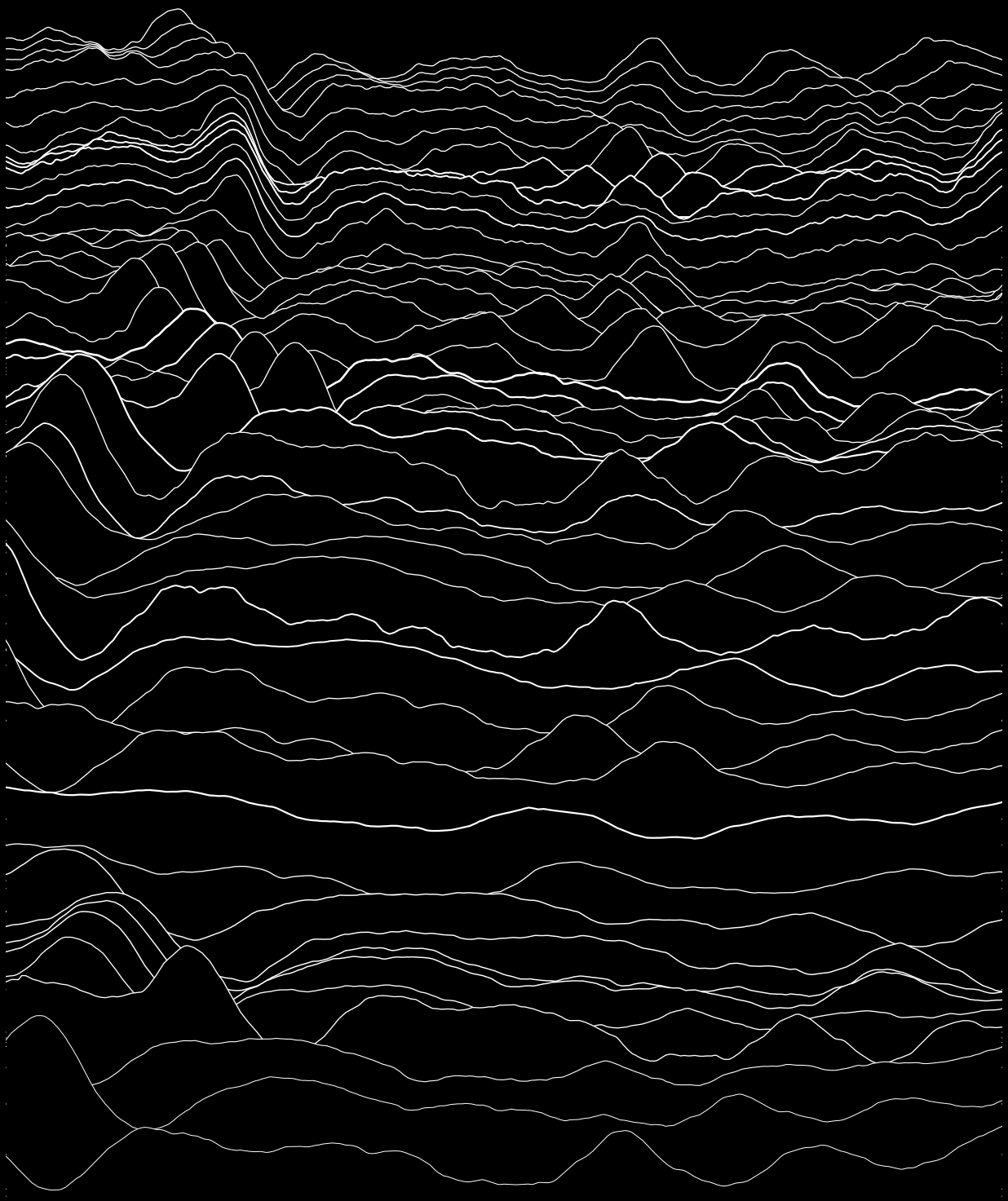
EXPECTATION-BASED RETRIEVAL AND INTEGRATION IN LANGUAGE COMPREHENSION



CHRISTOPH AURNHAMMER

Expectation-Based Retrieval and Integration in Language Comprehension

Dissertation

zur Erlangung des akademischen Grades eines

Doktors der Philosophie

an der Philosophischen Fakultät der
Universität des Saarlandes

Vorgelegt von

Christoph Aurnhammer

geb. 18.10.1992, Ulm

Saarbrücken, 2024

UNIVERSITÄT
DES
SAARLANDES

Betreuer:	Prof. Dr. Matthew W. Crocker Dr. Harm Brouwer
Erstgutachter:	Prof. Dr. Matthew W. Crocker
Zweitgutachter:	Prof. Dr. Axel Mecklinger
Drittgutacherin:	Prof. Dr. Seana Coulson
Dekanin der Philosophischen Fakultät:	Prof. Dr. Stefanie Haberzettl
Datum der Disputation:	19. Januar 2024
Promotionskommission:	Prof. Dr. Ingo Reich Prof. Dr. Matthew W. Crocker Prof. Dr. Axel Mecklinger Prof. Dr. Seana Coulson Prof. Dr. Bernd Möbius Dr. Francesca Delogu

"I have no idea where this will lead us, but I have a definite feeling it will be a place both wonderful and strange."

Agent Dale Cooper in Twin Peaks (Frost et al., 1991)

Abstract

To understand language, comprehenders must *retrieve* the meaning associated with the words they perceive from memory and they must *integrate* retrieved word meanings into a representation of utterance meaning. During incremental comprehension, both processes are constrained by what has been understood so far and hence are expectation-based mechanisms. Psycholinguistic experiments measuring the electrical activity of the brain have provided key evidence that may elucidate how the language comprehension system organises and implements expectation-based retrieval and integration. However, the field has converged neither on a generally accepted formalisation of these processes nor on their mapping to the two most salient components of the event-related potential signal, the N400 and the P600.

Retrieval-Integration theory offers a mechanistic account of the underpinnings of language comprehension and posits that retrieval is indexed by the N400 and integration is indexed by the P600. Following these core assumptions, this thesis demonstrates the expectation-based nature of language comprehension in which both retrieval (N400) and integration (P600) are influenced by expectations derived from an incrementally constructed utterance meaning representation. Critically, our results also indicate that lexical association to the preceding context modulates the N400 but not the P600, affirming the relation of the N400 to retrieval, rather than to integration. Zooming in on the role of integration, we reveal an important novel dimension to the interpretation of the P600 by demonstrating that P600 amplitude – and not N400 amplitude – is continuously related to utterance meaning plausibility. Finally, we examine the single-trial dynamics of retrieval and integration, establishing that words that are more effortful to retrieve tend to be more effortful to integrate, as evidenced by a within-trial correlation of N400 and P600 amplitude.

These results are in direct opposition to traditional and more recent proposals arguing that (1) the N400 indexes integration processes, (2) integration – as indexed by the N400 – is merely "quasi-compositional", and (3) the P600 is a reflection of conflicting interpretations generated in a multi-stream architecture. Rather, our findings indicate that (1) integration is continuously indexed by the P600, (2) integration is fully compositional, and (3) a single-stream architecture in which the N400 continuously indexes retrieval and the P600 continuously indexes integration is sufficient to account for the key ERP data. We conclude that retrieval and integration are two central mechanisms underlying language processing and that the N400 and the P600 should be considered part of the default ERP signature of utterance comprehension. Future study of expectation-based language processing should adopt a comprehension-centric view on expectancy and hence focus on integration effort, as indexed by the P600.

Zusammenfassung in deutscher Sprache

Um Sprache zu verstehen, müssen Menschen die Bedeutung einzelner Worte *abrufen* und sie müssen die Bedeutungen dieser Worte in eine Bedeutungsrepräsentation der Äußerung *integrieren*. Diese Prozesse erfolgen inkrementell: Mehr oder weniger jedes wahrgenommene Wort eines Satzes wird sofort einem Bedeutungsabrufungsprozess unterzogen und die abgerufene Wortbedeutung wird in die Äußerungsbedeutung integriert. Die inkrementelle Sprachverarbeitung ist dabei nicht allein von den wahrgenommen Informationen bestimmt sondern stark erwartungsbasiert: Das bislang Verstandene weckt Erwartungen darüber, was als nächstes kommuniziert wird. Zum Beispiel erleichtert das Verarbeiten des Teilsatzes „Gestern schärfte der Holzfäller die ...“ die Bedeutungsabrufung und Bedeutungsintegration für das Wort „Axt“ (Beispiel aus Kapitel 3). Lautet der Teilsatz jedoch „Gestern *aß* der Holzfäller die ...“ sollte keine Erleichterung für Abrufung und Integration des Wortes „Axt“ gegeben sein.

Zentraler Baustein hierfür ist die inkrementell erstellte Bedeutungsrepräsentation des Teilsatzes. Die Teilsatzbedeutung kann mögliche zukünftige Wortbedeutungen voraktivieren und dadurch deren Abrufung erleichtern. Ebenso kann die bislang erstellte Bedeutung der Äußerung die Integration von Wortbedeutungen in die angepasste Äußerungsbedeutung erleichtern, wenn die neuen Informationen dem Weltwissen gemäß erwartbar sind. Der Einfluss der Bedeutungsrepräsentation einer Äußerung auf Abrufung und Integration lässt sich mit dem generellen Begriff der *Erwartbarkeit* eines Wortes beschreiben. Diese Dissertation fußt auf der Annahme, dass das Sprachverständnis maßgeblich durch die erwartungsbasierten Prozesse der Bedeutungsabrufung und Bedeutungsintegration geprägt ist. Wenn diese beiden Prozesse tatsächlich maßgebliche Bestandteile des Sprachverständnisses sind, stellt sich die Frage, wie der kognitive Aufwand der Abrufung und der Integration gemessen werden kann.

Ein vielversprechender Ansatz um zu verstehen, wie Menschen Bedeutung abrufen und integrieren, wäre es, die „Hardware“, welche diese kognitiven Prozesse implementiert – nämlich das menschliche Gehirn – direkt zu messen, während Versuchspersonen Sprache verarbeiten. In der Tat wurden entscheidende Erkenntnisse über das Wie und Wann des Sprachverständnisses im Gehirn durch die Messung ereigniskorrelierter Potentiale (EKP) gewonnen. EKP werden aus dem Elektroenzephalogramm (EEG) berechnet und offenbaren die auf der Kopfhaut gemessene elektrische Aktivität des Gehirns im Verlauf der Zeit nach der Präsentation eines Stimulus. In den Experimenten, welche für diese Arbeit durchgeführt wurden, werden als Stimuli einzelne Worte, welche zusammen einen Satz formen, präsentiert. Dadurch lässt sich zum Beispiel das EKP erwartbarer Worte mit jenem nicht erwartbarer Worte vergleichen („Gestern [schärfte/aß] der Holzfäller die Axt“). Unterschiede in der Erwartbarkeit eines Wortes gehen im EKP – unter anderem – mit

Unterschieden in der Amplitude sogenannter EKP-Komponenten, zeitlich abgegrenzter Teile des EKPs, einher. Zwei EKP-Komponenten haben im Besonderen zu wichtigen Erkenntnissen für die Erforschung des Sprachverständnisses geführt: Die N400-Komponente, ein negativer Ausschlag des EKPs, welcher etwa 400 Millisekunden nach der Präsentation eines Stimulus seine maximale Amplitude erreicht, und die P600-Komponente, eine anhaltende, positive Abweichung des Signals, welche etwa ab 600 Millisekunden nach der Präsentation des Stimulus sichtbar wird. Seit der Entdeckung dieser EKP-Komponenten hat die elektrophysiologische Forschung die Sensitivität beider Komponenten hinsichtlich verschiedener sprachlicher sowie nicht-sprachlicher Variablen untersucht. Trotz der Vielzahl der EKP-Resultate, welche innerhalb der Sprachverarbeitungsforschung vorgelegt wurden, ist das Forschungsfeld weder bei einer allgemein anerkannten formellen Beschreibung der zum Sprachverständnis notwendigen Prozesse (z.B. Abrufung und Integration) noch zu einer unumstrittenen Zuordnung dieser Prozesse zu EKP-Komponenten (z.B. N400 und P600) angelangt. Die daraus resultierende Ungewissheit behindert Fortschritte in der Beschreibung der neurokognitiven Implementation des Sprachverständnisses, was in der Konsequenz die effektive Entwicklung experimenteller Sprachstudien sowie deren eindeutige Auswertung erschwert.

