300 ms after word onset) after which responses were facilitated gave us a first indication of automaticity. Further, the facilitation was present in a group that was discouraged from paying attention to the spoken word, and where the spoken word was as likely to cue the wrong location. A strategic integration of visual object and spoken word was thus not necessary to direct attention to the critical location. Importantly, however, the facilitation effect was significantly larger in a second group, for which the spoken word was more likely to cue the correct location than a false one; this group was also encouraged to take advantage of this.

The experiments in Chapter 4 and Chapter 5 investigated to what degree internal memory representations of visual objects are accessible for referential eye movements of different types. In the experiments, participants were presented sequences of object pictures in different locations of the screen, before they heard a sentence that either contained a restrictive verb and a referring noun phrase (Experiments S1, S2, and S3) or two referring noun phrases (Experiment W1). The experiments in Chapter 4 established that the target object is looked at in more trials than a comparison object, both during a restrictive verb, and during a referring noun phrase. Interactions between type of object (target or

comparitor) and position within the presentation sequence further demonstrated serial position effects on these eye movements. While eye movements during the verb exhibited a primacy effect, that is, more target fixations if the target object was seen early in the trial, referential eye movements showed a recency effect. The experiment in Chapter 5 further revealed a difference between eye movements in response to the name and in response to a category label. The direct comparison of target object fixations here yielded a stronger recency effect for the reference by name, while the primacy effect was stronger for the reference by category.

6.1.2. Anticipatory Eye Movements

Diverging from our description in Section 2.1.2, the different data patterns during and after the processing of a restrictive verb in Experiment S3 led us to consider eye movements during the verb separately from truly anticipatory eye movements. Choosing a truly anticipatory time region is inherently difficult, because it is defined by the end point, rather than its starting point. We allocated the starting point at the onset of the post-verbal adverb, which does not introduce any new referential material. For this post-verbal time window, we found significantly more looks to the location of the target than to the location to a comparitor object, irrespective of the serial position of the object in the presentation sequence. In contrast to the verbal time window and the referential time window discussed in the previous section, we did not find any serial position effects here.

6.2. Automaticity and Prediction in Situated Language Processing

In section 2.1.5 we presented two existing accounts of language mediated eye movements. Although they do not seem to be incompatible, they stress different aspects of the process. The featural overlap account (FOA, Altmann & Kamide, 2007) describes language mediated eye movements as an automatic process that arises as a byproduct of linguistic processing and scene processing. In their view, both linguistic processing and scene processing evoke representations consisting of activated multi-dimensional feature structures in memory. If a lexical representation shares features with the representation of a scene object, activation will spread from one to the other. As the former location is part of the feature structure representing the scene object, this location will also be re-activated which may induce an attentional shift towards it. The strength of the featural overlap here predicts the probability with which an eye movement towards the prior location is executed. The

coordinated interplay account (CIA, Knoeferle & Crocker, 2006, 2007), on the other hand, describes language-mediated eye movements as a search process which informs linguistic processing. After processing a given word in a sentence, the scene will be searched for referents for linguistic expressions already encountered as well as for anticipated referents. This search and co-indexation process in turn guides visual attention to ongoing actions the referent is part of and allows to acquire new information which is integrated with the existing interpretation. This way the visual information may also disambiguate between possible interpretations.

The main difference between the two accounts lies thus in the conceptualization and role of visual attention within linguistic processing. A clear prediction of the FOA is that language-induced shifts in visual attention arise automatically while processing the linguistic and visual stimulus. The CIA, on the other hand, suggests a top-down process. We interpret our results as showing both automatic aspects as well as top-down influences. In the following section we will again step through the processing of a sentence in the context of a visual scene to illustrate these two influences and explicate the observed emergence of serial position effects.

6.2.1. The Time Course of Activation and Prediction

We interpret the observed eye movement patterns as showing automatic activation as a by-product of linguistic processing, as well as top-down driven prediction of upcoming referents. In the first phase of our experiments that comprised the serial presentation of visual objects, we suppose that memory representations in the form of activated feature structures are built up for the individual objects. Depending on their position in the sequence, these representations vary in depth and overall activation. Objects seen early undergo deep processing resulting in rich and stable representations containing surface features as well as semantic features. Objects seen later in the trial only receive shallow processing. As a result, the representations are also shallow, containing mainly surface features like name and location. To illustrate these different representations and their degree of activation over the course of processing a sentence fragment, consider Figure 6.1. In the lower part of the figure, three different representations of the target ("pipe") and a comparitor ("knife") are sketched for the three positions first (pos1), middle (pos3) and last (pos6) in the presentation sequence at different stages of sentence processing. The size of the individual feature nodes correspond to their activation, whereas the overall level of activation is indicated by their vertical position in the figure. Note that these are not representations that compete directly in an individual trial, but correspond to the conditions in Experiment S3 in section 4.3. The three representations of the comparitor in

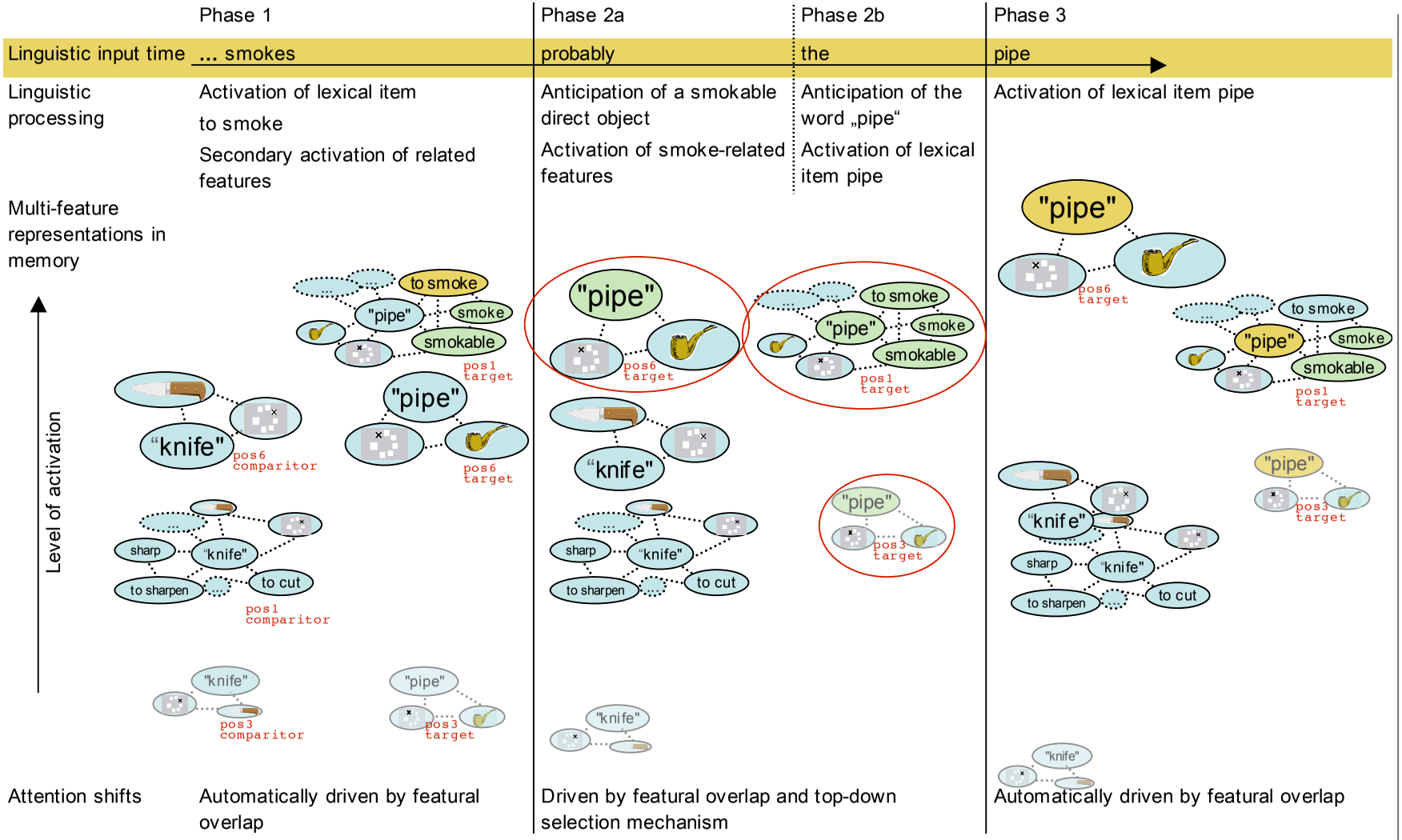


Figure 6.1.: Timecourse of activation patterns and top-down driven attention while processing the sentence fragment “ smokes probably the pipe”.

the left panel exemplify the different activation patterns of the same object we assume after the visual presentation and before the critical words are being processed. The representation for "knife" in first position is the most elaborate, containing multiple surface and semantic features that exhibit a moderate level of activation. In comparison, representations of "knife" seen in the middle or at the end of the sequence contain fewer features. Nevertheless, the representation of "knife" seen last exhibits the highest degree of activation, because no or little interfering information within the same modality was able to overwrite existing information. The representation for "knife" seen in the middle, on the other hand, is both shallow, and exhibits a generally low degree of activation.

During sentence processing, the activation status of these representations change due to lexical representations corresponding to the words being activated. In addition, the anticipation of linguistic entities also activates feature structures. The processing of the restrictive verb *smoke* (Phase 1 in 6.1) evokes an activation pattern that contains the phonological level as well as other features. *To smoke* is also part of the rich representation of the visual object pipe, when this object was seen early in the sequence. The reactivation of this feature therefore leads to an increase of the overall activation of this representation in comparison to a similar representation of a comparitor object, which is illustrated in Figure 6.1 by its lower vertical position. In this phase, we assume that the activation status predicts the probability of executing an eye movement to the former location of an object. The configuration in the left panel therefore suggests no difference in the amount of eye movements to target or comparitor location if seen in the middle or the end of the sequence: Few new fixations are expected on the location of an object seen in the middle, while both object locations are fairly likely to be fixated when the object was seen last. Importantly, the change in the activation status of the target representation if seen first predicts more eye movements in comparison to the comparitor object.

After the verb is processed, predictions about the upcoming linguistic material are formed. Our results suggest that these predictions entail at least the lexical level, illustrated in 6.1 by the reactivation of the feature "pipe" (Phase 2b). We further speculate that the lexical prediction is preceded by a conceptual prediction reactivating the affordance related features, if present (Phase 2a).[1] The reactivation causes an increase in activation of all target representations, while the comparitor representations slowly lose overall activation over time. The updated formation on the middle panel, however, does not account fully for the observed eye movements. In addition to the activation patterns, we therefore propose a top-down driven mechanism to influence the occurence of eye movements: The

[1] Our results do not allow us to distinguish between these two phases and only gives direct evidence for Phase 2b. This might partly be due to the sequential presentation of the visual objects, which encouraged the encoding and memorization of the name.

representation which exhibits the greatest overlap with a current prediction is selected, expressed by the red circles in Figure 6.1. Because of this mechanism, even representation with a low overall activation status are candidates for an eye movement, if they show some overlap with the predicted element.