Zur Lösung dieses Problems können komputationale Modelle des Sprachverständnisprozesses entwickelt werden, welche, erstens, die enthaltenen Prozesse (z.B. Abrufung und Integration) mit mathematischer Genauigkeit beschreiben. Aufgrund dieser exakten Beschreibungen können dann, zweitens, explizite und überprüfbare Vorhersagen für neuronale Indikatoren (z.B. N400 und P600) getroffen werden. Die zu Anfang ausgeführte Beschreibung des Sprachverarbeitungsprozesses durch die Funktionen der Bedeutungsabrufung und der Bedeutungsintegration entspricht dem komputationalem Retrieval-Integration-Modells der Elektrophysiologie des Sprachverständnisses (Brouwer et al., 2017; Brouwer et al., 2012, kurz RI-Modell). Gemäß dem RI-Modell indiziert die Amplitude der N400 die kognitive Leistung beim Abrufen von Wortbedeutungen, wobei negativere Werte höherem Aufwand entsprechen. Die Amplitude der P600 wird als Index der kognitiven Leistung bei der Beudeutungsintegration betrachtet, wobei positivere Werte höherem Aufwand entsprechen. Das Ziel dieser Dissertation ist es, spezifische Vorhersagen des RI-Modells empirisch zu validieren, wobei diese mit alternativen Interpretationen der N400 und P600 sowie mit alternativen Modellen verglichen werden.

Zu diesem Zwecke werden zunächst die EKP-Methode sowie wegweisende Resultate zusammengefasst (Kapitel 2). Basierend auf diesem Überblick werden die funktionalen Interpretationen der N400 und P600 sowie deren Rolle in Modellen der Elektrophysiologie der Sprachverarbeitung nachgezeichnet. Dem folgen drei Studien, welche entscheidende Hypothesen des RI-Modells empirisch untersuchen.

Die erste Hypothese betrifft die zentrale Rolle, welche die erwartungsbasierte Sprachverarbeitung innerhalb des RI-Modells einnimmt: Der Aufwand sowohl von Abrufung als auch von Integration sollte stark durch die Erwartbarkeit eines

Wortes moduliert werden. Neue erhobene EKP-Daten zeigen (Kapitel 3), dass unerwartete Worte tatsächlich sowohl die N400 als auch die P600 modulieren („Gestern [schärfte/aß] der Holzfäller [...] die Axt“). Die gleichzeitige Modulation von N400 und P600 bedeutet jedoch, dass aufgrund dieser Daten alleine nicht entschieden werden kann, welchem Prozess – Abrufung oder Integration – die beiden EKP-Komponenten entsprechen. Um dieses Problem zu lösen, wurde zusätzlich eine Manipulation der lexikalischen Assoziation vorgenommen („Gestern [schärfte/aß] der Holzfäller, [bevor er das Holz stapelte/bevor er den Film schaute], die Axt“). Der eingeschobene, assoziierte Nebensatz („bevor er das Holz stapelte“) sollte die Wortbedeutung des Zielwortes („Axt“) voraktivieren und dadurch dessen Abrufung zusätzlich erleichtern, jedoch ohne dabei Einfluss auf den Aufwand der Bedeutungsintegration zu nehmen. Die Ergebnisse zeigen, dass die Präsentation lexikalisch assoziierter Worte zu einer weiteren Reduktion der N400 führt, aber keinen Einfluss auf die P600 hat, was darauf hindeutet, dass die N400 Bedeutungsabrufung indiziert, während die P600 eindeutig der Bedeutungsintegration zuordenbar ist. Nachfolgend wurden Verhaltensstudien durchgeführt, in denen Lesezeiten gemessen wurden, welche ermitteln, wie lange Leser auf einzelnen Worten verweilen, was Aufschluss über den kognitiven Aufwand bei der Sprachverarbeitung geben kann. Diese Verhaltensdaten ähneln den Modulationsmustern der P600, was eine direkte Verbindung von Lesezeiten und der P600 mit dem Aufwand bei der Wortintegration nahelegt. Modulationen der Lesezeiten durch lexikalische Assoziation fielen kürzer und weniger reliabel aus, was es möglich erscheinen lässt, dass die etablierte Verbindung von Lesezeiten zur N400 nur korrelativ sein könnte. In der Summe stützen die erhobenen Lesezeitdaten die oben ausgeführte Interpretation der EKP Daten.

Eine zentrale Vorhersage des RI-Modells ist, dass die P600-Komponente von jedem Wort innerhalb einer Äußerung erzeugt wird und dass die Amplitude der P600 kontinuierlich den Aufwand der Integration indiziert. Als Teil dieser Dissertation werden erstmals EKP-Daten, welche diese Hypothese unterstützen, präsentiert. Eine post-hoc Analyse der EKP-Daten des ersten Experiments zeigt, dass sowohl die N400 als auch die P600 bei Zielworten der Kontrollkondition, welche keiner Manipulation unterlag, graduell mit der Erwartbarkeit des Zielwortes variieren. Dies würde nahelegen, dass die P600 nicht allein durch eindeutig unplausible Sätze hervorgerufen wird, sondern tatsächlich einen kontinuierlichen Index des Integrationsaufwandes darstellt. Die zweite experimentelle Studie ist speziell der Erforschung dieser Hypothese gewidmet (Kapitel 4). In diesem Experiment wird zunächst ein Kontextparagraph präsentiert, welcher den Beginn einer kurzen Geschichte enthält:

„Ein Tourist wollte seinen riesigen Koffer mit in das Flugzeug nehmen. Der Koffer war allerdings so schwer, dass die Dame am Check-in entschied, dem Touristen eine extra Gebühr zu berechnen. Daraufhin öffnete der Tourist seinen Koffer und warf

einige Sachen hinaus. Somit wog der Koffer des einfallsreichen Touristen weniger als das Maximum von 30 Kilogramm.“

Diesem Kontextparagraphen folgen abschließende Sätze, in welchen das Zielwort („Tourist“) plausibel, weniger plausibel, oder implausibel ist („Dann [verabschiedete / wog / unterschrieb] die Dame den Touristen...“). Eine zuerst durchgeführte Verhaltensstudie zeigt Verlangsamungen der Lesezeit als Funktion der Plausibilität, was die erfolgreiche Manipulation der Stimuli unterstreicht. Die Ergebnisse der danach durchgeführten EKP-Studie demonstrieren eindeutig, dass die Amplitude der P600 kontinuierlich als Funktion der Plausibilität variiert. Das experimentelle Design erlaubt zudem die Interpretation der N400 als Index der Bedeutungsabrufung zu überprüfen: Die wiederholte Präsentation des Zielwortes im vorangegangenen Kontextparagraph sollte die Bedeutungsabrufung in allen drei Konditionen gleichermaßen erleichtern - unabhängig von Unterschieden in der Plausibilität. In der Tat zeigen die EKP-Daten keinerlei Modulation der N400, was also die Zuordnung dieser EKP-Komponente zum Abrufungsprozess stützt.

Zusätzlich testet dieses Design die Vorhersagen einer Gruppe von alternativen Modellen des Sprachverständnisses, sogenannten Multi-Stream-Modellen. Multi-Stream-Modelle sagen eine verstärkte N400 für eine Kondition („Dann unterschrieb die Dame den *Tourist*“) und eine verstärkte P600 für eine andere Kondition („Dann wog die Dame den Tourist“) vorher. Dies ist abhängig davon, ob der implausible Satz eine alternative, plausible Interpretation nahelegt („Dann wog die Dame den Koffer“ anstelle von „Dann wog die Dame den Touristen“) oder nicht („Dann unterschrieb die Dame den Koffer“). Da keine der Konditionen eine verstärkte N400 hervorruft, wurde die Vorhersage der Multi-Stream-Modelle durch dieses zweite Experiment falsifiziert. Stattdessen bestätigen die Ergebnisse die Vorhersagen des *Single-Stream* RI-Modells und stellen starke Evidenzen für die Interpretation der P600 als kontinuierlichen Index der Bedeutungsintegration bereit.

Aus der Architektur des RI-Modells und der Erkenntnis, dass sowohl Bedeutungsabrufung als auch Bedeutungsintegration stark erwartungsbasiert sind, folgt eine weitere Vorhersage: Die Amplitude der N400 (je negativer die Amplitude desto höher der Abrufungsaufwand) und die Amplitude der P600 (je positiver die Amplitude desto höher der Integrationsaufwand) müssen negativ korreliert sein. Auf Prozessebene bedeutet dies: Worte, welche mehr Bedeutungsabrufung erfordern, sollten generell auch schwieriger zu integrieren sein. Diese Vorhersage steht wiederum im Kontrast zu Multi-Stream-Modellen, welche vorhersagen, dass durch jedes Wort entweder eine Verstärkung der N400 *oder* der P600 produziert werden sollte. Diese unterschiedlichen Hypothesen werden in neuen statistischen Analysen zuvor erhobener EKP-Daten überprüft (Kapitel 5). Die Resultate zeigen erstmals, dass die Amplituden der N400 und der P600 auf der Ebene einzelner EEG-Signale – und nicht nur auf der Ebene von durchschnittlichen EKP – korreliert sind. Diese Ergebnisse stärken damit weiter das RI-Modell und sind schwer mit der Architektur

eines Multi-Stream-Modells zu vereinbaren.

Zusammengefasst zeigt diese Doktorarbeit die separierbaren Einflüsse von lexikalischer Assoziation und Erwartbarkeit auf die N400. Die P600 wird dagegen nicht durch lexikalische Assoziationen moduliert, sondern reagiert darauf, wie stark die Satzbedeutung als Funktion der Erwartbarkeit und Plausibilität angepasst werden muss. Dabei ist die P600 keine kategorische Reaktion auf implausible Stimuli, sondern stellt einen kontinuierlichen Index des Bedeutungsintegrationsaufwandes dar. Des Weiteren konnte gezeigt werden, dass graduelle Modulationen der N400 und der P600 innerhalb einzelner EEG-Signale korrelieren, was auf die Organisation der erwartungsbasierten Prozesse Abrufung und Integration in einer Single-Stream-Architektur hindeutet. Für beide experimentellen Designs wurden neben EKP-Daten auch Lesezeitdaten erhoben, welche im Kontext verständnisbasierter Erwartbarkeit eine direkte Verbindung von Lesezeiten mit der P600 nahelegen.

Die Ergebnisse dieser Dissertation sind unvereinbar mit traditionellen sowie neueren Theorien, welche argumentieren, dass die N400 Aspekte der Bedeutungsintegration indiziert. Im Speziellen widersprechen die Ergebnisse mehreren Schlüsselhypothesen von Multi-Stream-Modellen, welche aussagen, dass die N400 strukturunsensible Integration indiziert, während die P600 Konflikte zwischen strukturunsensibler und struktursensibler Integration widerspiegelt. Stattdessen lassen sich die Resultate mit wesentlich weniger Annahmen durch das Single-Stream-Modell der Retrieval-Integration-Theorie erklären (siehe Diskussion in Kapitel 6). Demnach fußt das Sprachverständnis im Wesentlichen auf den Mechanismen der Bedeutungsabrufung sowie der Bedeutungsintegration, welche im EKP-Signal als N400- und P600-Komponente messbar sind. Beide Komponenten werden standardmäßig durch jedes Wort einer Äußerung hervorgerufen, wobei ihre Amplituden kontinuierlich den kognitiven Aufwand der Bedeutungsabrufung (N400) sowie der Bedeutungsintegration (P600) indizieren. Basierend auf den Ergebnissen dieser Dissertation ziehe ich den Schluss, dass eine an Erkenntnissen über das Sprach*verständnis* interessierte Forschung der P600 zentrale Bedeutung beimessen sollte.

Anhang A enthält eine theorieneutrale Abhandlung über die rERP Methode (Smith & Kutas, 2015a), einem statistischen Analyseverfahren, welches in der gesamten Dissertation zur Auswertung von EKP- und Lesezeitdaten zum Einsatz kommt. Alle Daten und sämtlicher Code, welche zur Reproduktion der Analysen und Graphiken dieser Arbeit, einschließlich des Anhangs, notwendig sind, werden im Thesis Repository bereitgestellt (https://www.github.com/caurnhammer/AurnhammerThesis). Jedwede Studien, welche mit menschlichen Partizipanten durchgeführt wurden, erhielten eine Ethik-Zulassung durch die Deutsche Gesellschaft für Sprachwissenschaft (DGfS). Teile dieser Arbeit basieren auf Veröffentlichungen in wissenschaftlichen Journalen (Kapitel 3: Aurnhammer et al., 2021; Kapitel 4: Aurnhammer, Delogu, et al., 2023; Kapitel 5: Aurnhammer, Crocker, and Brouwer, 2023).

Acknowledgements

First and foremost, I would like to extend my sincere gratitude to my supervisors, Matthew Crocker and Harm Brouwer, for the opportunity to conduct my doctoral research under their guidance. This endeavour would have been impossible without your continued support, feedback, and encouragement. Thank you, Matt, for welcoming me to your lab, and for the many formal and informal discussions we have had in the almost 5 years that have passed. Harm, thanks for the many science/hacking sessions we have had and for welcoming me to your (and Noortje's) house many times. To many more fish tacos and vinyl records.

I express my gratitude to Francesca Delogu and Miriam Schulz, who provided invaluable contributions to the work presented in this thesis and were great colleagues to work alongside. My sincere thanks go to them as well as to Elisabeth Süß, Noortje Venhuizen, Torsten Jachmann, and Wolfgang Aurnhammer for providing feedback on various parts of this dissertation. Not least, I would like to thank Derek Kunstmann for creating the amazing cover artwork showing my EEG data. Finally, I wish to express my gratitude to the members of my doctoral committee for the time and effort they took to evaluate my work.

During my time in Saarbrücken, I had the pleasure to work with many wonderful colleagues. Thank you, Torsten Jachmann, for bearing with my ramblings, Marjolein van Os, for tolerating my attempts at humour, Muqing Li, for sharing the way that lead us from Nijmegen to Saarland, Noortje Venhuizen, for teaching me the language of formal semantics (which I will not utter here), Heiner Drenhaus, for never letting me go hungry, and many others.

I express my deepest appreciation to my partner, Elisabeth Süß, my parents, Adelheid and Wolfgang Aurnhammer, and my brother Lukas Aurnhammer, who have supported me, always. Lastly, I would like to acknowledge my friends for connecting me to reality while I descended into the madness of science.

My work as a doctoral researcher at Saarland University was funded by the Collaborative Research Center "Information Density and Linguistic Encoding" (Project-ID 232722074 — SFB 1102), financed by the German Research Foundation (Deutsche Forschungsgemeinschaft).

Contents

List of Figures

List of Tables

For my parents.

Chapter 1

General Introduction

1.1 Building Blocks of Language Comprehension: Retrieval and Integration

To understand language, comprehenders must *retrieve* the meanings associated with the words they perceive from memory and they must *integrate* retrieved word meanings into a representation of utterance meaning. These processes are fundamentally incremental: More or less every incoming word in an utterance immediately triggers meaning retrieval and meaning integration. Crucially, as each incoming word contributes meaning in context of the incrementally unfolding utterance representation, certain continuations are more likely than others. That is, comprehension is guided not only by the perceived information but is also strongly expectation-based: Both processes – retrieval and integration – can be facilitated based on what has been understood so far. Thus, expectation-based processing improves the efficiency of the comprehension system because context often dictates which words and utterance meanings can and cannot continue the current utterance. For instance, after having constructed a meaning representation for the partial sentence "He spread his warm bread with ...", retrieval and integration are facilitated for the word "butter" (example from Kutas & Hillyard, 1980). However, if the sentence fragment is continued with the word "socks", no contextual facilitation of retrieval and integration is given.

The key requirement for expectation-based language comprehension is an incrementally constructed utterance meaning representation. The meaning representation of a partial sentence may preactivate the meanings of possible upcoming words and thereby facilitate their retrieval. Similarly, utterance meaning may facilitate integration of word meaning, if the new information is predictable given world knowledge. This dissertation is driven by the assumption that language comprehension is fundamentally shaped by the expectation-based retrieval and integration mechanisms – a position which is indeed consistent with surprisal theory, an information-theoretic formalisation of expectancy (Hale, 2001; Levy, 2008). If these two processes truly underlie language comprehension, the question arises of how these mechanisms are cognitively organised and how gradual differences in the facilitation of retrieval and integration can be measured.

1.2 Electrophysiological Correlates of Expectation-Based Language Processing

Key insights into the question of how comprehenders process language have been achieved through measuring event-related potentials (ERPs). Computed from the electroencephalogram recorded by electrodes placed on the scalp, ERPs reveal the electrical activity of the brain, extending over time after the presentation of a stimulus (such as the word "butter" in the above example). When large ensembles of neurons become active synchronously in response to a stimulus, this activity can become manifest as an ERP component, a reliably observed deflection of the event-related signal which varies in its morphological characteristics – e.g., its amplitude – and in how these morphological characteristics are differentially modulated by stimulus properties – e.g., becoming more negative or more positive for unexpected relative to expected words. In electrophysiological language research, two ERP components have played a major role in developing our understanding of language comprehension: The N400 and the P600.

The N400 is a negativity usually observed between 300 and 500 milliseconds post-stimulus onset, which was first described in a study manipulating the contextual congruency of a target word ("He spread the warm bread with butter/socks", Kutas and Hillyard, 1980). The amplitude of the N400 was found to be more negative for the incongruent word "socks" compared to the congruent word "butter". While it is a plausible assumption that contextual support may facilitate the processing of the target word "butter", it is not immediately obvious what it is specifically about the difference in congruence that induces the difference in neural activity for the expected relative to the unexpected target word. Clearly, presenting the word "socks" creates a sentence with a very implausible meaning, and, hence, the N400 may reflect the effort involved in updating the utterance meaning representation to represent a rather implausible scenario. However, compared to "butter", the word "socks" also receives little conceptual priming from the preceding context. Hence, the negativity could reflect lower retrieval facilitation for "socks" compared to "butter". Further, the overall unexpectedness of the word "socks" likely affects both retrieval and integration, such that any functional interpretation of the N400 based on this original finding is underdetermined.

Critically, in the ERPs reported by Kutas and Hillyard (1980), the signals differ not only in the N400 time window. Following the N400, around 600 milliseconds post-stimulus onset, the signals flip and the ERP of the incongruent words becomes more positive compared to the congruent words. While not discussed in the original article, this positivity has later been named the P600, when it was found in stimuli manipulating syntax (Hagoort et al., 1993; Osterhout & Holcomb, 1992). However, the stimuli used by Kutas and Hillyard (1980) do not manipulate syntax in any way: The sentence "He spread the warm bread with socks" is syntactically as well-formed

as the control condition. Rather, the manipulation is semantic in nature. Again, the question is which aspects of the linguistic manipulation – conceptual priming from the context, plausibility of utterance meaning, expectancy – modulate which aspects of the comprehension process – e.g., retrieval and integration – and how these processes map to the N400 and the P600, respectively. Since 1980, thousands of electrophysiological studies have been devised, partly with the goal to elucidate the question of the functional significance of the N400 and the P600. However, these studies have not resolved all uncertainty about the processes underlying the two components. Rather, the interpretations of the N400 and the P600 continue to be debated.

Retrieval-Integration (RI) theory (Brouwer et al., 2017; Brouwer et al., 2012) explicitly posits that retrieval and integration fundamentally underlie language comprehension and links retrieval to the N400 component (adopting earlier interpretations by Kutas & Federmeier, 2000, 2011; Lau et al., 2009; van Berkum, 2009, 2010) whereas the P600 is taken to index integration. However, these interpretations of the N400 and the P600 are not universally shared: The N400 has alternatively been interpreted to index integration (Brown & Hagoort, 1993, 2000; Hagoort et al., 2004) or both retrieval and integration (Baggio, 2018; Baggio & Hagoort, 2011; Lau et al., 2016; Nieuwland et al., 2020). The P600, on the other hand, has originally been described as an index of syntactic processing (Friederici, 1995; Hagoort et al., 1999; Kaan et al., 2000; Kaan & Swaab, 2003; Osterhout & Holcomb, 1992), or as an index of conflict monitoring or revision (Bornkessel-Schlesewsky & Schlesewsky, 2008; A. Kim & Osterhout, 2005; Kos et al., 2010; Kuperberg, 2007; van Herten et al., 2005). Crucially, different functional interpretations of the N400 and the P600 component give rise to very different architectures of the language comprehension system. For instance, while Retrieval-Integration theory proposes a single-stream model in which every word undergoes retrieval and integration, a set of multi-stream models make a strikingly different proposal: Multi-stream models posit that the language comprehension system effectively engages in two separate notions of integration in parallel (Bornkessel-Schlesewsky and Schlesewsky, 2008; A. Kim and Osterhout, 2005; Kos et al., 2010; Kuperberg, 2007; Michalon and Baggio, 2019; van Herten et al., 2005, similarly Li and Ettinger, 2023; Rabovsky et al., 2018; Ryskin et al., 2021). The N400 is taken to correspond to a plausibility-heuristic which attempts to construct utterance meaning while disregarding the structural properties of the utterance. Importantly, it is thus possible that this process constructs a *semantically attractive* interpretation rather than the literal interpretation. In parallel, a meaning representation of the input is also generated while adhering to morphosyntax. The P600 is taken to index not this *structure-sensitive* integration directly, but rather situations in which the *structure-sensitive* and *structure-insensitive* analyses conflict with each other. Thus, multi-stream models differ from RI theory in which processes are involved during comprehension, how they interact, and how they map to the N400 and the P600.

This divergence illustrates that arriving at a shared understanding of the interpretation of the N400 and the P600 has direct consequences for our understanding of the language comprehension architecture.

Thus, in sum, even though there is no shortage of empirical results detailing the sensitivities of the N400 and the P600, the field has converged neither on a generally accepted formalisation of the processes necessary for comprehension nor on their mapping to ERP components. The resulting uncertainty hinders progress in the description of the neurocognitive implementation of language comprehension, which in turn impedes the design of effective experimental studies and their unambiguous interpretation.

1.3 Empirical Investigation

Crucially, RI theory is instantiated as a *computational* model (Brouwer et al., 2017; Brouwer, Delogu, Venhuizen, and Crocker, 2021) which provides a mathematically precise description of the processes thought to underlie *both* the N400 and the P600. Because of this level of formal exactness, explanations of well-known relationships between linguistic stimulus properties and the event-related potentials they affect can be derived and, most importantly, predictions about future experiments investigating both the N400 and the P600 can be drawn. In order to progress the field of psycholinguistics in its goal to arrive at a shared understanding of the processes underpinning language comprehension, this dissertation empirically investigates key predictions of RI theory that provide strong contrasts to competing interpretations and models of the N400 and the P600.

Retrieval (N400) and Integration (P600) in Expectation-Based Comprehension Within the single-stream architecture of RI theory, the utterance meaning representation constructed so far strongly influences both retrieval and integration of the currently processed word. Indeed, understanding the central role of expectancy for the electrophysiological correlates of retrieval and integration may thus shed light on the neurophysiological basis of surprisal. The key role of utterance meaning based expectancy is investigated in a novel experimental design (Chapter 3, Design 1), in which we compare expected words ("Yesterday sharpened the lumberjack, [...], the axe", transliterated from German) to unexpected words ("Yesterday ate the lumberjack, [...], the axe"). We find that unexpected words induce an increase in both the N400 and the P600, thus yielding a biphasic N400-P600 effect pattern. Critically, due to the simultaneous modulation of both components, this result alone does not allow an unequivocal attribution of retrieval and integration to the N400 and the P600. In order to arrive at such an unambiguous attribution, we also manipulate lexical association of the target word to the immediately preceding context, in a way that modulates only retrieval and not integration effort. This is achieved by inserting semantically associated ("[...], before he the wood stacked, the axe") or unassociated

material ("[...], before he the movie watched, the axe") before the target word. We find that the additional lexical priming from the associated adverbial clause reduces N400 amplitude while not influencing the P600, hence supporting the link of the N400 to retrieval and of the P600 to integration, as proposed by RI theory. As we discuss in detail, competing theories and models only partially account for the ERP modulations found in this study, whereas RI theory directly predicts the observed N400 and P600 modulations.