Next, the processing of the second noun phrase activates the lexical representation of "pipe". This representation has common features with the representation of the visual object pipe irrespective of its serial position which causes another increase of activation for all target representations, while the shallow comparitor representations suffer from decay. As a consequence, the target representation for the object seen last exhibits the highest degree of activation and, importantly, a considerably higher degree of activation than the corresponding comparitor representation. The target representation for the object seen first is also comparatively activated. The difference between comparitor and target, however, is expected to be smaller, because the rich comparitor representation survives longer than the shallow one. For the object seen in the middle, the overall activation is still low, although we expect a difference in activation between target and comparitor. In this phase, we again suppose that the activation status mainly predicts the probability of an eye movement. In addition, already executed eye movements influence this probability: If the location was fixated before, it is less likely to receive a new inspection for two reasons. First, it is possible, that the location is still fixated. Our scheme of counting only new inspections thus does not take these cases into account. Second, a shift of attention to a location already attended to previously is less likely. Most new target inspections are therefore expected if it was seen late, as this representation is the most active one and attained its own highest degree of activation over the course of processing the sentence. For this reason, we were able to observe a recency effect in Experiment S3. Early objects are also good candidates, as the overall activation is high, but they already elicited an eye movement during the verb in more trials which reduces the probability of a new eye movement. As for objects seen in the middle, their representation contains a name feature, but the overall activation is rather low. Although the target might still be looked at more often than a comparitor, the recency effect suggests that the level of activation controls the probability of an eye movement.

The proposed time course of activation and prediction and its effects on language-mediated eye movements thus show signs of an automatic process induced by spreading activation between activated feature structures, as well as a top-down driven process that exploits all candidate feature structures with relatively little effect of their activation status. A more schematic illustration of these two influences can be found in Figure 6.2. Different representations of visual objects start out with different degrees of baseline activation: In

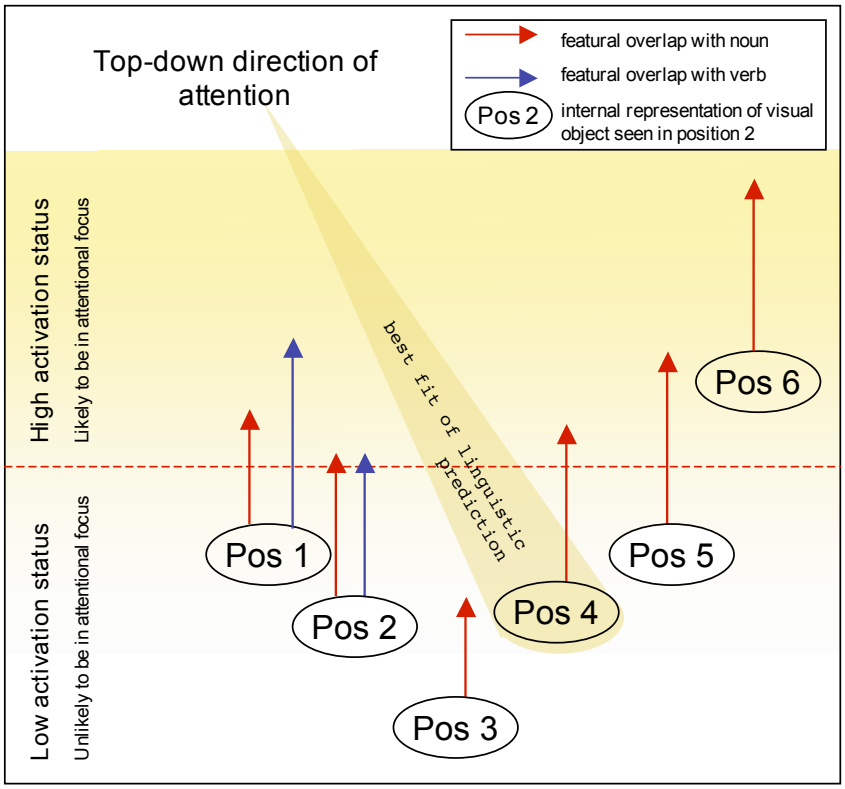


Figure 6.2.: Availability of memory representations for attention shifts: Baseline activation of object representations are depicted with arrows showing the influence of referential and associative processing. A top-down process can also directly select a representation, if it fits a linguistic prediction, as an example here the representation for the object in position 4

our experiments, this was mainly due to their serial position in the presentation sequence. In more realistic language processing environments there might be other factors involved. All representations with a high activation status (middle part of the figure) can induce a bottom-up shift of attention. If more than one representation reside in this area at the same time, these representations compete with higher activation increasing the likeliness of an eye movement. While there are only few representations that are within this range, language processing can change the pattern by boosting the activation of those representations that show featural overlap: The blue arrows indicate the consequence of (associative) featural overlap with the verb, while the red arrows show the result of featural overlap of the object's name with a noun. While a representation showing such overlap will experience a boost in activation regardless of its temporal position, it might or might not reach the threshold of sufficiently high activation to induce an eye movement. Importantly, a second mechanism, namely the top-down process that matches predictions with internal representations, can direct the focus of attention. For this mechanism, even representations below the activation

threshold for bottom-up attentional shifts are accessible, for instance the representation for the object in position 4, as in Figure 6.2.

Evidence for individual contributions of an automatic process and a top-down process was also found in the experiments in Chapter 3. The presence of an internal goal here considerably enhanced an existing automatic effect on covert visual attention. We interpreted this top-down effect as the influence of volition. In the case of anticipatory eye movements, it is not clear from our experiments, whether the effect is under control of volition, that is whether participants are actively searching the location the predicted object used to be in. Alternatively, the internal goal to find an anticipated referent could arise subconsciously when forming a linguistic prediction. Answering this question requires further experimentation.

Our interpretation of the observed eye movement patterns during sentences containing a restrictive verb rely heavily on the assumption that memory representations vary in depth and activation status based on the serial position. Independent evidence for the activation status comes from the literature on serial position effects (see section 2.2.3.1): Given an activated feature structure architecture, a differing activation status is the straight forward explanation of primacy and recency effects. In order to motivate the qualitatively different structures of representations based on their serial position, we draw on the theory of serial position effects proposed by Craik & Lockhart (1972), as well as our experimental findings in Chapter 5. We were able to show that the serial position predicts the accessibility of a memory representation for different linguistic stimuli.

6.2.2. Limitations of FOA and CIA

While above we interpreted our results with elements from both FOA and CIA, we will now shortly point why we do not think that one of them can explain all our data. Keep in mind, however, that the FOA and the CIA were originally formulated to account for different phenomena: The main focus of the FOA was to account for anticipatory and referential eye movements on a blank screen, i.e. drawing on internal representations. While the CIA also integrates the notion of working memory, the main goal was to account for the use of visual information in the course of situated language processing.

As described above, the FOA straightforwardly accounts for the different serial position effects we observed during referential processing, that is, in response to a verb or a noun phrase. The lack of serial position effects during the anticipatory adverb time window, however, is more problematic. If anticipatory eye movements rely on the overlap between activated affordance features in the objects' memory representations and the representation

of the linguistic input, there is no reason why serial position effects should differ between verb and adverb. In particular, the activation status which supposedly conditions the probability of executing an eye movement, should not have changed. One strategy to account for the anticipatory eye movements and their serial position patterns could be to suppose that predictions manifest themselves by augmenting the representation of the linguistic input. The representation of "The man will smoke" might activate lexical items likely to be mentioned next. This activation then overlaps with existing features in all kinds of representations of the visual object. In this case, however, we would expect the pattern during the adverb to match the pattern during the second noun phrase, which is not the case.

The CIA, on the other hand, is better able to account for the missing serial position effects during the adverb, by suggesting a top-down process meant to inform linguistic processing. Especially their implementation as a gate always selecting the best candidate is not dependent on the overall activation status any longer, if a threshold level of activation is reached. While in their implemented model the best fit is determined primarily based on the lexical level,[2] it is conceivable that minimally activated affordance features might also form the basis for such eye movements. With regard to referential eye movements, however, the CIA does not seem to explain the different serial position patterns we observed. In their model, referential eye movements and anticipatory eye movements follow the same top-down driven process and should therefore result in similar patterns. In addition, their conception of a working memory mechanism remains somewhat underspecified. They do not spell out the nature of underlying representations and do not seem to support them originally being different in strength and depth. While their notion of decay may explain recency effects, the occurrence of primacy effects is thus completely unexpected. Furthermore, the dominance of either recency or primacy depending on the triggering linguistic expression is not accounted for.

6.3. Implications for the Use of the Visual World Paradigm

Our experimental results and the insights we gained with regard to the underlying cognitive representations and processes leads us to reconsider aspects of the Visual World Paradigm regarding the generalization to more naturalistic situations, the interpretation of observed eye movements, and methodological details.

In section 2.1.6, we pointed out that the Visual World Paradigm in its canonical form is

[2] Although simple recurrent networks are able to learn categorical information, if provided with enough training data

in many aspects different from the kinds of situations in which we usually perceive language. Our experimental results establish that memory representations of visual objects are accessible for language-mediated eye movements, even if their use and organization requires some effort. This finding suggests that processes we observe within the Visual World Paradigm scale up to situations which require the use of internal memory representations, as for example an immersive environment with objects temporarily out of sight. On the other hand, we have seen that the depth and strength of a memory representation predicts its accessibility for distinct linguistic processes. We therefore expect eye movements that are triggered by expectations, referential matches, or merely semantically or otherwise related lexical units to show different vulnerability to the targeted object being out of sight.

Expanding on the last point, our results show that language-mediated eye movements cannot be described accurately as one uniform process. Instead, featural overlap and linguistic expectations may have independent influences on the direction of visual attention which makes it difficult to identify the word or expectation that triggered an eye movement. In particular, determining whether an eye movement is truly anticipatory or merely triggered by the verb semantics remains a controversial issue, with the two effects possibly overlaying each other. For future experimentation, this finding emphasizes the necessity of deconfounding referential with expectation-driven eye movements depending on the research question.

Finally, the finding that language-mediated attention is partly automatic, but can also be driven by internal goals and volition implies that the task used in a visual world experiment will affect the results. While the automatic influence will remain regardless of the task, a stronger, volitional influence may mask its effects. Our own results with a relatively weak task (serial picture-sentence verification) suggest that the top-down influences only depend on the semantic processing of the sentence. In summary, our results highlight the importance of methodological details for the use of the VWP and the interpretation of the results.

6.4. Conclusion

In this work, we presented six experiments that explored different aspects of the interplay of linguistic processing and visual attention in a context that required the use of memory representations. Two experiments on covert visual attention shifts induced by a picture-word pair shed light on the influence of a concurrent task. In particular, the automatic orientation effect we detected with a task that discouraged linguistic processing was

increased by a task that encouraged linguistic processing and the integration of the spoken word with the previously viewed picture. On the other hand, a task that encouraged an attentional shift away from a named object was not able to fully surpress an early orientation towards it. The remaining four experiments investigated the accessibility of internal object representations of differing depth and strength by manipulating the serial position of a target object in a presentation sequence that preceded the linguistic stimulus. We found different patterns depending on the relationship between linguistic expression and object (name, associated verb, category) and depending on the triggering process, which was either referential processing, or linguistic prediction. As a consequence, we suggest that the guidance of visual attention by language is not a uniform process. Our analysis combines aspects of two existing accounts to account for referential processing and linguistic prediction separately. While we attribute eye movements in response to a verb or a noun phrase to an automatic re-activation of the internal representation based on featural overlap, we propose that during prediction a top-down process selects the best fitting object as the target of a possible eye movement.

Our results indicate that processes of situated language processing as observed in the Visual World Paradigm generalize to settings where the use of internal memory representations is necessary. This suggests that such processes take place in more naturalistic language comprehension situations, too. The subtle influences of a concurrent task, of underlying representation structures, and of triggering linguistic processing stages stress the importance of including non-linguistic cognitive mechanisms in a comprehensive model of situated language processing.

Bibliography

Abrams, R. A. & Christ, S. E. (2003). Motion onset captures attention. *Psychological Science*, *14*(5), 427–432.

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the Time Course of Spoken Word Recognition Using Eye Movements: Evidence for Continuous Mapping Models. *Journal of Memory and Language*, *38*(4), 419–439.

Altmann, G. T. M. (2004). Language-mediated eye movements in the absence of a visual world: the 'blank screen paradigm'. *Cognition*, *93*(2), 79–87.