In two separate behavioural experiments that presented the same stimuli, we also recorded reading times in a self-paced paradigm. We conducted the same experiment twice, differing only in the task that participants completed (binary plausibility ratings vs. comprehension questions). Behaviourally, reduced expectancy slowed reading across spillover regions, regardless of the task. Reduced association only slowed readers on one spillover region and only for the comprehension question task. Taken together, the unique sensitivity of the P600 to expectancy in the current design and the reliable role of expectancy for reading times support a close relation of reading times to the P600 within a framework of "comprehension-centric" surprisal (Venhuizen et al., 2019; Brouwer, Delogu, Venhuizen, and Crocker, 2021).

The P600 as a Continuous Index of Integration Effort While some theories of language comprehension make no direct predictions about the P600 (e.g., hybrid theories assigning both retrieval and integration to the N400) or hypothesise the P600 to be a binary index of conflicting analyses (multi-stream models), a key prediction of RI theory is that the P600 is elicited by every word and that its amplitude continuously indexes the effort of updating the utterance meaning representation constructed so far with the meaning of the current word. First evidence in support of this hypothesis is presented in a post-hoc analysis examining the ERPs of the control condition of Design 1. Within the control condition, target words generally lead to plausible interpretations, yet, these interpretations vary in their expectancy, as operationalised by Cloze probability (range = 0.17 - 1). The post-hoc analysis suggests that the differential expectancy of the target words in the baseline condition elicits both a graded N400 response - replicating earlier work (Frank et al., 2015; Kutas & Hillyard, 1984) - and, as a novel result, a graded P600 response.

To corroborate these exploratory results, we developed a design which is entirely dedicated to the prediction that the amplitude of the P600 is a continuous index of integration effort (Chapter 4, Design 2). We first present a context paragraph, which repeatedly mentions the target word ("A tourist wanted to take his huge suitcase onto the airplane. [...]", transliterated from German). This context paragraph is followed by a final sentence in which the target word is plausible, less plausible, or implausible ("Then [dismissed / weighed / signed] the lady the tourist"). On RI theory, this design should induce no N400 modulations, as the lexical repetition of the target word in the preceding context paragraph should facilitate target word retrieval equally strongly in all three conditions. The P600, on the other hand, should

become more positive across the three conditions, thus scaling inversely with plausibility. An initial self-paced reading study resulted in graded reading times across conditions, thereby validating that the stimuli induce graded integration effort.

Additionally, the design captures the predictions of multi-stream models. In the less plausible continuation, the design makes a plausible alternative interpretation available through semantic attraction ("Then weighed the lady the suitcase"), which, according to multi-stream models, should induce only an increase in P600 amplitude and not in N400 amplitude. In the implausible continuation, *no* plausible alternative interpretation is available, and hence, multi-stream models predict an increase in N400 amplitude but not in P600 amplitude.

The manipulation of plausibility resulted in no N400 effects between conditions. Rather, the stimuli elicited graded P600s that match the plausibility difference across the three conditions – a relation which was modelled statistically by continuous plausibility ratings. Hence, these results confirm the key prediction of RI theory that the P600 – and not the N400 – is a continuous index of integration effort. We found no evidence for categorical N400/P600 increases induced by the absence/presence of semantic attraction, as predicted by multi-stream models, thereby disconfirming one of their key predictions. In a post-hoc analysis, we successfully regressed P600 amplitudes on the per-item reading times recorded in the behavioural experiment, further corroborating the primary link of the P600 – and not the N400 – to reading times and comprehension-centric surprisal.

Single-trial Neurodynamics Reveal N400 and P600 Coupling in Language Comprehension Retrieval-Integration theory posits that the N400 *and* P600 are elicited by every word in an utterance and that their amplitudes continuously index retrieval and integration, respectively – relations that were confirmed in the first two studies. Indeed, many ERP studies found biphasic N400-P600 patterns for incongruent relative to congruent words (Van Petten & Luka, 2012). We argue that both RI theory and multi-stream models can explain biphasic patterns at the *effect* level – comparing average waveforms of incongruent and congruent target words – but that their accounts can be disentangled at the single-trial level of EEG recordings for individual target words (Chapter 5).

The single-trial prediction of Retrieval-Integration theory is that, generally, words that are more effortful to retrieve should also be more effortful to integrate. This differs strikingly from the prediction of multi-stream models: At the single-trial level, any given word should induce either an increase in N400 amplitude *or* an increase in P600 amplitude, but typically not both (for a detailed explanation see Chapter 5). These diverging predictions can be expressed as correlations: RI theory predicts a negative correlation, in that more negative N400 amplitudes should co-occur with more positive P600 amplitudes on the same trials. The prediction of

multi-stream models, on the other hand, corresponds to a positive correlation between the N400 and the P600, as N400 amplitudes that are more negative should co-occur with P600 amplitudes that are more negative than average.

We investigate these diverging predictions by revisiting the biphasic ERP data from Design 1 ("Yesterday [sharpened/ate] the lumberjack [...] the axe"). Using a novel regression approach, we demonstrate that N400 amplitude is predictive of P600 amplitude at the single-trial level, in that more negative N400 amplitudes predict more positive P600 amplitudes. Crucially, the single-trial interrelation of the N400 and the P600 which we found for the biphasic effect of Design 1 is also present in ERPs from a study which resulted only in monophasic effects between conditions (Delogu et al., 2019, see Brouwer, Delogu, and Crocker, 2021; Delogu et al., 2021, for a discussion of component overlap with regard to these monophasic results).

Our approach demonstrates that competing effect-level explanations can be dissociated by specifying testable predictions at the single-trial level and highlights the importance of statistical analysis at the single-trial, rather than at the effect level. In particular, the finding that increases in N400 and P600 amplitude are coupled within-trial supports the single-stream model proposed by Retrieval-Integration theory and appears inconsistent with the processing architecture proposed by many Multi-stream models, which predict that any given word should elicit either an N400 increase or a P600 increase.

Summary Within the field of language comprehension research, thousands of studies that investigate the sensitivities of the N400 and the P600 component to linguistic materials have been put conducted. The field has, however, not arrived at a shared understanding of the processes necessary for comprehension and the mapping of possible processes to ERP components remains debated. This thesis aims to elucidate these issues by investigating key predictions of Retrieval-Integration theory that strongly contrast with those of competing models.

We found that the N400 is independently modulated by lexical association to the preceding context and is continuously related to target word expectancy. No N400 effects between conditions are observed when target words are strongly and equally primed through repetition. The P600 is continuously modulated by the expectancy and the plausibility of the interpretation the target word induces. Both components were found to be correlated on the single-trial level, where trials with more negative N400 amplitudes also induced more positive P600 amplitudes. While reading times can be modulated by lexical association, their elicitation patterns most reliably resembled those of the P600.

These results (discussed in full detail in Chapter 6) are in direct opposition to both traditional and more recent proposals arguing that (1) the N400 in part indexes integrative processes, (2) integration – as indexed by the N400 – is operating

in a structure-insensitive manner, and (3) the P600 indexes the resolution of conflicts within a multi-stream architecture in which structure-insensitive and structure-sensitive integration processes construct interpretations in parallel. Rather, our results indicate that (1) integration is indexed by the P600, (2) integration is fully compositional and structure-sensitive, and (3) a single-stream model, Retrieval-Integration theory, in which the N400 continuously indexes retrieval and the P600 continuously indexes integration is sufficient to account for the key ERP data. We argue that the N400 and the P600 should be considered as part of the default signature of language comprehension and that future study of a comprehension-centric view on language processing should focus on the P600.

1.4 Method and Reproducibility

Throughout this thesis, we apply regression-based ERP estimation (rERPs, Smith & Kutas, 2015a), a technique that decomposes ERPs into the contributions made by different stimulus properties at full temporal and spatial resolution. Appendix A provides a practical and theory-neutral description of the rERP method and its application to both ERPs and reading times. Code and data required to reproduce the analyses of this thesis, including the appendix, are made publicly available: https://github.com/caurnhammer/AurnhammerThesis. This dissertation is published open-access.

1.5 Ethics and Funding

All studies involving human participants were conducted with ethics approval of the Deutsche Gesellschaft für Sprachwissenschaft (DGfS). This work was funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Project-ID 232722074—SFB 1102.

1.6 Publications

Several chapters of this dissertation are adapted from peer-reviewed, open-access journal articles. Complete bibliographic information for the last article is pending at the time of writing.

- Chapter 3: Aurnhammer, C., Delogu, F., Schulz, M., Brouwer, H., Crocker, M. W. (2021). Retrieval (N400) and integration (P600) in expectation-based comprehension. *PLOS ONE*, *16*(9), e0257430. https://doi.org/10.1371/journal.pone.0257430
- Chapter 4: Aurnhammer, C., Delogu, F., Brouwer, H., Crocker, M. W. (2023). The P600 as a continuous index of integration effort. *Psychophysiology*, *60*(9), e014302. https://doi.org/10.1111/psyp.14302

- Chapter 5: Aurnhammer, C., Crocker, M. W., Brouwer, H. (2023). Single-trial neurodynamics reveal N400 and P600 coupling in language comprehension. *Cognitive Neurodynamics*. https://doi.org/10.1007/s11571-023-09983-7

Chapter 2

The Electrophysiology of Language Comprehension: An Overview

2.1 Why Event-Related Potentials

For the study of the cognitive functions that underpin language comprehension, it is an inviting prospect to measure the "hardware" that implements these functions directly. That is, researchers have been interested in going beyond measurements of human behaviour, such as reaction times, reading times, or eye movements, and in moving to an understanding of how exactly the brain itself is involved in the human ability to comprehend and produce language. In the history of psycholinguistics, the study of the brain commenced already in the 19th century, with the descriptions of language deficiencies induced by brain damage, a process with relied on post-mortem dissection of the brains of afflicted patients (see Levelt, 2013). By now, many neuroscientific measures are available that allow researchers to study the brain in vivo, meaning that the activity of the human brain can be measured during online, incremental language comprehension. Electroencephalography (EEG) recordings – measured from the human brain for the first time in 1924 (Berger, 1929) – and the event-related potentials (ERPs) derived from them have proven particularly insightful, specifically because they offer a temporal resolution that is high enough to determine stimulus induced brain activity at millisecond scale.

While raw EEG signals are commonly used in some applications (e.g., the resting state EEG in medical contexts), they are characterised by a low signal-to-noise ratio that imposes limits on its applicability. What made EEG recordings an indispensable tool for cognitive science was the discovery that by averaging many event-locked EEG recordings, the systematic *event-related* variation can be separated from the unsystematic variation, including noise (Luck, 2005). An entire array of ERP components – waves following and overlapping with each other – are typically elicited by presenting a stimulus and research is concerned with establishing how these components vary systematically in morphological properties such as their amplitude, onset, offset, peak latency, frequency content, etc., as a function of stimulus properties. Further, by placing a grid of electrodes around the scalp of a participant, it is possible to determine whether ERP components are stronger over some

electrode sites than others, giving rise to specific scalp distributions that add a spatial dimension to EEG data (which is not to be equated with specific brain regions, see below). Importantly, ERP components often overlap with each other temporally and spatially which means that, in the overlapping regions, their amplitudes add positively if they are of the same polarity, or they partially cancel out if they are of opposite polarity (Luck, 2005). Because of this spatiotemporal *component overlap*, the observed components may not always directly reflect the underlying, latent component structure (Brouwer & Crocker, 2017). In sum, event-related potentials consist of ERP components which vary systematically along the amplitude, time, frequency, and spatial dimensions.

Aside from its practical advantages of being a low-invasive, cheap, and easily deployed brain measurement technique, the strength of EEG is the high temporal resolution that the analogue signal recorded from the electrodes provides, which is usually digitally sampled at a resolution in the order of one or two milliseconds. This temporal resolution allows researchers to determine *what* processes are invoked during stimulus processing through experimental manipulations and to determine *when* these processes are active. Like every method, EEG also has unique disadvantages: Most importantly, EEG is very limited in its ability to add to discussions about localisation, i.e., the question of which brain regions are involved in specific cognitive processes. While many attempts at source modelling have been undertaken – i.e., techniques that aim to determine which set of neural generators induce the activity observed at the scalp – these approaches have been heavily criticised: Because there is an infinite number of forward solutions, i.e., combinations of neural generators, that could result in the observed scalp distributions, it is difficult to decide between them – an issue known as the *inverse problem* (Luck, 2005). Thus, while there is a reliable spatial dimension to ERPs – manifest as differences in amplitude over different electrodes – these should not be mistaken to correspond directly to neural generators located in specific areas of the brain.

Paramount to arrive at correct interpretations of ERPs is an understanding of their physical basis. A single ERP emerges as the summated post-synaptic potential of a large group of neurons which fire synchronously in response to an event. This summated signal can be picked up on the scalp by placing a specific electrode setup on it. The continuous signal, unfolding over time after presenting a stimulus, expresses the difference between two electrodes – the active and the reference electrode[1] – and is digitised to single, scalar values according to some sampling rate. This process yields a series of single scalar values, one for each combination of time sample, electrode, participant, and experimental trial. In general, EEG data reflect the brain activity of many cognitive processes, not all of which are relevant to the experiment the participant is taking part in, and, additionally, the signal is influenced by many external factors, such as eye movements, voltage drifts resulting from, e.g.,

[1] Technically, what is amplified is the difference between the difference of reference and ground electrode and the difference of active and ground electrode (Luck, 2005).

variability in skin conductance, or electrical devices nearby the electrodes, all of which can introduce artefacts and noise into the recording. Typically, the recordings extracted for individual trials in an experiment – e.g., the activity following visual presentation of a word – undergo artefact rejections steps, in which either an automatic, semi-automatic, or completely manual process is applied to determine whether any given trial should be included in the final data set used for visualisation and statistical analysis. In many ERP experiments, researchers are interested in determining differences in ERP components between conditions. In the conventional ERP averaging process for a within-subjects design, it is customary to first compute the average ERP wave for each participant and each condition. Following this, the per-condition waveforms are computed from the per-participant, per-condition averages. Due to this two-step procedure, each participant contributes equally to the final condition averages, regardless of the amount of data rejection on that subject. By averaging the EEG signals collected from many trials, random variation is filtered out, whereas variation that is systematically elicited by the stimulus remains. As outlined above, the resulting average ERPs are characterised by ERP components and experiments often aim to elicit a difference in the morphological properties of one or more ERP components between conditions.

2.2 Language-Elicited ERP Components

The ERP response to language is characterised by several reliably modulated components. Of particular interest to language comprehension research are the N400, a negativity observed between 300 and 500 milliseconds after presentation of a stimulus, and the P600, a parietally peaking, sustained positivity emerging from around 500 milliseconds post-stimulus onset. Figure 2.1 displays the ERP response to an experimental design with two conditions, contrasting expected with unexpected target words ("Yesterday [sharpened / ate] the lumberjack [...] the axe"; see Chapter 3). The unexpected target words elicited both a more negative N400 amplitude and a more positive P600 amplitude on average, compared to the expected target words in the baseline condition. While there are also several other language-related components such as the mismatch negativity (MNN), the N200, the left anterior negativity (LAN; see Kaan, 2007, for an overview about these components), the P300 (see Osterhout et al., 1996), as well as a frontal late positivity (see Federmeier et al., 2007; Kuperberg et al., 2020; Van Petten & Luka, 2012), this dissertation largely focuses on the N400 and the P600. Importantly, the N400 and the P600 are of opposite polarity and thus can in principle be affected by component overlap attenuating their amplitudes (see Brouwer and Crocker, 2017, for discussion and Brouwer, Delogu, and Crocker, 2021; Delogu et al., 2021, for empirical evidence). In fact, the non-overlapping time windows of the observed waveforms in biphasic data (cf. Figure 2.1) could be an artefact of their partial cancellation in an overlapping time window:

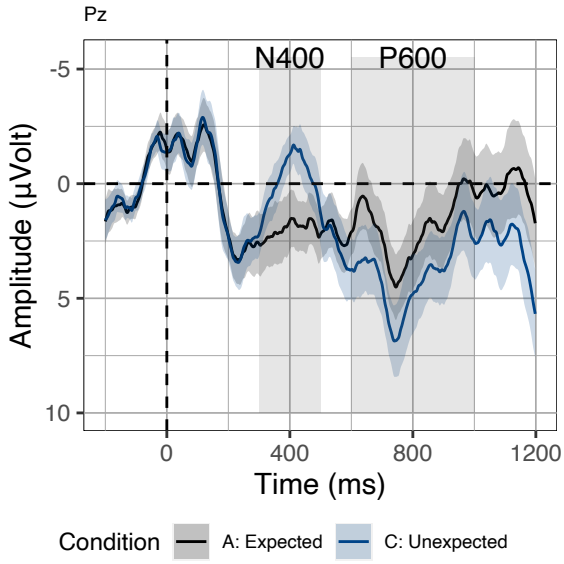


FIGURE 2.1: ERP data averaged for two conditions which differed in target word expectancy, eliciting an N400 effect (300 - 500 ms) and a P600 effect (600 - 1000 ms) between conditions ("Yesterday sharpened / ate the lumberjack [...] the axe"; see Chapter 3). Error ribbons indicate confidence intervals around the mean, computed from the standard error across subjects multiplied by 1.96.

If an N400 effect *and* a P600 effect occur relative to the same baseline, it is impossible to observe both of them on the same time samples within an electrode. Further, N400 effects that are not followed by a P600 effect are sometimes characterised by a later offset latency (e.g., see Delogu et al., 2021, for such a sustained negativity), and, similarly, P600 effects that are not preceded by an N400 increase are characterised by an earlier onset latency (e.g., see Chapter 4).

Both the N400 and the P600 are sensitive to many different linguistic manipulations. These findings have informed attempts to attribute specific cognitive functions – e.g., semantic access, syntactic structure building, semantic integration, etc. – to the two ERP components. Due to the multitude of ERP results that often seem incompatible, the debate about the functional interpretation of the N400 and the P600 components is still ongoing (see Delogu et al., 2019, for an overview). The following sections provide a brief overview of key N400 and P600 results as well as a discussion of their functional interpretations and how these interpretations have been translated into computational models of the electrophysiology of language comprehension.

2.2.1 N400

The N400 was first described in a study by Kutas and Hillyard (1980) which examined the electrophysiological response to incongruous sentence completions ("He spread the warm bread with socks") relative to congruous control sentences ("He spread the warm bread with butter"). Measuring on the word "socks", the authors

hypothesised to elicit a P300 effect for the unexpected word "socks" relative to the expected word "butter". However, they discovered the N400 effect instead. Over the past forty years, the N400 has become the most frequently studied ERP component of language processing. It is elicited by spoken, written, and signed words (Kutas & Federmeier, 2011). Importantly, it is also found outside of language settings, when faces, gestures, pictures, mathematical symbols, videos, or sounds are presented, suggesting that it is an index of the processing of meaningful stimuli in general (Kutas & Federmeier, 2011). In sum, the N400 varies systematically in its amplitude in that stimuli which receive contextual support elicit less negative amplitudes.

While the original experiment (Kutas & Hillyard, 1980) measured the ERP response to a target word in a sentence context, the N400 is also elicited and modulated when presenting single words in isolation. When presenting single words, attenuations in N400 amplitude have been observed for more frequent words (Rugg, 1990) and for words containing frequent syllables (Barber et al., 2004). Further, semantically richer words, i.e., words with more semantic features, more semantic associations, and more concrete meaning, elicit a more negative N400 response than semantically less rich words (Kounios and Holcomb, 1994; West and Holcomb, 2000, however, see Kounios et al., 2009). Words with larger orthographic neighbourhood size – i.e., words that are orthographically similar to many other words – were found to elicit a larger N400 amplitude than words with few orthographic neighbours (Holcomb et al., 2002). Perhaps surprisingly, the same effect is found for pseudowords (Laszlo & Federmeier, 2011). That is, strings of letters that bear orthographic similarity to existing words but are not lexicalised, i.e., not linked to conceptual knowledge via convention, also elicit a larger N400 amplitude when they have many orthographic neighbours. While the N400 response to illegal strings is smaller than that to legal strings, the N400 is, however, clearly elicited even by illegal strings (Laszlo & Federmeier, 2008). In priming studies, where a target word is preceded by a "prime" word, N400 reductions were observed for target words that were preceded by a semantically related word, compared to targets word preceded by an unrelated prime (Franklin et al., 2007; Rugg, 1985).

Most N400 results from single-word presentation studies appear to generalise to utterance processing as well: Semantic association/relatedness (Chapter 3; Federmeier and Kutas, 1999; Kutas, 1993), frequency (Van Petten, 1993), and orthographic neighbourhood size (Laszlo & Federmeier, 2009, 2011) also modulate the N400 for words presented within sentences. However, as the context provided by the sentence becomes richer, contextual effects appear to start overriding the influence of the words in isolation. For instance, the influence of word frequency on the N400 is larger at the start of a sentence but dwindles as the sentence continues (Van Petten & Kutas, 1990) and priming effects are sometimes overridden by strong message-level constraints (Coulson et al., 2005).

Indeed, the expectancy of a word in a sentence (Kutas & Hillyard, 1980; Kutas &

Hillyard, 1984) or discourse context (van Berkum, 2009) is one of the strongest predictors of the N400 (Dambacher et al., 2006; DeLong et al., 2011; Kutas & Hillyard, 1984) and is inversely and continuously related to its amplitude (see also Chapter 2 for a contextualisation of expectancy effects on ERPs in the surprisal literature). That is, the N400 is not a binary reflection of a semantic incongruency but a graded response to stimulus expectancy (Kutas & Hillyard, 1980). Word expectancy is often operationalized as Cloze probability (Taylor, 1953), computed from the number of times the target word was produced as a completion for a preceding sentence fragment by a group of participants in a norming study conducted prior to the ERP experiment (see Kutas & Hillyard, 1984, for the first application of Cloze probability in an ERP study). A corpus-based approach computes word probabilities conditioned on the preceding words by using next-word prediction language models. Indeed, conditional word probabilities (typically transformed to their negative logarithm and referred to as surprisal) computed from language models are predictive of N400 amplitude in naturalistic sentence reading data (Frank et al., 2015), i.e., studies in which no artificially constructed conditions are employed. Experimental manipulations often make use of uncommon sentences which are not attested in corpora and hence tend to rely on Cloze probabilities rather than on language models. The sensitivity of the N400 to contextual constraint however also reaches limits: For instance, implausible, negated sentences do not induce N400 modulation over and above what is explained by semantic relatedness ("A robin [is / is not] a [bird / vehicle]", Fischler et al., 1983, but see Nieuwland and Kuperberg, 2008; Palaz et al., 2020). Further, even though context strongly determines N400 amplitude, target words that are matched for expectancy are not modulated by the strength of constraint imposed on the target word by the context (Federmeier et al., 2007). Importantly, in a series of studies contrasting unexpected and implausible target words to expected, plausible words – for instance by reversing thematic roles – no N400 effects were observed (see Bornkessel-Schlesewsky & Schlesewsky, 2008; Brouwer et al., 2012; Kuperberg, 2007, for reviews). These studies often elicited a P600 effect when an N400 effect was expected and will be referred to as "semantic P600" studies henceforth. Due to these surprising results, semantic P600 studies were highly influential on theoretical and computational accounts of the electrophysiology of language comprehension.

2.2.2 P600

The P600 (or syntactic positive shift, SPS; Hagoort et al., 1993) was originally observed for manipulations of syntax by Osterhout and Holcomb (1992). This later component has been found to be elicited by syntactic violations, garden path sentences, or syntactically complex sentences (beim Graben et al., 2008; Friederici and Mecklinger, 1996; Hagoort et al., 1993; Kaan et al., 2000; Kaan and Swaab, 2003; Osterhout and Holcomb, 1992; Osterhout et al., 1994; Osterhout and Mobley, 1995; see Gouvea et al., 2010, for a review).