Altmann, G. T. M. (2011). Language can mediate eye movement control within 100 milliseconds, regardless of whether there is anything to move the eyes to. *Acta Psychologica*, *137*(2), 190–200.

Altmann, G. T. M. & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–64.

Altmann, G. T. M. & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language*, *57*(4), 502–518.

Andersson, R., Ferreira, F., & Henderson, J. M. (2011). I see what you're saying: the integration of complex speech and scenes during language comprehension. *Acta Psychologica*, *137*(2), 208–16.

Atkinson, R. C. & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation. Advances in research and theory*, volume 2 (pp. 89–195). New York: Academic Press.

Baddeley, A. D. (2000). The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423.

Baddeley, A. D. & Hitch, G. J. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation*, volume 8 of *Psychology of Learning and Motivation* chapter 3, (pp. 47–89). New York: Academic Press.

Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. (1997). Deictic codes for the embodiment of cognition. *The Behavioral and Brain Sciences*, *20*(4), 723–42.

Bates, D. (2008). Linear mixed model implementation in lme4. *Statistics*, *2010*(Figure 1), 1–32.

Burkell, J. A. & Pylyshyn, Z. W. (1997). Searching through subsets: a test of the visual indexing hypothesis. *Spatial Vision*, *11*(2), 225–258.

Cavanagh, P. & Alvarez, G. a. (2005). Tracking multiple targets with multifocal attention. *Trends in cognitive sciences*, *9*(7), 349–54.

Cooper, R. M. (1974). The Control of Eye Fixation by the Meaning of Spoken Language. *Cognitive Psychology*, *107*, 84–107.

Cowan, N. (2001). The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *The Behavioral and brain sciences*, *24*(1), 87–114; discussion 114–85.

Craik, F. & Lockhart, R. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 671–684.

Dahan, D. & Tanenhaus, M. K. (2005). Looking at the rope when looking for the snake: conceptually mediated eye movements during spoken-word recognition. *Psychonomic bulletin & review*, *12*(3), 453–9.

Driver, J., Davis, G., Ricciardelli, P., Kidd, P., Maxwell, E., & Baron-Cohen, S. (1999). Gaze Perception Triggers Reflexive Visuospatial Orienting. *Visual Cognition*, *6*(5), 509–540.

Ellsiepen, E., Knoeferle, P., & Crocker, M. W. (2008). Incremental syntactic disambiguation using depicted events : plausibility , co-presence and dynamic presentation. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society, Washington D.C., USA*, number 2005, (pp. 2398–2403).

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, *211*, 1–28.

Friesen, C. K., Ristic, J., & Kingstone, A. (2004). Attentional effects of counterpredictive gaze and arrow cues. *Journal of experimental psychology. Human perception and performance*, *30*(2), 319–29.

Frisson, S., Rayner, K., & Pickering, M. J. (2005). Effects of contextual predictability and transitional probability on eye movements during reading. *Journal of experimental psychology. Learning memory and cognition*, *31*(5), 862–877.

Glanzer, M. & Cunitz, A. R. (1966). Two storage mechanisms in free recall. *Journal Of Verbal Learning And Verbal Behavior*, *5*(4), 351–360.

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

Henderson, J. M. (1992). Visual attention and eye movement control during reading and picture viewing. In K. Rayner (Ed.), *Eye movements and visual cognition: Scene perception and reading* (pp. 260–283). New York: Springer-Verlag.

Henderson, J. M. & Ferreira, F. (2004). Scene perception for psycholinguists. In J. M. Henderson & F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world* (pp. 1–58). Psychology Press New York.

Hollingworth, A. & Henderson, J. M. (2002). Accurate visual memory for previously attended objects in natural scenes. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(1), 113–136.

Hommel, B., Pratt, J., Colzato, L., & Godijn, R. (2001). Symbolic control of visual attention. *Psychological Science*, *12*(5), 360–365.

Huettig, F. & Altmann, G. T. M. (2005). Word meaning and the control of eye fixation: semantic competitor effects and the visual world paradigm. *Cognition*, *96*(1), B23–32.

Huettig, F. & Altmann, G. T. M. (2007). Visual-shape competition during language-mediated attention is based on lexical input and not modulated by contextual appropriateness. *Visual Cognition*, *15*(8), 985–1018.

Itti, L. & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision research*, *40*(10-12), 1489–506.

Jonides, J. (1981). Voluntary versus automatic control over the mind's eye's movement. In J. B. Long & A. D. Baddeley (Eds.), *Attention and performance IX*, volume 9 chapter 11, (pp. 187–203). Erlbaum.

Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, *49*(1), 133–156.

Kamide, Y., Scheepers, C., & Altmann, G. T. M. (2003). Integration of syntactic and semantic information in predictive processing: cross-linguistic evidence from German and English. *Journal of psycholinguistic research*, *32*(1), 37–55.

Knoeferle, P. & Crocker, M. W. (2006). The coordinated interplay of scene, utterance, and world knowledge: Evidence from eye tracking. *Cognitive science*, *30*(3), 481–529.

Knoeferle, P. & Crocker, M. W. (2007). The influence of recent scene events on spoken comprehension: Evidence from eye movements. *Journal of Memory and Language*, *57*(4), 519–543.

Knoeferle, P., Crocker, M. W., Scheepers, C., & Pickering, M. J. (2005). The influence of the immediate visual context on incremental thematic role-assignment: evidence from eye-movements in depicted events. *Cognition*, *95*(1), 95–127.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word Frequencies in Written and Spoken English: based on the British National Corpus*. Longman.

Luck, S. J. & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279–81.

Magnuson, J. S., Dixon, J. a., Tanenhaus, M. K., & Aslin, R. N. (2007). The dynamics of lexical competition during spoken word recognition. *Cognitive science*, *31*(1), 133–56.

Mayberry, M. R., Crocker, M. W., & Knoeferle, P. (2009). Learning to attend: a connectionist model of situated language comprehension. *Cognitive science*, *33*(3), 449–96.

McClelland, J. L., Elman, J. L., & Diego, S. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1–86.

McDonald, S. A. & Shillcock, R. C. (2003). Low-level predictive inference in reading: the influence of transitional probabilities on eye movements. *Vision Research*, *43*(16), 1735–1751.

McElree, B. (2006). Accessing recent events. *Psychology of learning and motivation*, *46*(06), 155–200.

Miles, W. R. (1930). Ocular dominance in human adults. *Journal of General Psychology*, *3*, 412–430.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, *63*, 81–97.

Murdock, Bennet B., J. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, *64*(5), 482–488.

Nairne, J. S. (2002). Remembering over the short-term: the case against the standard model. *Annual review of psychology*, *53*, 53–81.

Navarrete, E. & Costa, A. (2005). Phonological activation of ignored pictures: Further evidence for a cascade model of lexical access. *Journal of Memory and Language*, *53*(3), 359–377.

Oberauer, K. (2002). Access to Information in Working Memory : Exploring the Focus of Attention. *Cognition*, *28*(3), 411–421.

O'Regan, J. (1992). Solving the *real* mysteries of visual perception: The world as an outside memory. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *46*(3), 461.

O'Reilly, R. C. O., Braver, T. S., & Cohen, J. D. (1999). A Biologically-Based Computational Model of Working Memory. In A. Miyake & P. Shah (Eds.), *Models of Working Memory* (pp. 375–411). Cambridge: Cambridge University Press.

Oztekin, I., Davachi, L., & McElree, B. (2010). Are representations in working memory distinct from representations in long-term memory? Neural evidence in support of a single store. *Psychological science*, *21*(8), 1123–33.

Posner, M. I. (1980). Orienting of attention. *The Quarterly Journal of Experimental Psychology*, *32*(1), 3–25.

Postman, L. & Phillips, L. W. (1965). Short-term temporal changes in free recall. *Quarterly Journal of Experimental Psychology*, *17*(2), 132–138.

Pylyshyn, Z. W. (2001). Visual indexes, preconceptual objects, and situated vision. *Cognition*, *80*(1-2), 127–58.

Pylyshyn, Z. W. & Storm, R. W. (1988). Tracking multiple independent targets: evidence for a parallel tracking mechanism. *Spatial vision*, *3*(3), 179–97.

Rensink, R. A., Regan, J. K. O., & Clark, J. J. (1997). To see or not to see : The Need for Attention to Perceive Changes in Scenes. *Psychological Science*, *8*(5), 1–6.

Richardson, D. C. & Spivey, M. J. (2000). Representation, space and Hollywood Squares: looking at things that aren't there anymore. *Cognition*, *76*(3), 269–295.

Salverda, A. P. & Altmann, G. T. M. (2011). Attentional capture of objects referred to by spoken language. *Journal of experimental psychology. Human perception and performance*, *37*(4), 1122–33.

Scholl, B. J. & Pylyshyn, Z. W. (1999). Tracking multiple items through occlusion: clues to visual objecthood. *Cognitive psychology*, *38*(2), 259–90.

Simons, D. J. & Levin, D. T. (1998). Failure to detect changes to people during a real-world interaction. *Psychonomic Bulletin & Review*, *5*(4), 644–649.

Spivey, M. J. & Geng, J. J. (2001). Oculomotor mechanisms activated by imagery and memory: eye movements to absent objects. *Psychological research*, *65*(4), 235–41.

Tanenhaus, M. K., Magnuson, J. S., Dahan, D., & Chambers, C. G. (2000). Eye movements and lexical access in spoken-language comprehension: Evaluating a linking hypothesis between fixations and linguistic processing. *Journal of psycholinguistic research*, *29*(6), 557–80.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634.

Tipples, J. (2002). Eye gaze is not unique: Automatic orienting in response to uninformative arrows. *Psychonomic bulletin & review*, *9*(2), 314–8.

Tipples, J. (2008). Orienting to counterpredictive gaze and arrow cues. *Attention, Perception, & Psychophysics*, *70*(1), 77–87.

Yee, E. & Sedivy, J. C. (2006). Eye movements to pictures reveal transient semantic activation during spoken word recognition. *Journal of experimental psychology. Learning, memory, and cognition*, *32*(1), 1–14.

Zelinsky, G. J. & Murphy, G. L. (2000). Synchronizing visual and language processing: an effect of object name length on eye movements. *Psychological Science*, *11*(2), 125–131.

Zimmer, H. D. (1998). Spatial information with pictures and words in visual short-term memory. *Psychological research*, *61*(4), 277–84.

Appendix A.