Importantly, however, P600 effects were later also obtained for semantic manipulations (Münte et al., 1998) and featured prominently in semantic P600 studies (Hoeks et al., 2004; A. Kim and Osterhout, 2005; Kuperberg, 2007; Nieuwland and van Berkum, 2005; van Herten et al., 2005; for reviews see Bornkessel-Schlesewsky and Schlesewsky, 2008; Brouwer et al., 2012; Kuperberg, 2007). For instance, P600 effects (rather than N400 effects) were found for semantic role violations ("the hearty meal was devouring/devoured", A. Kim and Osterhout, 2005) and for thematic role reversals ("the javelin has the athletes thrown" vs. "the javelin was by the athletes thrown", Hoeks et al., 2004). Later studies revealed that this phenomenon is not specific to materials in which a *semantic illusion* could be a plausible explanation of the absence of the N400 effect (Chow & Phillips, 2013; Nieuwland & van Berkum, 2005, see Chapter 4 for further discussion), i.e., a situation in which comprehenders temporarily employ a plausible interpretation (e.g., the athletes throwing the javelin, when reading "the javelin has the athletes thrown"), indicating a more general role of the P600 in language comprehension.

Indeed, P600 effects are also observed for many canonical semantic incongruencies, where biphasic N400-P600 effect patterns are often, but not always, elicited (see Van Petten and Luka, 2012, for an overview and Brouwer and Crocker, 2017, for a discussion with regard to component overlap). More generally, the P600 was found to be modulated by the semantic and pragmatic processing effort involved in establishing a coherent representation, e.g., by introducing new discourse referents (Burkhardt, 2006, 2007), irony (Regel et al., 2010; Spotorno et al., 2013), the semantics of visual stories (Cohn & Kutas, 2015; Sitnikova et al., 2008; Võ & Wolfe, 2013), topic shifts (Xu & Zhou, 2016), accented words (Dimitrova et al., 2012), scalar implicature (Spychalska et al., 2016), metonymy (Schumacher, 2013), and noun-phrase meaning composition (Fritz & Baggio, 2020, 2022). Just like the N400, the P600 appears to be a domain-general rather than a language-specific index of processing effort (see Leckey & Federmeier, 2020, for discussion), as evidenced by elicitations in visual scenes (Cohn & Kutas, 2015; Sitnikova et al., 2008; Võ & Wolfe, 2013), music (Patel et al., 1998), non-linguistic sequences (Christiansen et al., 2012; Lelekov-Boissard & Dominey, 2002), and arithmetic (Núñez-Peña & Honrubia-Serrano, 2004).

Strikingly, P600 effects elicited by semantic manipulations are not as reliably observed as N400 effects. For instance, while all studies on semantic incongruencies reviewed by Van Petten and Luka (2012) elicited an N400 effect, only a subset elicited a late positivity (further subdivided into frontally and parietally peaking positivities; see below for discussion). One aspect of an explanation of this phenomenon is spatiotemporal component overlap (Luck, 2005) between the N400 and the P600 (Brouwer & Crocker, 2017), leading to the partial cancellation of their amplitudes. Indeed, Brouwer, Delogu, and Crocker (2021) demonstrated that component overlap between a sustained negativity and the P600 explains the absence of an expected P600 effect in a study by Delogu et al. (2019). Further corroborating this explanation, Delogu et al. (2021) demonstrated that the P600 effect resurfaces when conditions are

matched such that target words elicit similar N400 amplitudes. Also in line with this argument is the finding that P600 effects elicited by syntactic processing difficulties are more reliably observed, because these manipulations typically do not induce differences in N400 amplitude.

A second aspect that may explain why Semantic P600s effects are not always observed is that the P600 has been shown to be sensitive to task demands. In ERP experiments, participants may or may not be instructed to complete a task in between reading sentences, such as stating whether a probe word was present in the previous sentence, rating the sentence for plausibility, or replying to comprehension questions. For manipulations of semantics, the amplitude of the P600 was found to be reduced when task demands were lightened or removed (Geyer et al., 2006; Kolk et al., 2003; Osterhout et al., 1996; Schacht et al., 2014; see Brouwer et al., 2012; Kuperberg, 2007, for discussion). Indeed, Van Petten and Luka (2012) excluded studies that employed an explicit task during the experiment, which may thus also explain why only a subset of the studies they reviewed elicited a positivity in the P600 time window.

It is worth noting that in a subset of the late positivities reviewed by Van Petten and Luka (2012), amplitudes were peaking frontally, rather than parietally, where they would be expected for canonical P600s. This frontal late positivity is typically elicited when a sentence context constrains strongly for a specific word which is then not presented and replaced by an unexpected word (Brothers et al., 2015; DeLong et al., 2014; DeLong et al., 2011; Federmeier et al., 2007; Kuperberg et al., 2020; Quante et al., 2018; Thornhill and Van Petten, 2012, see also earlier data by Kutas, 1993, and Stone et al., 2023, for evidence that found a parietal rather than a frontal distribution). This investigation is however complicated by the fact that the manipulations eliciting frontal positivities typically also elicit an increase in N400 amplitude, possibly resulting in spatiotemporal overlap between components of opposite sign (but see Chapter 4 for a potential case with no overlapping N400 effect). While it has been argued that the frontal and parietal late positivities index functionally distinct cognitive processes (Kuperberg et al., 2020; Van Petten & Luka, 2012), the interpretations assigned to them are often similar (see Kuperberg et al., 2020, for an overview) and frontal and posteriorly peaking positivities may be argued to be part of a family of late positivities that index related aspects of integrative language processing (see Brouwer et al., 2012).

2.3 From ERPs to Theories

In summary, both the N400 and the P600 components are modulated by many linguistic and non-linguistic manipulations. Due to the multitude of influences on each component, identifying which specific cognitive process underlies the N400 and the P600 has proven difficult. Over the years, several major functional interpretations

of the N400 and the P600, as well as computational models thereof, have been proposed.

2.3.1 Interpreting the N400

The first major interpretation of the N400 component posited that it is an index of semantic integration (Brown & Hagoort, 1993, 2000; Hagoort et al., 2004), which is also referred to as post-lexical integration (Kutas & Federmeier, 2011) or unification (Hagoort et al., 2009). That is, the N400 was taken to index the incremental update of an unfolding utterance meaning representation with novel information. This interpretation accounts for the strong relation of N400 amplitude to word expectancy, as integrating novel information into an utterance meaning representation should be more effortful for unexpected than for expected words ("He spread the warm bread with socks/butter").

However, Semantic P600 results pose a challenge for a strong integration view of the N400: Studies contrasting plausible with implausible sentences, e.g., by reversing thematic roles, did not induce an N400 effect. This gave rise to a set of multi-stream models (Bornkessel-Schlesewsky & Schlesewsky, 2008; A. Kim & Osterhout, 2005; Kos et al., 2010; Kuperberg, 2007; Michalon & Baggio, 2019; van Herten et al., 2006; van Herten et al., 2005), which maintain that the N400 is an index of integrative processing by explaining the absence of the N400 as the result of a semantic illusion (a term adapted from Erickson & Mattson, 1981), i.e., the idea that comprehenders temporarily perceive the input (e.g., "the javelin has the athletes thrown") as semantically plausible (e.g., by constructing the interpretation that the athletes threw the javelin; see Figure 2.2). Critically, the explanation of absent N400s as a result of a semantic illusion requires that the integrative processing underlying this component is agnostic to morpho-syntactic constraints, e.g., by ignoring structural cues about which nouns take the agent and patient roles in role reversals. In the plausibility-driven integrative process that multi-stream models take to underlie the N400, meaning is instead computed by applying a plausibility heuristic that operates only on the content words of the input (e.g., "athletes + javelin + throwing"; see Figure 2.2). Thus, the absence of N400 effects for implausible items critically hinges on the availability of a semantically plausible alternative interpretation, computed, e.g., by reversing thematic roles to form a plausible interpretation. Indeed, this plausibility-based meaning construction can be understood as an instance of *shallow*/"*good-enough*" processing (Ferreira, 2003; Ferreira et al., 2002; Ferreira & Patson, 2007) or "quasi-compositional" integration (Rabovsky and McClelland, 2020; see also the models of Li and Ettinger, 2023; Ryskin et al., 2021, for a formulation within a noisy-channel framework). The linking of the N400 to a notion of integration based on a plausibility heuristic has, however, been further challenged by studies demonstrating that N400 effects for implausible relative to plausible target

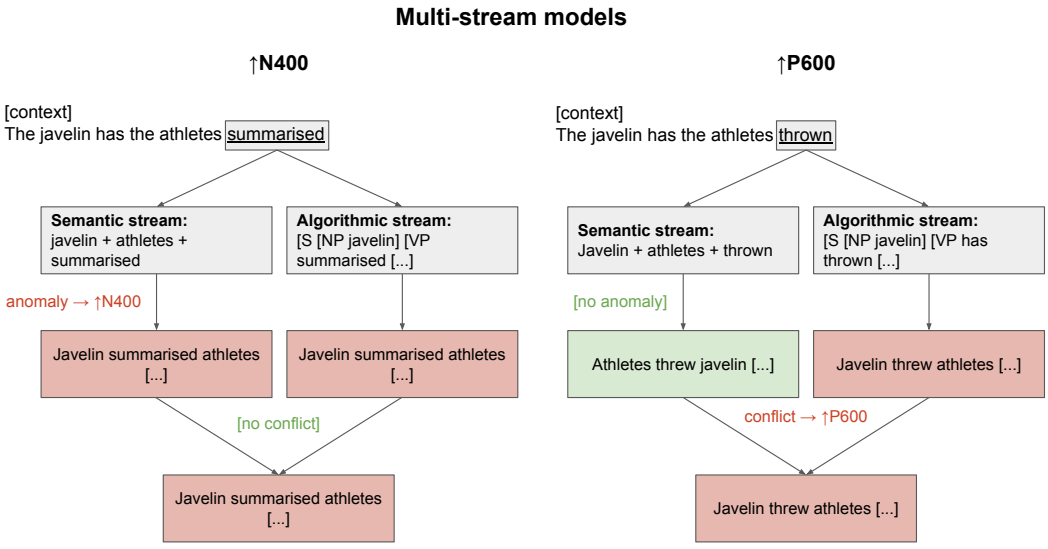


FIGURE 2.2: Abstracted schematic of multi-stream accounts of the N400 effect and the P600 effect. The precise terminology and mechanisms different multi-stream models propose for the *semantic stream* and the *algorithmic stream* vary. Stimuli are examples from two conditions in Hoeks et al. (2004) who observed a biphasic effect for "the javelin has the athletes summarised" and a P600 effect for "The javelin has the athletes thrown" relative to the baseline condition "The javelin was by the athletes thrown".

words can be absent, even if there is no semantically plausible alternative interpretation (Chow and Phillips, 2013; Delogu et al., 2019; Nieuwland and van Berkum, 2005; see Chapter 4 for discussion and novel data).

Importantly, both notions of integration ascribed to the N400 (either constrained by morpho-syntax or not) are not obviously reconcilable with N400 results from single-word processing studies. For instance, as reviewed above, word or syllable frequency modulate N400 amplitude for individually presented words and so do words and pseudowords with varying orthographic neighbourhood sizes. It is not evident why individually presented words should trigger attempts at meaning integration and why meaning integration would be modulated by factors such as frequency and orthographic neighbourhood size (see van Berkum, 2009, for discussion).

The second major interpretation of the N400, the retrieval or semantic access view of the N400 (Brouwer et al., 2012; Kutas & Federmeier, 2000; Lau et al., 2009; Lau et al., 2008; van Berkum, 2009, 2010), is indeed strongly informed by these and similar results from single-word presentation and priming studies. Retrieval is posited as the process by which word forms trigger access of word meaning, the conceptual knowledge associated with a word form, in long-term memory. On the retrieval view, N400 modulations elicited by frequency and semantic relatedness are

straightforwardly explained as gradual facilitation of semantic access: Word meanings of frequent words are more easily accessed, and presenting a semantically related prime word may partly pre-activate the word meaning of target words. The orthographic neighbourhood size effect of pseudowords offers an interesting dimension of the retrieval view of the N400, namely that words are not identified first before an attempt at semantic access is made. Strongly informed by single-word presentation studies are two computational models that account for the retrieval process thought to underlie the N400: The Semantic Activation Model (Cheyette & Plaut, 2017; Laszlo & Armstrong, 2014; Laszlo & Plaut, 2012) and the Semantic Attractor model (Rabovsky & McRae, 2014).