Experimental Material

A.1. Covert Visual Attention Experiments

Verbal Material Experiment A1

Item number	left object	right object	TrialType uninformative group	TrialType predictive group	SOA
1	pipe	ear	compatible	compatible	200
2	paint	lamp	compatible	compatible	500
3	cake	hat	compatible	compatible	800
4	spring	match	compatible	compatible	200
5	leaf	ring	compatible	compatible	500
6	bug	glue	compatible	compatible	800
7	cup	ball	compatible	compatible	200
8	kite	rake	compatible	compatible	500
9	wood	bird	compatible	compatible	800
10	crib	broom	compatible	compatible	200
11	beer	pie	compatible	compatible	500
12	jam	bow	compatible	compatible	800
13	sink	jeans	compatible	compatible	200
14	bench	shelf	compatible	compatible	500
15	seat	case	compatible	compatible	800
16	cross	rose	compatible	compatible	200
17	tub	pram	compatible	compatible	500
18	wheel	gate	compatible	compatible	800
19	cork	sweets	compatible	compatible	200
20	egg	tea	compatible	compatible	500
21	fly	soap	compatible	compatible	800
22	mouse	nose	compatible	compatible	200
23	salt	disk	compatible	compatible	500

Item number	left object	right object	TrialType uninformative group	TrialType predictive group	SOA
24	comb	snail	compatible	compatible	800
25	boot	rock	compatible	incompatible	200
26	mask	sword	compatible	incompatible	500
27	dice	clasp	compatible	incompatible	800
28	crisps	scoop	compatible	incompatible	200
29	thorn	clamp	compatible	incompatible	500
30	plate	wine	compatible	incompatible	800
31	juice	straw	compatible	incompatible	200
32	skate	vase	compatible	incompatible	500
33	nail	stamp	compatible	incompatible	800
34	drill	rug	compatible	incompatible	200
35	glass	clock	compatible	incompatible	500
36	shell	flame	compatible	incompatible	800
37	board	card	incompatible	incompatible	200
38	crown	shirt	incompatible	incompatible	500
39	cab	palm	incompatible	incompatible	800
40	pear	bib	incompatible	incompatible	200
41	belt	tray	incompatible	incompatible	500
42	string	watch	incompatible	incompatible	800
43	whisk	grape	incompatible	incompatible	200
44	rim	whip	incompatible	incompatible	500
45	jar	pill	incompatible	incompatible	800
46	thread	drop	incompatible	incompatible	200
47	phone	chair	incompatible	incompatible	500
48	foot	sign	incompatible	incompatible	800
49	bike	flag	compatible	compatible	200
50	box	scale	compatible	compatible	500
51	bee	toe	compatible	compatible	800
52	bell	can	compatible	compatible	200
53	toast	fork	compatible	compatible	500
54	chess	torch	compatible	compatible	800
55	bolt	sock	compatible	compatible	200
56	sand	net	compatible	compatible	500
57	fish	plant	compatible	compatible	800
58	frog	cone	compatible	compatible	200
59	pen	tap	compatible	compatible	500
60	glove	brush	compatible	compatible	800
61	book	jug	compatible	compatible	200
62	tin	bone	compatible	compatible	500
63	sponge	lime	compatible	compatible	800

Item number	left object	right object	TrialType uninformative group	TrialType predictive group	SOA
64	dart	scarf	compatible	compatible	200
65	screw	purse	compatible	compatible	500
66	fence	rope	compatible	compatible	800
67	tooth	chain	compatible	compatible	200
68	hose	axe	compatible	compatible	500
69	leek	prawn	compatible	compatible	800
70	shoe	milk	compatible	compatible	200
71	fan	gun	compatible	compatible	500
72	nest	duck	compatible	compatible	800
73	pin	thumb	compatible	incompatible	200
74	cage	throne	compatible	incompatible	500
75	spoon	peach	compatible	incompatible	800
76	hook	brow	compatible	incompatible	200
77	doll	mug	compatible	incompatible	500
78	lock	cheese	compatible	incompatible	800
79	frame	dress	compatible	incompatible	200
80	pan	owl	compatible	incompatible	500
81	shot	worm	compatible	incompatible	800
82	bulb	yarn	compatible	incompatible	200
83	desk	boat	compatible	incompatible	500
84	pants	harp	compatible	incompatible	800
85	plane	bridge	incompatible	incompatible	200
86	stone	mouth	incompatible	incompatible	500
87	couch	shark	incompatible	incompatible	800
88	tent	sack	incompatible	incompatible	200
89	tape	key	incompatible	incompatible	500
90	saw	robe	incompatible	incompatible	800
91	bread	pot	incompatible	incompatible	200
92	tree	car	incompatible	incompatible	500
93	globe	spade	incompatible	incompatible	800
94	lid	rice	incompatible	incompatible	200
95	trunk	stool	incompatible	incompatible	500
96	moth	drain	incompatible	incompatible	800
97	plane	coat	compatible	compatible	200
98	slide	rope	compatible	compatible	500
99	clock	knife	compatible	compatible	800
100	jam	doll	compatible	compatible	200
101	net	disk	compatible	compatible	500
102	kite	saw	compatible	compatible	800
103	drop	cheese	compatible	compatible	200

Item number	left object	right object	TrialType uninformative group	TrialType predictive group	SOA
104	suit	card	compatible	compatible	500
105	rock	bird	compatible	compatible	800
106	throne	sink	compatible	compatible	200
107	fish	bridge	compatible	compatible	500
108	wheel	boot	compatible	compatible	800
109	tent	brush	compatible	compatible	200
110	bolt	ant	compatible	compatible	500
111	pill	can	compatible	compatible	800
112	bag	key	compatible	compatible	200
113	toast	juice	compatible	compatible	500
114	axe	sponge	compatible	compatible	800
115	sock	bulb	compatible	compatible	200
116	cage	jeans	compatible	compatible	500
117	frame	rose	compatible	compatible	800
118	car	book	compatible	compatible	200
119	vase	spade	compatible	compatible	500
120	crown	nose	compatible	compatible	800
121	sign	chair	compatible	incompatible	200
122	shelf	moon	compatible	incompatible	500
123	broom	snail	compatible	incompatible	800
124	peach	dice	compatible	incompatible	200
125	rug	pan	compatible	incompatible	500
126	brow	jar	compatible	incompatible	800
127	thumb	bee	compatible	incompatible	200
128	globe	couch	compatible	incompatible	500
129	bowl	pen	compatible	incompatible	800
130	comb	drain	compatible	incompatible	200
131	lime	spoon	compatible	incompatible	500
132	pipe	beer	compatible	incompatible	800
133	stamp	fly	incompatible	incompatible	200
134	flame	bench	incompatible	incompatible	500
135	torch	thread	incompatible	incompatible	800
136	foot	plant	incompatible	incompatible	200
137	bus	ear	incompatible	incompatible	500
138	board	cup	incompatible	incompatible	800
139	sword	fence	incompatible	incompatible	200
140	tape	wood	incompatible	incompatible	500
141	skate	chess	incompatible	incompatible	800
142	cab	rice	incompatible	incompatible	200
143	milk	ring	incompatible	incompatible	500

Item number	left object	right object	TrialType uninformative group	TrialType predictive group	SOA
144	pot	bell	incompatible	incompatible	800
145	mug	drill	compatible	compatible	200
146	bow	pin	compatible	compatible	500
147	tap	owl	compatible	compatible	800
148	bed	tree	compatible	compatible	200
149	mouse	lock	compatible	compatible	500
150	seat	box	compatible	compatible	800
151	hose	cork	compatible	compatible	200
152	cross	dress	compatible	compatible	500
153	grape	cone	compatible	compatible	800
154	bug	pear	compatible	compatible	200
155	plate	leaf	compatible	compatible	500
156	bone	fan	compatible	compatible	800
157	belt	tin	compatible	compatible	200
158	watch	shirt	compatible	compatible	500
159	tea	ball	compatible	compatible	800
160	moth	leek	compatible	compatible	200
161	glove	trunk	compatible	compatible	500
162	bean	nest	compatible	compatible	800
163	scale	match	compatible	compatible	200
164	string	case	compatible	compatible	500
165	screw	dart	compatible	compatible	800
166	tray	palm	compatible	compatible	200
167	yarn	stool	compatible	compatible	500
168	maize	scarf	compatible	compatible	800
169	glue	bib	compatible	incompatible	200
170	knob	clasp	compatible	incompatible	500
171	frog	whisk	compatible	incompatible	800
172	sieve	crisps	compatible	incompatible	200
173	phone	wine	compatible	incompatible	500
174	jug	pram	compatible	incompatible	800
175	pail	wig	compatible	incompatible	200
176	desk	cake	compatible	incompatible	500
177	glass	stone	compatible	incompatible	800
178	rake	thorn	compatible	incompatible	200
179	pie	toe	compatible	incompatible	500
180	shoe	paint	compatible	incompatible	800
181	lid	bat	incompatible	incompatible	200
182	shell	tie	incompatible	incompatible	500
183	hat	salt	incompatible	incompatible	800

Item number	left object	right object	TrialType uninformative group	TrialType predictive group	SOA
184	purse	swan	incompatible	incompatible	200
185	lamp	bike	incompatible	incompatible	500
186	sack	worm	incompatible	incompatible	800
187	flag	nail	incompatible	incompatible	200
188	sweets	pants	incompatible	incompatible	500
189	duck	cap	incompatible	incompatible	800
190	mask	straw	incompatible	incompatible	200
191	bread	chain	incompatible	incompatible	500
192	gun	egg	incompatible	incompatible	800

Verbal material Experiment A2

Item number	left object	right object	TrialType	neutral word	SOA
1	flame	bench	incompatible		300
2	glove	trunk	incompatible		300
3	string	case	incompatible		300
4	drop	cheese	incompatible		300
5	watch	shirt	incompatible		300
6	sock	bulb	incompatible		300
7	stamp	fly	incompatible		1200
8	cage	jeans	incompatible		1200
9	bug	pear	incompatible		1200
10	desk	cake	incompatible		1200
11	glue	bib	incompatible		1200
12	foot	plant	incompatible		1200
13	moth	leek	incompatible		300
14	bow	pin	incompatible		300
15	tent	brush	incompatible		300
16	cab	rice	incompatible		300
17	sweets	pants	incompatible		300
18	glass	stone	incompatible		300
19	axe	sponge	incompatible		1200
20	tape	wood	incompatible		1200
21	vase	spade	incompatible		1200
22	throne	sink	incompatible		1200
23	net	disk	incompatible		1200
24	clock	knife	incompatible		1200
25	flag	nail	incompatible		300
26	yarn	stool	incompatible		300

Item number	left object	right object	TrialType	neutral word	SOA
27	gun	egg	incompatible		300
28	mug	drill	incompatible		300
29	sieve	crisps	incompatible		300
30	gate	tooth	incompatible		300
31	star	church	incompatible		1200
32	wheel	boot	incompatible		1200
33	kite	saw	incompatible		1200
34	sign	chair	incompatible		1200
35	pot	bell	incompatible		1200
36	bag	key	incompatible		1200
37	tea	ball	incompatible		300
38	pie	toe	incompatible		300
39	torch	thread	incompatible		300
40	maize	scarf	incompatible		300
41	tap	owl	incompatible		300
42	lamp	bike	incompatible		300
43	broom	snail	incompatible		1200
44	car	book	incompatible		1200
45	slide	rope	incompatible		1200
46	milk	ring	incompatible		1200
47	crown	nose	incompatible		1200
48	phone	wine	incompatible		1200
49	weights	stairs	incompatible		300
50	toast	juice	incompatible		300
51	shell	tie	incompatible		300
52	jug	pram	incompatible		300
53	duck	cap	incompatible		300
54	rock	bird	incompatible		300
55	pail	wig	incompatible		1200
56	tray	palm	incompatible		1200
57	plate	leaf	incompatible		1200
58	thumb	bee	incompatible		1200
59	lid	bat	incompatible		1200
60	bowl	pen	incompatible		1200
61	skate	chess	incompatible		300
62	grape	cone	incompatible		300
63	globe	couch	incompatible		300
64	shoe	paint	incompatible		300
65	cross	dress	incompatible		300
66	sword	fence	incompatible		300
67	hose	cork	incompatible		1200

Item number	left object	right object	TrialType	neutral word	SOA
68	rug	pan	incompatible		1200
69	fish	bridge	incompatible		1200
70	drum	hook	incompatible		1200
71	lime	spoon	incompatible		1200
72	board	cup	incompatible		1200
73	bread	chain	compatible		300
74	mouse	lock	compatible		300
75	suit	card	compatible		300
76	seat	box	compatible		1200
77	bone	fan	compatible		1200
78	brow	jar	compatible		1200
79	belt	tin	compatible		1200
80	frog	whisk	compatible		1200
81	peach	dice	compatible		1200
82	bed	tree	compatible		300
83	pill	can	compatible		300
84	frame	rose	compatible		300
85	scale	match	neutral	cube	300
86	plane	coat	neutral	shorts	300
87	hat	salt	neutral	van	300
88	pipe	beer	neutral	pig	1200
89	knob	clasp	neutral	ship	1200
90	bolt	ant	neutral	house	1200
91	jam	doll	neutral	sand	1200
92	sack	worm	neutral	crib	1200
93	bean	nest	neutral	crane	1200
94	shelf	moon	neutral	braid	300
95	mask	straw	neutral	dart	300
96	bus	ear	neutral	vest	300
97	trunk	stool	incompatible		300
98	dice	clasp	incompatible		300
99	belt	tray	incompatible		300
100	tent	sack	incompatible		300
101	sink	jeans	incompatible		300
102	mouse	nose	incompatible		300
103	lock	cheese	incompatible		1200
104	boot	rock	incompatible		1200
105	tape	key	incompatible		1200
106	salt	disk	incompatible		1200
107	cap	bean	incompatible		1200
108	wheel	gate	incompatible		1200