The retrieval view has also been applied to account for data from sentence/discourse comprehension studies. As previously reviewed, it was found that during sentence processing, message level constraints often appear to start overriding single-word features such as frequency or relatedness, and, generally, word expectancy is one of the strongest determinants of N400 amplitude. The strong relation of the N400 to expectancy is however not at odds with a retrieval view of the N400: Indeed, the degree to which words can be expected, given the context, could very plausibly facilitate the effort involved in accessing word meaning in long-term memory. For instance, in the neurocomputational instantiation of RI theory (Brouwer and Crocker, 2017; Brouwer, Delogu, Venhuizen, and Crocker, 2021) these message-level influences on the N400 are posited to occur as a result of an incrementally constructed utterance meaning representation that pre-activates aspects of upcoming word meaning. Strikingly, the retrieval view of the N400 is compatible with the absence of N400 effects in semantic illusion studies: In cases of role-reversals, the target word may, in fact, have been similarly primed by both contexts ("the javelin has the athletes thrown" vs. "the javelin was by the athletes thrown"), resulting in equal retrieval facilitation. In general, on the retrieval view, N400 amplitudes are predicted to vary whenever target word meaning is differentially preactivated. As a variant of this retrieval view, Debruille (2007) argue that the N400 indexes the inhibition of semantic features after stimulus presentation rather than their activation – a view which is at odds with the finding that constraint does not modulate the N400 when word expectancy is held constant (Federmeier et al., 2007, see Kutas and Federmeier, 2011, for discussion).

On a third major account, the "hybrid" account (Baggio & Hagoort, 2011; Lau et al., 2016; Nieuwland et al., 2020), the N400 is taken to index both retrieval and integration (refered to as "pre-activation" and "unification", respectively, by Baggio & Hagoort, 2011). Whereas the pure integration view of the N400 is difficult to reconcile with the N400 data from single-word presentation studies, the hybrid view can explain these findings in terms of retrieval, while maintaining that the N400 also indexes integration. However, the hybrid view does not directly account for the absence of N400 effects in Semantic P600 studies: Even when conditions are

matched for priming and thus equally facilitate retrieval, the difference in plausibility between "the javelin has the athletes thrown" and "the javelin was by the athletes thrown" should lead to differential integration difficulty that should result in a difference in N400 amplitude. Hence, in order to account for the absence of the N400 effect in Semantic P600 studies, the hybrid view of the N400 would have to adopt the structure-insensitive notion of integration that is also assumed by multi-stream models. A final interpretation of the N400 that also does not fall clearly in either retrieval or integration are proposals that relate the N400 to predictive coding. Under predictive coding (Friston, 2005), stimulus-induced activation and error signals are passed up and down cortical hierarchies. Indeed, a recent proposal employed a predictive coding perspective to account for the sensitivities of the N400 component (Eddine et al., 2022). It remains to be seen how the core building blocks of retrieval and integration can be implemented within a predictive coding architecture.

2.3.2 Interpreting the P600

The wide array of P600 results for different manipulations has caused a lively debate about the functional interpretation of the P600. In response to the original results elicited by syntactic manipulations, the P600 has been interpreted as a marker for syntax indexing the revision, repair, or re-analysis of (morpho-)syntactic structure (Friederici, 1995; Hagoort et al., 1999; Osterhout & Holcomb, 1992). Generalising these ideas, Kaan et al. (2000) and Kaan and Swaab (2003) proposed the P600 as an index of syntactic integration, i.e., as a processing index that reflects the effort involved in syntactic structure building in general, rather than reflecting revision processes that ensue only if syntactic analysis was unsuccessful (see also Fitz & Chang, 2019, for a computational model). Critically, however, Semantic P600 studies challenged not only the interpretation of the N400 as an index of semantic integration but also the proposal that the P600 uniquely indexes syntactic integration. An initial result by A. Kim and Osterhout (2005, "the hearty meal was devoured/devouring") was still interpreted through the lens of syntax by postulating that accepting a semantically plausible reading leads the input to be perceived as syntactically ill-formed, thus inducing a P600 effect relative to baseline. Further results, however, clearly exclude an explanation of the observed P600 effects as a reflection of syntactic processing difficulty: "the javelin has the athletes thrown" is syntactically as well formed as the control condition "the javelin was by the athletes thrown" (Hoeks et al., 2004).

In response to Semantic P600 studies, several theories have been proposed on which the P600 is taken to index conflict monitoring, conflict resolution, or a revision mechanism (Bornkessel-Schlesewsky and Schlesewsky, 2008; A. Kim and Osterhout, 2005; Kos et al., 2010; Kuperberg, 2007; van Herten et al., 2005), within a multi-stream architecture (see Figure 2.2, right). Crucially, on these multi-stream accounts, the absence of an N400 increase for a semantically implausible target word, induced, for instance, by the presence of a semantically attractive alternative interpretation, is a

necessary condition for observing a P600 increase: An incongruous target word that passes the semantic stream unnoticed will be detected in the structure-sensitive algorithmic stream. Importantly, the P600 is not taken to index processing difficulty in the algorithmic processing stream directly; in fact, multi-stream models do not propose a direct neural correlate for the algorithmic stream. Rather, the conflict between the analyses generated by the two streams is taken to be reflected by the P600, and, hence, only if the incongruous word has been analysed as plausible in the semantic stream but not in the algorithmic stream should a P600 increase ensue. More recent work (Rabovsky & McClelland, 2020; Ryskin et al., 2021) follows similar lines of reasoning by arguing that the P600 may index a revision mechanism that is activated when an implausible word passes the main integration stream unnoticed, effectively invoking a multi-stream architecture.

Importantly, the interpretation of the P600 as an index of conflicting analyses mostly accounts for a specific set of Semantic P600 studies (Bornkessel-Schlesewsky & Schlesewsky, 2008; A. Kim & Osterhout, 2005; Kos et al., 2010; Kuperberg, 2007; van Herten et al., 2005). While a conflict monitoring perspective may also explain P600 results from stimuli that induced *syntactically* infelicitous analyses, it is less clear how P600 increases in response to syntactically complex, but well-formed materials, relative to syntactically simple materials would be explained. Further, within the domain of semantically induced P600 increases, the multi-stream architecture, within which the conflict monitoring view of the P600 has been formulated, suggests that the absence of an N400 increase is prerequisite for an increase in P600 amplitude. However, P600 increases are often observed together with N400 increases where they manifest as biphasic effects between conditions and it is not clear how these simultaneous N400 and P600 increases would be accounted for on a multi-stream account (see Van Petten and Luka, 2012, for an overview and Brouwer et al., 2012, as well as Chapter 5 for discussion). More generally, semantic P600s were found for many semantic manipulations and appear not to be systematically tied to the presence of a semantically attractive alternative interpretation (see the review above).

Taken together, the sensitivity of the P600 to the meaning conveyed by the message – as evidenced by results from semantics, pragmatics, and syntax – has led to the interpretation of the P600 as a general index of integration (Brouwer et al., 2017; Brouwer et al., 2012). Notably, this integration view can thus also explain the full breadth of syntactic P600 findings, since syntactically violating as well as syntactically complex sentences should, generally, be more difficult to comprehend, and therefore induce more integration effort (but see Leckey et al., 2023, for a discussion of possibly distinct semantic and syntactic P600s). The integration view of the P600 bears some similarity to the algorithmic processing proposed by multi-stream models, in that integration is assumed to take into account morpho-syntactic information. However, the P600 is taken to index integration effort directly, rather than to reflect the conflict between analyses. Critically, and in contrast to multi-stream

models, the integration account of the P600 is not dependent on the availability of a semantically attractive alternative interpretation and, hence, this view is not limited to semantic P600s elicited by implausible sentences that make a semantically attractive alternative interpretation available.

2.3.3 Retrieval-Integration Theory: Predictions

The wide range of linguistic elicitations of the N400 and the P600 has led to a multitude of theoretical interpretations of the two components. Critically, these interpretations inform different language comprehension models which differ vastly in their architectural choices. Additionally, only few theories offer a unified account of *both* the N400 and the P600 that specifies the cognitive process(es) thought to underlie each component, as well as their interaction.

One theoretical account that does offer a unified model of the N400 and the P600 is Retrieval-Integration (RI) theory (Brouwer et al., 2017; Brouwer et al., 2012), which combines the retrieval view of the N400 (as also proposed by Kutas & Federmeier, 2000, 2011; Lau et al., 2009; Lau et al., 2008; van Berkum, 2009, 2010) with the novel proposal that the P600 indexes integration. That is, the N400 is taken to index the effort involved in accessing word meaning in long-term memory, whereas the P600 is taken to index the effort involved in updating an incrementally constructed utterance meaning representation with this retrieved word meaning. Critically thus, these processes are thought to directly interlock with each other in a single-stream (rather than a multi-stream) architecture (see Figure 2.3 for a representation of the computational model instantiation of the theory proposed by Brouwer et al., 2017; Brouwer, Delogu, Venhuizen, and Crocker, 2021).

The three following empirical investigations test key predictions of RI theory for the role of the N400 and the P600 in incremental, expectation-based language comprehension while providing critical contrasts to competing interpretations and accounts of the N400 and the P600, therefore seeking to offer a dissociation between competing accounts of the N400 and the P600. RI theory assumes that both retrieval and integration are strongly constrained by expectations generated based on the utterance meaning representation constructed so far, thus predicting both the N400 and the P600 (as well as reading times; Brouwer, Delogu, Venhuizen, and Crocker, 2021) to be modulated by expectancy. Critically, this bears a potential confound for the mapping of retrieval/integration to N400/P600, which is addressed in an experiment crossing expectancy and association (Chapter 3). For the P600, RI theory explicitly predicts that its amplitude should *continuously* index integration effort. We examine this question in an experiment that manipulates utterance meaning plausibility on three levels, while also testing diverging predictions made by multi-stream models (Chapter 4). Due to the central role of expectancy for both retrieval and integration, RI theory predicts that unexpected words should be more effortful to retrieve *and* more effortful to integrate. Hence, N400 amplitude and P600 amplitude

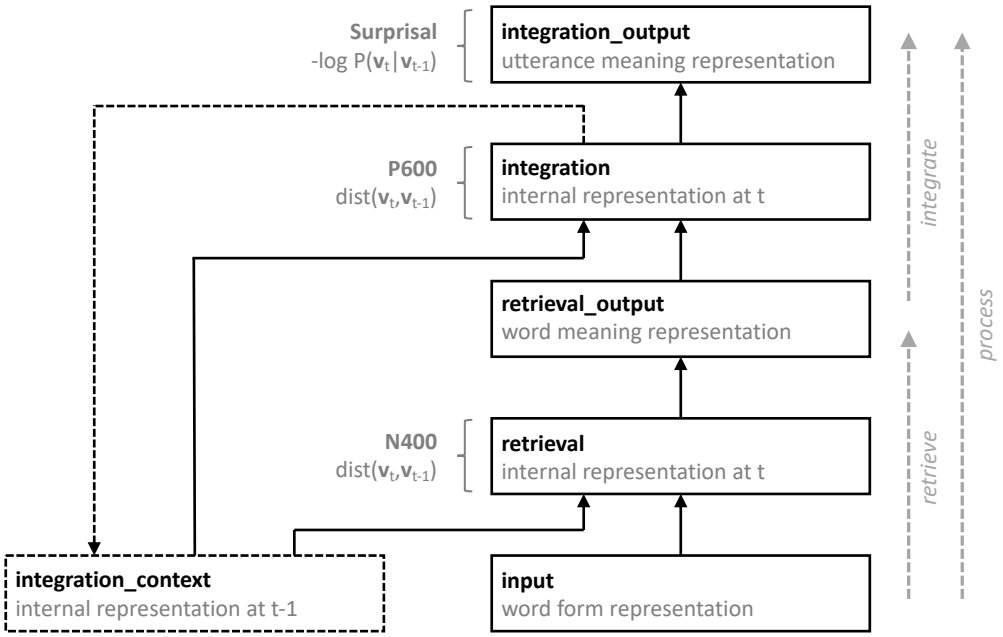


FIGURE 2.3: Schematic architecture of the neurocomputational instantiation of the Retrieval-Integration theory, implementing word-by-word language processing. Figure adapted from Brouwer, Delogu, Venhuizen, and Crocker (2021).

should be negatively correlated *within-trial*. As we will argue, this prediction directly contradicts the single-trial dynamics proposed by multi-stream models. We address this research question in Chapter 5 by applying a novel statistical analysis approach to the data obtained for Chapter 3. Finally, the sum of the results is discussed with regard to the overarching goal of dissociating the RI account of the N400 and the P600 from competing theories (Chapter 6).