Item number	left object	right object	TrialType	neutral word	SOA
109	broom	ant	incompatible		300
110	chess	hinge	incompatible		300
111	skate	vase	incompatible		300
112	fan	gun	incompatible		300
113	pin	thumb	incompatible		300
114	cab	palm	incompatible		300
115	thread	drop	incompatible		1200
116	screw	purse	incompatible		1200
117	hook	brow	incompatible		1200
118	frog	cone	incompatible		1200
119	bulb	yarn	incompatible		1200
120	plane	bridge	incompatible		1200
121	tree	car	incompatible		300
122	wood	bird	incompatible		300
123	pear	bib	incompatible		300
124	rim	whip	incompatible		300
125	bug	glue	incompatible		300
126	leek	clamp	incompatible		300
127	pen	tap	incompatible		1200
128	stone	clock	incompatible		1200
129	mask	sword	incompatible		1200
130	globe	spade	incompatible		1200
131	crisps	scoop	incompatible		1200
132	bread	pot	incompatible		1200
133	kite	rake	incompatible		300
134	sponge	lime	incompatible		300
135	toast	fork	incompatible		300
136	cake	hat	incompatible		300
137	beer	pie	incompatible		300
138	bolt	sock	incompatible		300
139	bell	can	incompatible		1200
140	box	scale	incompatible		1200
141	moth	drain	incompatible		1200
142	whisk	grape	incompatible		1200
143	suit	card	incompatible		1200
144	shot	worm	incompatible		1200
145	cross	rose	incompatible		300
146	mouth	coat	incompatible		300
147	pipe	ear	incompatible		300
148	seat	case	incompatible		300
149	spring	match	incompatible		300

Item number	left object	right object	TrialType	neutral word	SOA
150	doll	mug	incompatible		300
151	nest	duck	incompatible		1200
152	frame	dress	incompatible		1200
153	fish	plant	incompatible		1200
154	cage	throne	incompatible		1200
155	bee	toe	incompatible		1200
156	comb	snail	incompatible		1200
157	pan	owl	incompatible		300
158	couch	knob	incompatible		300
159	tin	bone	incompatible		300
160	cork	sweets	incompatible		300
161	jar	pill	incompatible		300
162	bench	shelf	incompatible		300
163	juice	straw	incompatible		1200
164	tooth	chain	incompatible		1200
165	book	jug	incompatible		1200
166	spoon	peach	incompatible		1200
167	jam	bow	incompatible		1200
168	cane	pram	incompatible		1200
169	plate	wine	compatible		300
170	string	watch	compatible		300
171	lid	rice	compatible		300
172	egg	tea	compatible		1200
173	desk	boat	compatible		1200
174	pants	harp	compatible		1200
175	fly	soap	compatible		1200
176	bike	flag	compatible		1200
177	paint	lamp	compatible		1200
178	saw	robe	compatible		300
179	net	tie	compatible		300
180	shoe	milk	compatible		300
181	drill	rug	neutral	cube	300
182	shell	flame	neutral	van	300
183	hose	axe	neutral	shorts	300
184	nail	stamp	neutral	wing	1200
185	fence	rope	neutral	braid	1200
186	leaf	ring	neutral	dart	1200
187	crown	shirt	neutral	sand	1200
188	glove	brush	neutral	crib	1200
189	scarf	torch	neutral	ship	1200
190	phone	chair	neutral	crane	300

Item number	left object	right object	TrialType	neutral word	SOA
191	cup	ball	neutral	house	300
192	foot	sign	neutral	vest	300
193	frog	clamp	incompatible		300
194	tree	bird	incompatible		300
195	net	moon	incompatible		300
196	pen	egg	incompatible		300
197	plant	chair	incompatible		300
198	toast	fence	incompatible		300
199	cross	match	incompatible		1200
200	drill	bow	incompatible		1200
201	cage	straw	incompatible		1200
202	shell	disk	incompatible		1200
203	box	spring	incompatible		1200
204	peach	scoop	incompatible		1200
205	slide	jeans	incompatible		300
206	phone	bridge	incompatible		300
207	cup	book	incompatible		300
208	hose	scarf	incompatible		300
209	pill	bat	incompatible		300
210	mug	hook	incompatible		300
211	lamp	tie	incompatible		1200
212	crown	scale	incompatible		1200
213	gate	beer	incompatible		1200
214	watch	rose	incompatible		1200
215	moth	robe	incompatible		1200
216	mouse	shirt	incompatible		1200
217	string	lock	incompatible		300
218	doll	tap	incompatible		300
219	tape	ear	incompatible		300
220	wine	ring	incompatible		300
221	whisk	rake	incompatible		300
222	rim	pail	incompatible		300
223	jam	toe	incompatible		1200
224	stone	weights	incompatible		1200
225	fish	sign	incompatible		1200
226	saw	grape	incompatible		1200
227	frame	sink	incompatible		1200
228	yarn	soap	incompatible		1200
229	chess	maize	incompatible		300
230	pear	glue	incompatible		300
231	key	ball	incompatible		300

Item number	left object	right object	TrialType	neutral word	SOA
232	shelf	bench	incompatible		300
233	dress	nose	incompatible		300
234	bread	tin	incompatible		300
235	nest	owl	incompatible		1200
236	crisps	spade	incompatible		1200
237	pan	rug	incompatible		1200
238	sack	ant	incompatible		1200
239	paint	leaf	incompatible		1200
240	sweets	hinge	incompatible		1200
241	sock	worm	incompatible		300
242	pot	chain	incompatible		300
243	torch	thread	incompatible		300
244	tea	rock	incompatible		300
245	flame	brush	incompatible		300
246	cake	boat	incompatible		300
247	belt	rice	incompatible		1200
248	skate	harp	incompatible		1200
249	stamp	bulb	incompatible		1200
250	flag	trunk	incompatible		1200
251	shoe	desk	incompatible		1200
252	plate	shot	incompatible		1200
253	bowl	fan	incompatible		300
254	suit	plane	incompatible		300
255	cane	pram	incompatible		300
256	lid	bean	incompatible		300
257	bike	salt	incompatible		300
258	bolt	fly	incompatible		300
259	brow	jar	incompatible		1200
260	can	pie	incompatible		1200
261	couch	purse	incompatible		1200
262	axe	squash	incompatible		1200
263	seat	glass	incompatible		1200
264	pipe	wheel	incompatible		1200
265	kite	wreath	compatible		300
266	screw	cork	compatible		300
267	cone	broom	compatible		300
268	bib	jug	compatible		1200
269	bag	wood	compatible		1200
270	clock	cheese	compatible		1200
271	swan	globe	compatible		1200
272	throne	fork	compatible		1200

Item number	left object	right object	TrialType	neutral word	SOA
273	tent	stool	compatible		1200
274	sword	juice	compatible		300
275	lime	spoon	compatible		300
276	drop	knife	compatible		300
277	cab	duck	neutral	shorts	300
278	palm	tray	neutral	cube	300
279	bug	whip	neutral	van	300
280	gun	boot	neutral	pig	1200
281	clasp	knob	neutral	ship	1200
282	mask	rope	neutral	wing	1200
283	comb	snail	neutral	dart	1200
284	glove	nail	neutral	crib	1200
285	hat	milk	neutral	crane	1200
286	sponge	vase	neutral	lamb	300
287	dice	sieve	neutral	vest	300
288	pin	thumb	neutral	sand	300

A.2. Materials for Experiments S1 and S2

	Item	sentences	objects
1a	version 1	Der Mann raucht vermutlich die Pfeife. *The man smokes presumably the pipe* 'The man will presumably smoke the pipe'	man (*character*), pipe (*target 1*), knife (*target 2*),
	version 2	Der Mann schärft noch heute das Messer. *The man sharpens still today the knife* 'The man will sharpen the knife today'	cane, coat, bottle screw, hat
1b	version 1	Der Vater raucht vermutlich die Zigarette. *The man smokes presumably the cigarette* 'The man will presumably smoke the cigarette'	man, cigarette, scissors,
	version 2	Der Vater schärft noch heute die Schere. *The man sharpens still today the scissors* 'The man will sharpen the scissors today'	dustpan, helmet, waistcoat, calculator
2a	version 1	Die Hausfrau spült gerade das Besteck. *The housewife washes just now the silverware* 'The housewife is just now cleaning the silverware'	woman, silverware, cauliflower,
	version 2	Die Hausfrau kocht gerade den Blumenkohl. *The housewife cooks just now the cauliflower* 'The housewife is just now cooking the cauliflower'	oven glove, hand brush, tablecloth, ironing board
2b	version 1	Die Mutter spült wohl den Teller. *The mother washes perhaps the plate* 'The mother will perhaps clean the plate'	woman, plate, egg,
	version 2	Die Mutter kocht bald das Ei. *The mother cooks soon the egg* 'The mother will soon boil the egg'	paper towels, coffee grinder, chair, broom
3a	version 1	Der Koch salzt gerade die Suppe. *The chef salts just now the soup* 'The chef is just now salting the soup '	chef, soup, glass,
	version 2	Der Koch zerbricht bestimmt das Glas. *The chef breaks certainly the glass* 'The chef will certainly break the glass'	table, grater, cheese slicer, chef's hat
3b	version 1	Die Frau salzt gerade die Nudeln. *The woman salts just now the pasta* 'The woman is just now salting the pasta'	woman, pasta, cup,
	version 2 '	Die Frau zerbricht bestimmt die Tasse. *The woman breaks certainly the cup* 'The woman will certainly break the cup'	toilette paper, plug, pram, blazer

Item	sentences version 1	sentences version 2	objects
4a	Die Studentin liest heute das Buch. *The student reads today the book* 'The student will read the book today'	Die Studentin verschliesst jetzt den Spind. *The student locks now the locker* 'The student will now lock the locker'	student, book, locker, pineapple, ring, pencil, chocolate
4b	Die Frau liest heute die Zeitung. *The woman reads today the newspaper* 'The woman will read the newspaper today'	Die Frau verschliesst jetzt die Tür. *The woman locks now the door* 'The woman will now lock the door'	woman, newspaper, door, funnel, strawberry, blanket, paintbrush
5a	Der Mann fährt demnächst das Motorrad. *The man rides shortly the motorbike* 'The man will shortly ride the motorbike.'	Der Mann isst später das Butterbrot. *The man eats later the sandwich* 'The man will later eat the sandwich.'	man, motorcycle, sandwich, record player, shirt, light bulb, saw
5b	Der Junge fährt demnächst das Fahrrad. *The boy rides shortly the bike* ' The boy will shortly ride the bike'	Der Junge isst später den Apfel *The boy eats later the apple* 'The boy will later eat the apple'	boy, bike, apple, cassette, magnifying glass , fishing rod, piggy bank
6a	Die Sekretärin verschickt bald den Brief. *The secretary dispatches soon the letter* 'The secretary will soon dispatch the letter'	Die Sekretärin trinkt später den Kaffee. *The secretary drinks later the coffee* 'The secretary will later drink the coffee'	woman, coffee, letter, stapler, paper clip, laptop, folder
6b	Der Großvater verschickt bald das Päckchen. *The grandfather dispatches soon the parcel* 'The grandfather will soon dispatch the parcel'	Der Großvater trinkt später das Bier *The grandfather drinks later the beer* 'The grandfather will later drink the beer'	man, parcel, beer, hammer, binoculars , folding ruler, glasses
7a	Der Onkel grillt bestimmt das Würstchen. *The uncle grills certainly the sausage* ' The uncle will certainly grill the sausage'	Der Onkel pflanzt später den Strauch. *The uncle plants later the bush* 'The uncle will later plant the bush'	man, sausage, bush, lawn chair, chain saw , windmill, sunshade
7b	Der Vater grillt heute den Maiskolben. *The father grills today the corn* 'The father will grill the corn today'	Der Vater pflanzt demnächst den Baum. *The father plants shortly the tree* 'The father will shortly plant the tree'	man, corn cob, tree, salad, cobweb, sunbed, bottle

Item	sentences version 1	sentences version 2	objects
8a	Die Tante giesst gerade die Pflanze. *The aunt waters just now the plant* 'The aunt is just now watering the plant'	Die Tante flickt vermutlich den Pullover. *The aunt mends presumably the sweater* 'The aunt will presumably mend the sweater'	woman, plant, sweater, comb, hall clock, tv, cupboard
8b	Die Frau giesst gerade den Kaktus. *The woman waters just now the cactus* 'The woman is just now watering the cactus'	Die Frau flickt vermutlich die Hose. *The woman mends presumably the trousers* 'The woman will presumably mend the trousers'	woman, cactus, trousers, necklace, cake, cd, bed
9a	Die Urlauberin packt schon bald den Koffer. *The vacationist packs already soon the suitcase* 'The vacationist will soon pack the suitcase '	Die Urlauberin verpasst gleich das Flugzeug. *The vacationist misses now the plane* 'The vacationist will now miss the plane'	woman, suitcase, plane, longdrink, palm , lake, hotdog
9b	Der Manager packt schon bald die Aktentasche. *The manager packs already soon the briefcase* 'The manager will soon pack the briefcase'	Der Manager verpasst gleich den Zug. *The manager misses now the train* 'The manager will now miss the train'	man, briefcase, train, fan, telephone , desk , glass
10a	Die Frau öffnet bestimmt die Dose. *The woman opens certainly the can* 'The woman will certainly open the can'	Die Frau pflückt vermutlich die Kirsche. *The woman picks presumably the cherry* 'The woman will presumably pick the cherry'	woman, can, cherry, butter, colander , watering can, peeler
10b	Das Mädchen öffnet sicher das Geschenk. *The girl opens surely the present* 'The girl will surely open the present'	Das Mädchen pflückt wahrscheinlich die Blume. *The girl picks probably the flower* 'The girl will probably pick the flower'	girl, present, flower, swing, paddle pond , toy train, puzzle
11a	Die Frau bügelt heute die Bluse. *The woman irons today the blouse* 'The woman will iron the blouse today'	Die Frau brät gerade den Fisch. *The woman frys just now the fish* 'The woman is just now frying the fish'	woman, blouse, fish, rolling pin, tea pot, buttons, vase
11b	Die Frau bügelt heute den Rock. *The woman irons today the skirt* 'The woman will iron the skirt today'	Die Frau brät gerade das Kotelett. *The woman frys just now the chop* 'The woman is just now frying the chop'	woman, skirt, chop, melon, cutting board, lipstick , hand bag

Item	sentences version 1	sentences version 2	objects
12a	Der Cowboy lädt vermutlich die Pistole. *The cowboy charges presumably the pistol* 'The cowboy will presumably charge the pistol'	Der Cowboy wirft offenbar das Lasso. *The cowboy tosses apparently the lasso* 'The cowboy will apparently toss the lasso'	cowboy, pistol, lasso, fence, guitar, tent, bonfire
12b	Der Student lädt vermutlich das Handy. *The student charges presumably the cell phone* 'The student will presumably charge the cell phone'	Der Student wirft offenbar den Ball. *The student tosses apparently the ball* 'The student will apparently toss the ball'	student, cell phone, ball, hamburger, black board, backpack, office chair
13a	Der Sportler verspritzt jetzt gleich den Sekt. *The athlete splashes now soon the champagne* 'The athlete will now splash the champagne'	Der Sportler entzündet gleich die Fackel. *The athlete ignites now the torch* 'The athlete will now ignite the torch'	athlete, champagne, torch, goggles, cup, whistle, tennis racket
13b	Die Großmutter verspritzt sicher das Wasser. *The grandmother splashes surely the water* 'The grandmother will surely splash the water'	Die Großmutter entzündet jetzt die Kerze. *The grandmother ignites now the candle* 'The grandmother will now ignite the candle'	grandmother, water, candle, onion, whisk, apron, ball of wool
14a	Die Touristin unterschreibt jetzt die Postkarte. *The tourist signs now the postcard.* 'The tourist will now sign the postcard'	Die Touristin schält gerade die Banane. *The tourist peels just now the banana* 'the tourist is just now peeling the banana'	tourist, postcard, banana, camera, passport, tooth brush, sun glasses
14b	Die Hausfrau unterschreibt jetzt die Rechnung. *The housewife signs now the bill* 'The housewife will now sign the bill'	Die Hausfrau schält gerade die Karotte. *The housewife peels just now the carrot* 'The housewife is just now peeling the carrot'	woman, bill, carrot, mirror, ladder, belt, curtain
15a	Die Frau erntet schon bald den Kürbis. *The woman harvests already soon the pumpkin* 'The woman will soon harvest the pumpkin'	Die Frau schließt vermutlich das Kästchen. *The woman closes presumably the box* 'The woman will presumably close the box'	woman, pumpkin, box, hammock, lawn mower, cookie, rake
15b	Die Frau erntet schon bald die Bohnen. *The woman harvests already soon the beans* 'The woman will soon harvest the pumpkin'	Die Frau schliesst vermutlich das Fenster. *The woman closes presumably the window* 'The woman will presumably close the window'	woman, beans, window, shelf, bathtub, napkin, landing net

A.3. Materials Experiment W

Item	Sentence	scene objects
1	Zu sehen war die Sauce/der Ketchup und auch das Möbelstück. *There was the sauce/the ketchup and also the furniture*	ketchup, tractor, tricycle, fern, bib, juice
2	Hast Du das Gebäck/den Keks gesehen und das runde Objekt? *Did you see the pastry/the cookie and the round object?*	cookie, boot, football, tea pot, spoon,lichees
3	Zu sehen war das Gemüse/die Tomate und die Blume. *There was the vegetable/the tomato and the flower.*	tomato, cow, jacket, arm, clarinette, wine
4	Zu sehen war das Spielzeug/das Puzzle und das orange Objekt. *There was the toy/the puzzle and the orange object.*	puzzle, chicken, camper van, toast, pig, cucumber
5	Hast Du das Sportgerät/den Federball und das blaue Objekt gesehen? *Did you see the sports equipment/the shuttlecock and the blue object?*	shuttlecock, tea, polar bear,guitar, donut, jam
6	Zu sehen war das Insekt/die Fliege und auch das rote Objekt. *There was the insect/the fly and also the red object.*	fly, crown, pear, tree whisk, pocket watch
7	Erinnerst Du Dich an die Pflanze/der Kaktus und an das schwarze Objekt? *Do you remember the plant/the cactus and the black object?*	caktus, coat, ball, telephone, screw driver, roll
8	Erinnerst Du Dich an das Kleidungsstück/den Handschuh und an die Tür? *Do you remember the clothing item and the door?*	glove, tennis racket, bulb, leek, pepper mill, eggcup
9	Hast Du das Essen/die Suppe gesehen und das quadratische Objekt? *Have you seen the food/the soup and the quadratic object?*	soup, pistol, hourglass, spade, owl, balloon
10	Zu sehen war das Tier/das Pferd und die Schere. *There was the animal/the horse and the scissors.*	horse, beer, sword potato, leg, cream
11	Erinnerst Du Dich an den Körperteil/das Auge und das viereckige Objekt? *Do you remember the part of the body/the eye and the square object?*	eye, clover leaf, ship, pocketknife hot-water bottle, flag
12	Erinnerst Du Dich an den Vogel/die Ente und an das grüne Objekt? *Do you remember the bird/the duck and the green object?*	duck, dresser, pliers, croissant watering can, dress

Item	Sentence	scene objects
13	Erinnerst Du Dich an das Geschirr/den Teller und auch an die Laterne? *Do you remember the tableware/the plate and also the lantern?*	plate, ladybird, licorice, penguin, strawberry, rubber duck
14	Du hast sicher das Stofftier/den Teddy bemerkt und das Küchengerät. *You have certainly noticed the stuffed animal/the teddy and the kitchen device*	teddy saddle, cup, scarf coffee mill, tie
15	Du hast das Haustier/die Katze gesehen und das Gänseblümchen. *You have seen the pet/the cat and the daisy.*	cat, crash helmet, drum, asparagus, milk, rubber boot
16	Erinnerst Du Dich an das Getränk/den Kaffee und an das dreieckige Objekt? *Do you remember the beverage/the coffee and the triangular object?*	coffee, finger, bus, tiger, onion, sock
17	Zu sehen war das Obst/die Kirsche und außerdem das längliche Objekt. *There was the fruit/the cherry and also the longish object.*	cherry, bee, hedgehog, pan screw, helicopter
18	Hast Du das Gebäude/die Kirche gesehen und die Tasse? *Did you see the building/the church and the cup?*	church, mushroom, sewing machine, cup, stag, grater
19	Hast Du das Elektrogerät/die Waschmaschine und den Schuh gesehen? *Did you see the electric appliance/the washing machine and the shoe?*	washing machine, violin, french fries, sandal, peach, flag
20	Du hast bestimmt die Frucht/die Kiwi bemerkt und das runde Objekt. *You have probably noticed the fruit/the kiwi and the round object.*	kiwi, clothespin, doll, iron clock, sea horse
21	Du hast bestimmt das Genussmittel/die Schokolade und das blaue Objekt bemerkt. *You have probably noticed the semiluxury food/the chocolate and the blue object.*	chocolate, bag, leather jacket, lion, worm, dog house
22	Hast Du den Nachtisch/den Obstsalat und das längliche Objekt gesehen? *Did you see the desert/the fruit salad and the longish object?*	fruit salad, truck, eagle, zebra wrench, house

Item	Sentence	scene objects
23	Jetzt hast Du das Gartengerät/den Rasenmäher und das gelbe Objekt gesehen. *Now you have seen the gardening tool/the lawn mower and the yellow object.*	lawn mower, banana, cupboard, elephant, pineapple, top hat
24	Du hast sicher das Sitzmöbel/den Schaukelstuhl und das grüne Objekt bemerkt. *You have probably noticed the seating furniture/the rocking chair and the green object.*	rocking chair corkscrew, dagger, bottle giraffe, parrot
25	Du hast das Gerät/den Fernseher und auch das orange Objekt gesehen. *You have seen the apparatus/the TV and also the orange object.*	TV, hamburger, pump, butterfly carrot, barrette
26	Hast Du das Knabberzeug/die Salzstangen und den Behälter gesehen? *Did you see the snack/the saltsticks and the container?*	saltsticks, lamp, tower, fence, trumpet, pitchfork
27	Hast Du das Fahrzeug/das Motorrad und das grüne Objekt gesehen? *Did you see the vehikle/the motorcycle and the green object?*	motorcycle piano, wine glass, snail, colander, nut
28	Erinnerst Du Dich an das Instrument/Saxophon und den Fisch? *Do you remember the instrument/the saxophone and the fish?*	saxophone, fish, glas, radio, bread, temple
29	Hast Du den Körperteil/den Fuss gesehen und das schwarze Objekt? *Did you see the part of the body/the foot and the black object?*	foot, dip, bed, chair castle, hat
30	Zu sehen war die Kopfbedeckung/die Mütze und das Gefäß. *There was the headdress/the hat and the vessel.*	hat, orange, hotdog wallet, box, chocolate
31	Zu sehen war das Milchprodukt/der Käse und das längliche Objekt. *There was the dairy product/the cheese and the longish object.*	cheese, elbow, slipper, ant nail, chest
32	Erinnerst Du Dich an die Lichtquelle/die Taschenlampe und an den Frosch? *Do you remember the illuminant/the flashlight and the frog?*	flashlight, apple, stool, suitcase chips, baseball cap

Appendix B.

Model Summaries of Generalized Linear Mixed Effect Models

Experiment S1

Table B.1.: Model summary for VerbEnd time region
N = 2110; log-likelihood = -876.4
$ins \sim OBJ + POS + OBJ * POS + (1|subj) + (OBJ + POS|item)$

Predictor	Coefficiant	Std. Error	z value	p-value		
(Intercept)	-2.46	0.34	-7.27	3.66e-13	***	*baseline condition*
target	1.01	0.38	2.67	0.01	**	
pos1	-0.26	0.37	-0.70	0.49		*simple*
pos2	-0.59	0.39	-1.50	0.13		*effect*
pos4	-0.51	0.39	-1.33	0.18		*terms*
pos5	-0.30	0.37	-0.81	0.42		
pos6	-0.55	0.40	-1.39	0.16		
target:pos1	0.27	0.46	0.60	0.55		
target:pos2	0.92	0.47	1.95	0.05	.	
target:pos4	0.50	0.47	1.06	0.29		*interaction*
target:pos5	0.70	0.45	1.57	0.12		*terms*
target:pos6	1.50	0.46	3.25	0.00	**	

Table B.2.: Main effect model summary for VerbEnd time region
N = 2110; log-likelihood = -882.1
$ins \sim OBJ + POS + (1|subj) + (OBJ + POS|item)$

Predictor	Coefficiant	Std. Error	z value	p-value		
(Intercept)	-2.92	0.31	-9.53	< 2e-16	***	*baseline condition*
target	1.67	0.24	6.82	9.11e-12	***	
pos1	-0.07	0.23	-0.32	0.75		*simple*
pos2	0.07	0.22	0.30	0.77		*effect*
pos4	-0.16	0.23	-0.73	0.47		*terms*
pos5	0.20	0.21	0.91	0.36		
pos6	0.54	0.23	2.38	0.02	*	

Table B.3.: Model summary for VERB time region
N = 2110; log-likelihood = -419.5
$ins \sim OBJ + POS + OBJ * POS + (1|subj) + (OBJ + POS|item)$

Predictor	Coefficiant	Std. Error	z value	p-value		
(Intercept)	-3.80	0.48	-7.85	4.15e-15	***	*baseline condition*
target	0.64	0.52	1.24	0.22		
pos1	0.09	0.51	0.18	0.86		*simple*
pos2	-1.99	0.77	-2.58	0.01	**	*effect*
pos4	-0.89	0.71	-1.26	0.21		*terms*
pos5	-0.32	0.57	-0.57	0.57		
pos6	-0.44	0.54	-0.82	0.41		
target:pos1	-0.17	0.68	-0.25	0.80		
target:pos2	1.93	0.88	2.19	0.03	*	
target:pos4	0.97	0.83	1.17	0.24		*interaction*
target:pos5	0.76	0.70	1.09	0.28		*terms*
target:pos6	1.16	0.67	1.72	0.09	.	

Table B.4.: Main effect model summary for VERB time region
N = 2110; log-likelihood = -423.7
$ins \sim OBJ + POS + (1|subj) + (OBJ + POS|item)$

Predictor	Coefficiant	Std. Error	z value	p-value		
(Intercept)	-4.23	0.45	-9.47	< 2e-16	***	*baseline condition*
target	1.27	0.29	4.38	1.17e-05	***	
pos1	-0.02	0.35	-0.06	0.95		*simple*
pos2	-0.44	0.37	-1.20	0.23		*effect*
pos4	-0.19	0.41	-0.48	0.63		*terms*
pos5	0.21	0.35	0.60	0.55		
pos6	0.39	0.32	1.20	0.23		

Table B.5.: Model summary for NP2 time region
N = 2110; log-likelihood = -674.9
$ins \sim OBJ + POS + OBJ * POS + (POS|subj) + (OBJ|item)$

Predictor	Coefficiant	Std. Error	z value	p-value		
(Intercept)	-3.23	0.39	-8.19	2.7e-16	***	*baseline condition*
target	1.22	0.45	2.68	0.01	**	
pos1	-0.52	0.55	-0.95	0.34		*simple*
pos2	0.03	0.49	0.06	0.96		*effect*
pos4	-0.15	0.50	-0.30	0.77		*terms*
pos5	-0.15	0.50	-0.30	0.76		
pos6	-0.34	0.53	-0.65	0.52		
target:pos1	0.29	0.63	0.46	0.64		
target:pos2	0.48	0.57	0.85	0.40		
target:pos4	0.13	0.58	0.23	0.82		*interaction*
target:pos5	0.69	0.58	1.19	0.23		*terms*
target:pos6	1.22	0.59	2.06	0.04	*	

Table B.6.: Main effect model summary for NP2 time region
N = 2110; log-likelihood = -677.9
$ins \sim OBJ + POS + (POS|subj) + (OBJ|item)$

Predictor	Coefficiant	Std. Error	z value	p-value		
(Intercept)	-3.57	0.30	-12.05	$<$ 2e-16	***	*baseline condition*
`target`	1.72	0.25	6.87	6.28e-12	***	
`pos1`	-0.36	0.31	-1.17	0.24		*simple*
`pos2`	0.35	0.26	1.35	0.18		*effect*
`pos4`	-0.09	0.28	-0.32	0.75		*terms*
`pos5`	0.33	0.25	1.29	0.20		
`pos6`	0.59	0.25	2.34	0.02	*	

Experiment S2

Predictor	Coefficiant	Std. Error	z value	p-value			
(Intercept)	-3.52	0.42	-8.48	<2e-16	***	}	*baseline condition*
`target`	0.35	0.49	0.71	0.48		⎫	
`pos1`	-0.32	0.55	-0.58	0.56		⎪	*simple*
`pos2`	-0.16	0.53	-0.30	0.77		⎪	*effect*
`pos4`	-0.15	0.54	-0.28	0.78		⎪	*terms*
`pos5`	-0.10	0.55	-0.19	0.85		⎪	
`pos6`	-0.68	0.58	-1.16	0.25		⎭	
`target:pos1`	1.45	0.66	2.20	0.03	*	⎫	
`target:pos2`	1.00	0.66	1.52	0.13		⎪	
`target:pos4`	1.05	0.66	1.59	0.11		⎪	*interaction*
`target:pos5`	0.95	0.68	1.41	0.16		⎪	*terms*
`target:pos6`	1.60	0.70	2.30	0.02	*	⎭	

Table B.7.: Model summary for VerbEnd time region (N = 2136; log-likelihood = -552.6)
Model: $ins \sim OBJ + POS + OBJ * POS + (OBJ + POS|subj) + (1|item)$

Predictor	Coefficiant	Std. Error	z value	p-value			
(Intercept)	-4.25	0.35	-12.19	<2e-16	***	}	*baseline condition*
`target`	1.39	0.19	7.13	1e-12	***	⎫	
`pos1`	0.75	0.31	2.42	0.02	*	⎪	*simple*
`pos2`	0.55	0.32	1.70	0.09	.	⎪	*effect*
`pos4`	0.59	0.34	1.75	0.08	.	⎪	*terms*
`pos5`	0.57	0.33	1.74	0.08	.	⎪	
`pos6`	0.53	0.32	1.64	0.10		⎭	

Table B.8.: Main effect model summary for VerbEnd time region
(N = 2136; log-likelihood = -555.8)
Model: $ins \sim OBJ + POS + (OBJ + POS|subj) + (1|item)$

Predictor	Coefficiant	Std. Error	z value	p-value			
(Intercept)	-3.59	0.43	-8.30	<2e-16	***	}	*baseline condition*
target	0.11	0.55	0.19	0.85		⎫	
pos1	-1.04	0.70	-1.49	0.14		⎪	*simple*
pos2	-1.22	0.72	-1.68	0.09	.	⎬	*effect*
pos4	-0.76	0.66	-1.16	0.25		⎪	*terms*
pos5	-0.07	0.60	-0.12	0.91		⎪	
pos6	-1.30	0.79	-1.65	0.10	.	⎭	
target:pos1	1.93	0.81	2.37	0.02	*	⎫	
target:pos2	2.00	0.81	2.48	0.01	*	⎪	
target:pos4	1.42	0.76	1.87	0.06	.	⎬	*interaction*
target:pos5	0.48	0.76	0.64	0.53		⎪	*terms*
target:pos6	1.54	0.88	1.75	0.08	.	⎭	

Table B.9.: Model summary for VERB time region (N = 2136; log-likelihood = -427)
Model: $ins \sim OBJ + POS + OBJ * POS + (POS|subj) + (OBJ + POS|item)$

Predictor	Coefficiant	Std. Error	z value	p-value			
(Intercept)	-4.37	0.37	-11.94	< 2e-16	***	}	*baseline condition*
target	1.35	0.24	5.51	3.52e-08	***	⎫	
pos1	0.33	0.36	0.91	0.36		⎪	*simple*
pos2	0.20	0.42	0.48	0.63		⎬	*effect*
pos4	0.18	0.42	0.42	0.67		⎪	*terms*
pos5	0.18	0.40	0.45	0.65		⎪	
pos6	-0.23	0.46	-0.49	0.62		⎭	

Table B.10.: Main effect model summary for VERB time region
(N = 2136; log-likelihood = -431.9)
Model: $ins \sim OBJ + POS + (POS|subj) + (OBJ + POS|item)$

Predictor	Coefficiant	Std. Error	z value	p-value			
(Intercept)	-6.48	1.27	-5.09	3.62e-07	***	}	*baseline condition*
target	-0.53	1.58	-0.34	0.73		⎫	
pos1	1.49	1.45	1.03	0.30		⎪	*simple*
pos2	1.73	1.46	1.19	0.23		⎬	*effect*
pos4	1.33	1.55	0.86	0.39		⎪	*terms*
pos5	0.85	1.72	0.50	0.62		⎪	
pos6	1.64	1.43	1.15	0.25		⎭	
target:pos1	1.75	1.73	1.01	0.31		⎫	
target:pos2	1.10	1.76	0.63	0.53		⎪	
target:pos4	2.23	1.83	1.22	0.22		⎬	*interaction*
target:pos5	2.76	1.97	1.41	0.16		⎪	*terms*
target:pos6	1.89	1.71	1.10	0.27		⎭	

Table B.11.: Model summary for NP2 time region (N = 2136; log-likelihood = -246.6)
Model: $ins \sim OBJ + POS + OBJ * POS + (OBJ + POS|subj) + (1|item)$

Predictor	Coefficiant	Std. Error	z value	p-value			
(Intercept)	-7.54	0.92	-8.21	2.22e-16	***	}	*baseline condition*
`target`	1.32	0.44	3.00	0.00	**	⎫	
`pos1`	2.51	0.89	2.81	0.00	**	⎪	*simple*
`pos2`	2.32	0.90	2.58	0.01	**	⎪	*effect*
`pos4`	2.68	0.89	3.02	0.00	**	⎪	*terms*
`pos5`	2.64	0.93	2.85	0.00	**	⎪	
`pos6`	2.74	0.87	3.15	0.00	**	⎭	

Table B.12.: Main effect model summary for NP2 time region
(N = 2136; log-likelihood = -247.8)
Model: $ins \sim OBJ + POS + (OBJ + POS|subj) + (1|item)$

Predictor	Coefficiant	Std. Error	z value	p-value			
(Intercept)	-5.74	1.03	-5.55	2.86e-08	***	}	*baseline condition*
`target`	-0.45	1.40	-0.32	0.75		⎫	
`pos1`	0.13	1.26	0.10	0.92		⎪	*simple*
`pos2`	1.53	1.16	1.32	0.19		⎪	*effect*
`pos4`	0.60	1.37	0.44	0.66		⎪	*terms*
`pos5`	-0.32	1.58	-0.20	0.84		⎪	
`pos6`	0.76	1.21	0.63	0.53		⎭	
`target:pos1`	2.34	1.57	1.49	0.14		⎫	
`target:pos2`	0.93	1.52	0.61	0.54		⎪	
`target:pos4`	2.73	1.68	1.63	0.10		⎪	*interaction*
`target:pos5`	3.47	1.85	1.87	0.06	.	⎪	*terms*
`target:pos6`	2.16	1.54	1.40	0.16		⎭	

Table B.13.: Model summary for NP2400 time region (N = 2136; log-likelihood = -275.5)
Model: $ins \sim OBJ + POS + OBJ * POS + (OBJ + POS|subj) + (1|item)$

Predictor	Coefficiant	Std. Error	z value	p-value			
(Intercept)	-7.08	0.76	-9.33	< 2e-16	***	}	*baseline condition*
`target`	1.62	0.40	4.06	4.96e-05	***	⎫	
`pos1`	1.77	0.70	2.53	0.01	*	⎪	*simple*
`pos2`	2.14	0.73	2.93	0.00	**	⎪	*effect*
`pos4`	2.48	0.75	3.29	0.00	**	⎪	*terms*
`pos5`	2.27	0.72	3.16	0.00	**	⎪	
`pos6`	2.22	0.69	3.21	0.00	**	⎭	

Table B.14.: Model summary for NP2+400 time region (N = 2136; log-likelihood = -278.3)
Model: $ins \sim OBJ + POS + (OBJ + POS|subj) + (1|item)$

Experiment S3

Table B.15.: Model summary for VerbEnd time region
N = 1440; log-likelihood = -664
$ins \sim OBJ + POS + OBJ * POS + (OBJ + POS|subj) + (OBJ + POS|item)$

Predictor	Coefficiant	Std. Error	z value	p-value	
(Intercept)	-2.38	0.34	-7.06	1.69e-12	***
`target`	0.93	0.29	3.17	0.00	**
`pos1`	0.15	0.32	0.48	0.63	
`pos6`	0.56	0.31	1.83	0.07	.
`target:pos1`	0.05	0.37	0.14	0.89	
`target:pos6`	-0.08	0.35	-0.24	0.81	

Table B.16.: Main effect model summary for VerbEnd time region
N = 1440; log-likelihood = -664.1
$ins \sim OBJ + POS + OBJ * POS + (OBJ + POS|subj) + (OBJ + POS|item)$

Predictor	Coefficiant	Std. Error	z value	p-value	
(Intercept)	-2.37	0.31	-7.72	1.19e-14	***
`target`	0.91	0.19	4.79	1.67e-06	***
`pos1`	0.20	0.23	0.86	0.39	
`pos6`	0.51	0.22	2.30	0.02	*

Table B.17.: Model summary for Verb time region
N = 1440; log-likelihood = -502.6
$ins \sim OBJ + POS + OBJ * POS + (OBJ + POS|subj) + (OBJ|item)$

Predictor	Coefficiant	Std. Error	z value	p-value	
(Intercept)	-2.82	0.30	-9.28	<2e-16	***
`target`	0.50	0.33	1.51	0.13	
`pos1`	0.07	0.37	0.19	0.85	
`pos6`	0.46	0.33	1.40	0.16	
`target:pos1`	0.19	0.45	0.43	0.67	
`target:pos6`	0.11	0.42	0.26	0.80	

Table B.18.: Main effect model summary for VERB time region
N = 1440; log-likelihood = -503.1
$ins \sim OBJ + POS + OBJ * POS + (OBJ + POS|subj) + (OBJ|item)$

Predictor	Coefficiant	Std. Error	z value	p-value	
(Intercept)	-2.88	0.25	-11.41	< 2e-16	***
`target`	0.62	0.18	3.41	0.00	***
`pos1`	0.19	0.23	0.81	0.42	
`pos6`	0.52	0.22	2.40	0.02	*

Table B.19.: Model summary for NP2 time region
N = 1440; log-likelihood = -410.7
$ins \sim OBJ + POS + OBJ * POS + (OBJ + POS|subj) + (OBJ + POS|item)$

Predictor	Coefficiant	Std. Error	z value	p-value	
(Intercept)	-3.60	0.44	-8.27	< 2e-16	***
`target`	1.13	0.44	2.59	0.01	**
`pos1`	-0.69	0.48	-1.44	0.15	
`pos6`	-0.07	0.48	-0.15	0.88	
`target:pos1`	1.07	0.55	1.93	0.05	.
`target:pos6`	0.70	0.52	1.33	0.18	

Table B.20.: Main effect model summary for NP2 time region
N = 1440; log-likelihood = -412
$ins \sim OBJ + POS + OBJ * POS + (OBJ + POS|subj) + (OBJ + POS|item)$

Predictor	Coefficiant	Std. Error	z value	p-value	
(Intercept)	-4.09	0.43	-9.56	< 2e-16	***
`target`	1.69	0.33	5.17	2.29e-07	***
`pos1`	0.28	0.27	1.00	0.32	
`pos6`	0.53	0.34	1.57	0.12	

Table B.21.: Model summary for VERBEXACT time region
N = 1440; log-likelihood = -306.6)
$ins \sim OBJ + POS + OBJ * POS + (OBJ + POS|subj) + (OBJ|item)$

Predictor	Coefficiant	Std. Error	z value	p-value	
(Intercept)	-3.35	0.38	-8.80	<2e-16	***
`target`	-0.27	0.50	-0.55	0.58	
`pos1`	-0.85	0.59	-1.43	0.15	
`pos6`	0.46	0.43	1.07	0.28	
`target:pos1`	1.06	0.70	1.52	0.13	
`target:pos6`	0.74	0.59	1.25	0.21	

Table B.22.: Main effect model summary for VerbExact time region
N = 1440; log-likelihood = -307.7
$ins \sim OBJ + POS + OBJ * POS + (OBJ + POS|subj) + (OBJ|item)$

Predictor	Coefficiant	Std. Error	z value	p-value	
(Intercept)	-3.70	0.37	-10.05	< 2e-16	***
target	0.34	0.28	1.22	0.22	
pos1	-0.26	0.47	-0.55	0.58	
pos6	0.87	0.33	2.63	0.01	**

Table B.23.: Model summary primacy test for VerbExact time region
N = 960; log-likelihood = -169.3
$ins \sim OBJ + POS + OBJ * POS + (OBJ + POS|subj) + (1|item)$

Predictor	Coefficiant	Std. Error	z value	p-value	
(Intercept)	-3.43	0.38	-8.93	< 2e-16	***
target	-0.52	0.54	-0.96	0.33	
pos1	-1.29	0.67	-1.93	0.05	.
target:pos1	2.00	0.77	2.59	0.01	**

Table B.24.: Model summary recency test for VerbExact time region
N = 960; log-likelihood = -221.1
$ins \sim OBJ + POS + OBJ * POS + (OBJ + POS|subj) + (OBJ|item)$

Predictor	Coefficiant	Std. Error	z value	p-value	
(Intercept)	-3.32	0.37	-8.90	<2e-16	***
target	-0.56	0.53	-1.07	0.28	
pos6	0.48	0.44	1.11	0.27	
target:pos6	0.77	0.60	1.28	0.20	

Table B.25.: Model summary for AdvExact time region
N = 1440; log-likelihood = -410.3
$ins \sim OBJ + POS + OBJ * POS + (OBJ + POS|subj) + (POS|item)$

Predictor	Coefficiant	Std. Error	z value	p-value	
(Intercept)	-3.79	0.42	-8.99	< 2e-16	***
target	1.02	0.42	2.40	0.02	*
pos1	0.82	0.45	1.81	0.07	.
pos6	1.18	0.43	2.74	0.01	**
target:pos1	-0.33	0.54	-0.60	0.55	
target:pos6	-0.53	0.52	-1.03	0.30	

Table B.26.: Main effect model summary for AdvExact time region
N = 1440; log-likelihood = -410.8
$ins \sim OBJ + POS + (OBJ + POS|subj) + (POS|item)$

Predictor	Coefficiant	Std. Error	z value	p-value	
(Intercept)	-3.55	0.32	-11.09	< 2e-16	***
`target`	0.67	0.20	3.33	0.00	***
`pos1`	0.60	0.26	2.26	0.02	*
`pos6`	0.83	0.26	3.23	0.00	**

Table B.27.: Model summary primacy test for AdvExact time region
N = 960; log-likelihood = -250.1
$ins \sim OBJ + POS + OBJ * POS + (OBJ + POS|subj) + (OBJ + POS|item)$

Predictor	Coefficiant	Std. Error	z value	p-value	
(Intercept)	-4.20	0.50	-8.34	< 2e-16	***
`target`	1.49	0.47	3.21	0.00	**
`pos1`	1.08	0.50	2.15	0.03	*
`target:pos1`	-0.57	0.57	-1.00	0.32	

Table B.28.: Model summary of recency test for AdvExact time region
N = 960; log-likelihood = -273
$ins \sim OBJ + POS + OBJ * POS + (OBJ + POS|subj) + (OBJ + POS|item)$

Predictor	Coefficiant	Std. Error	z value	p-value	
(Intercept)	-3.86	0.44	-8.87	< 2e-16	***
`target`	0.93	0.45	2.07	0.04	*
`pos6`	1.26	0.44	2.84	0.00	**
`target:pos6`	-0.51	0.52	-0.98	0.32	

Table B.29.: Model summary for NounExact time region
N = 1440; log-likelihood = -299.7
$ins \sim OBJ + POS + OBJ * POS + (OBJ + POS|subj) + (OBJ + POS|item)$

Predictor	Coefficiant	Std. Error	z value	p-value	
(Intercept)	-3.90	0.44	-8.87	<2e-16	***
`target`	1.03	0.52	1.97	0.05	*
`pos1`	-0.86	0.57	-1.52	0.13	
`pos6`	-1.09	0.62	-1.75	0.08	.
`target:pos1`	0.57	0.66	0.86	0.39	
`target:pos6`	1.43	0.68	2.11	0.03	*

Table B.30.: Main effect model summary for NounExact time region
N = 1440; log-likelihood = -300.8
$ins \sim OBJ + POS + OBJ * POS + (OBJ + POS|subj) + (OBJ + POS|item)$

Predictor	Coefficiant	Std. Error	z value	p-value	
(Intercept)	-4.41	0.43	-10.30	< 2e-16	***
`target`	1.62	0.40	4.08	4.6e-05	***
`pos1`	-0.22	0.34	-0.64	0.52	
`pos6`	0.10	0.35	0.29	0.77	

Table B.31.: Model summary primacy effect for NounExact time region
N = 960; log-likelihood = -188.2
$ins \sim OBJ + POS + OBJ * POS + (OBJ + POS|subj) + (OBJ + POS|item)$

Predictor	Coefficiant	Std. Error	z value	p-value	
(Intercept)	-3.94	0.47	-8.46	<2e-16	***
`target`	0.76	0.60	1.27	0.20	
`pos1`	-0.93	0.62	-1.51	0.13	
`target:pos1`	1.05	0.72	1.46	0.14	

Table B.32.: Model summary recency test for NounExact time region
N = 960; log-likelihood = -211.7
$ins \sim OBJ + POS + OBJ * POS + (OBJ + POS|subj) + (OBJ + POS|item)$

Predictor	Coefficiant	Std. Error	z value	p-value	
(Intercept)	-3.75	0.42	-9.02	< 2e-16	***
`target`	0.73	0.51	1.43	0.15	
`pos6`	-1.40	0.67	-2.08	0.04	*
`target:pos6`	1.95	0.71	2.73	0.01	**