Chapter 3

Retrieval (N400) and Integration (P600) in Expectation-Based Comprehension

The contents of this chapter, with the exception of Experiment 2, were published in a peer-reviewed journal article (Aurnhammer et al., 2021).

3.1 Introduction

Theories of sentence comprehension have recently focused on expectation-based processing and the notion of surprisal (Hale, 2001; Kuperberg & Jaeger, 2016; Levy, 2008; Venhuizen et al., 2019). Surprisal theory posits that the cognitive effort induced by a word is proportional to its expectancy in context, and has been shown to account for a wide spectrum of behavioural processing phenomena (Aurnhammer & Frank, 2019a, 2019b; Boston et al., 2008; Brouwer et al., 2010; Demberg & Keller, 2008; Frank, 2009; Hale, 2001; Levy, 2008; Roark et al., 2009; Smith & Levy, 2008). Crucially, however, properties of words other than their expectancy, such as the association of a word with the preceding context (Kutas & Federmeier, 2011), are known to also influence online indices of comprehension. Given the central role of expectancy in current theories and the linking hypothesis of surprisal theory, an important open question is whether it is possible to identify processing correlates that are specifically sensitive to expectancy/surprisal and insensitive to association, as well as the time course of these neural and behavioural correlates.

In the electrophysiological domain, expectancy-related measures, such as surprisal and cloze probability, have typically been linked to the N400 component (Delogu et al., 2017; DeLong et al., 2005; Frank et al., 2015; Kutas et al., 1984), a negative voltage deflection peaking around 400 milliseconds post stimulus onset, the amplitude of which is inversely related to the expectedness of a word in context. The N400 is, however, sensitive to many other linguistic (and non-linguistic) factors beyond expectancy as well, such as frequency (Van Petten & Kutas, 1990), orthographic

neighbourhood size (Laszlo & Federmeier, 2009, 2011), and lexical association (Kutas, 1993). As a consequence, many studies that have been interpreted as evidence for expectancy effects - based for example on manipulations of cloze or n-gram probability - are confounded with simple association. For instance, in the sentence manipulation "He spread the warm bread with socks/butter" (Kutas & Hillyard, 1980), the word "socks" is not only unexpected with regard to the meaning of the entire sentence, but it is also not related semantically. That is, "socks" is semantically unassociated to the prior context words, irrespective of their compositional meaning as an utterance, whereas the other target word, "butter", is both semantically expected and associated. In consequence, the N400 has functionally been interpreted as reflecting semantic integration (Brown & Hagoort, 1993, 2000; Hagoort et al., 2004), lexical retrieval (Brouwer et al., 2012; Kutas & Federmeier, 2000, 2011; Lau et al., 2009; Lau et al., 2008; van Berkum, 2009, 2010), or both integration and retrieval on more recent "hybrid" accounts (Baggio & Hagoort, 2011; Lau et al., 2016; Nieuwland et al., 2020).

Another salient component of the event-related potential (ERP) signal is the P600, a positive going shift becoming apparent from around 500 milliseconds post stimulus onset, which has initially been identified as a component that is sensitive to structural processing. Theories of the P600 have associated it with the reanalysis of existing (morpho-)syntactic structure (e.g., Friederici, 1995; Hagoort et al., 1999; Osterhout & Holcomb, 1992), with syntactic integration difficulty (e.g., Kaan et al., 2000; Kaan & Swaab, 2003), conflict monitoring/resolution (Bornkessel-Schlesewsky & Schlesewsky, 2008; A. Kim & Osterhout, 2005; Kos et al., 2010; Kuperberg, 2007; van Herten et al., 2005), and more recently with semantic integration processes (Brouwer et al., 2017; Brouwer et al., 2012).

The retrieval view of the N400 and the semantic integration account of the P600 are at the core of the Retrieval-Integration (RI) account of language comprehension (Brouwer et al., 2017; Brouwer et al., 2012; Brouwer & Hoeks, 2013; Hoeks & Brouwer, 2014) and RI theory predicts these two components to be differentially affected by association and expectancy. As a specific case of general memory retrieval, lexical retrieval is the process by which the meaning of a word is accessed in long-term memory and, on RI theory, is taken to be indexed by the N400. As such, the sensitivity of the N400 to linguistic properties like frequency, orthographic neighbourhood size, as well as association and expectancy, is explained by the influence of these properties on the ease with which word meanings are retrieved. In particular, words that are associated with the prior context, or that are more expected given the unfolding utterance interpretation, are easier to retrieve from long-term memory. Integration, on the other hand, is linked to the P600. Integrative processing is conceptualised as the cognitive process that incorporates the meaning of a new word into a compositional representation of the meaning of the utterance, as constructed so far. Crucially, the resultant meaning representation is assumed to provide the relevant contextual cues for the facilitated retrieval of potentially upcoming word

meanings.

A key strength of the account is therefore that it makes simultaneous predictions regarding effects in both ERP components. In fact, the decomposition of language comprehension into retrieval and integration is made even more explicit in the computational instantiation of RI theory. In this model, retrieval is instantiated by the function

$$\textit{retrieve(word form, utterance context)} \mapsto \textit{word meaning} \quad [\sim \text{N400}] \tag{3.1}$$

which maps an incoming orthographic/acoustic *word form* onto a representation of *word meaning*, while taking the unfolding *utterance context* – the utterance meaning constructed prior to the current word – into account (Brouwer, Delogu, Venhuizen, & Crocker, 2021). The output of this function serves as input to the function

$$\textit{integrate(word meaning, utterance context)} \mapsto \textit{utterance meaning} \quad [\sim \text{P600}] \tag{3.2}$$

which integrates the retrieved *word meaning* into the unfolding *utterance context*, to produce an updated *utterance meaning*. While the *retrieve* and *integrate* functions, which, respectively, underlie the N400 and the P600 component, may both be influenced by the overall expectancy of a word, this is for different reasons. In the case of the former, it is because the expectancy of an incoming word may facilitate retrieving its meaning from long-term memory, while in the case of the latter, it affects the effort involved in updating the unfolding utterance meaning representation with this retrieved meaning.

Indeed, the effort involved in updating utterance representations has been the focus of surprisal theory. The original formalisation of surprisal theory focused on syntactic comprehension (Hale, 2001) and has been generalised as the relative entropy, or Kullback-Leibler Divergence (Kullback & Leibler, 1951), of a new probability distribution over syntactic analyses (operationalised as parse trees of a probabilistic context-free grammar) resulting from the current word, compared to the previous probability distribution (Levy, 2008). In light of this characterisation, one would thus expect structurally-induced surprisal effects, i.e., syntactic integration difficulty, to be reflected in an increase in P600 amplitude (Hagoort et al., 1993; Osterhout & Holcomb, 1992). However, building upon the considerable evidence that the P600 also indexes semantic integration difficulty (as predicted by the RI account), Venhuizen et al. (2019) have recently proposed that the P600 component more broadly indexes comprehension-centric surprisal – the negative log-probability of the utterance meaning representation after processing a word; that is, they propose that the P600 amplitude induced by an incoming word is proportional to how unlikely the interpretation is after processing this word, given the interpretation before encountering it. This surprisal measure is influenced by both linguistic experience and

knowledge about the world (Venhuizen et al., 2019). As Brouwer, Delogu, Venhuizen, and Crocker (2021) argue, this view of the P600 as reflecting comprehension-centric surprisal follows directly from RI theory. Just as syntactic models determine the likelihood of alternative analyses based on linguistic experience, the RI model recovers interpretations that reflect the distributional characteristics of the utterances it is exposed to (Brouwer, Delogu, Venhuizen, & Crocker, 2021).

The most recent instantiation of RI theory thus predicts the P600 component of the ERP signal, which indexes the amount of effort involved in updating the unfolding utterance meaning representation with the retrieved meaning of an incoming word, to be the locus that is specifically sensitive to expectancy/surprisal effects (Brouwer, Delogu, Venhuizen, & Crocker, 2021) and insensitive to association effects. That is, integration effort is assumed to increase to the extent that the utterance meaning representation resulting from integrating this word meaning is semantically, pragmatically, or structurally unexpected, given the utterance meaning representation prior to integration. Given that the retrieval processes underlying the N400 are, among other factors, also sensitive to expectancy, previously reported N400 effects of surprisal are unsurprising; that is, RI generally predicts both N400 (retrieval) and P600 (integration) amplitude to increase as a function of unexpectedness (although sufficient priming can eliminate the N400 effect even for unexpected words; see below). RI theory is thus in line with the linking of surprisal to the N400 via retrieval (as also proposed by Frank et al., 2015). In sum, on the RI account, the P600, as an index of compositional, semantic, and integrative processes, should therefore be sensitive primarily to the expectancy of a new word with regard to the current utterance meaning representation, and crucially, insensitive to association. Further, the RI account predicts the N400, as an index of lexical retrieval, to be sensitive to both lexical association and expectancy.

This raises the question of how we can test the prediction that the P600 is the component that is specifically sensitive to expectancy/surprisal, while the N400 is sensitive to both association and expectancy. In the extreme case ("He spread the warm bread with butter/socks"), where the manipulations of lexical association and expectancy are completely overlapping, it is impossible to tease apart the contributions of lexical association and expectancy to the N400. At the other extreme, evidence comes from constellations in which expectation and association disagree; that is, when expectancy is low, but association is high – e.g., "De vos die op the stroper .." (lit.: "The fox that on the poacher hunted" meaning that the fox hunted the poacher) relative to "De stroper die op the vos joeg..." (lit. "the poacher that on the fox hunted...", van Herten et al., 2005) – unexpected words result in N400 amplitudes similar to expected words, showing no difference in retrieval difficulty (cf. the 'Semantic Illusion' or 'Semantic P600' literature; e.g., see Bornkessel-Schlesewsky & Schlesewsky, 2008; Brouwer et al., 2012; Delogu et al., 2019; Kuperberg, 2007, for reviews). Crucially, for both of these kinds of manipulations, P600 effects have

been observed in response to unexpected words (for an overview, see Bornkessel-Schlesewsky & Schlesewsky, 2008; Brouwer et al., 2012; Hoeks et al., 2004; Kuperberg, 2007).

An open question, however, is how precisely association and expectancy combine in affecting N400 amplitude; that is, the picture that emerges from studies investigating the combination of association and expectancy between these extremes is less clear. Some studies found that association has no influence when the sentence is incongruent (Camblin et al., 2007; Khachatryan et al., 2014; Khachatryan et al., 2018). Others, by contrast, found a stronger effect of association for incongruent targets, when presented to the right visual field (left hemisphere, Coulson et al., 2005). Similarly, it was found that in syntactically correct but not meaningful sentences, word associations do play a role for the N400 (Schwartz et al., 2003; Van Petten, 1993; Van Petten et al., 1997). Further, a reduction in N400 amplitude was observed for event-related compared to event-unrelated contextually anomalous target words (Metusalem et al., 2012). Indeed, arguments against the role of association in semantic violations contrast starkly with the results observed in the aforementioned literature in which high association eliminates an N400 effect for unexpected words (e.g., Delogu et al., 2019, where high association leads an otherwise contextually improbable target word to not increase N400 amplitude). Other studies focused on specific aspects like visual half-field paradigms (Coulson et al., 2005), individual differences (Boudewyn et al., 2012), or later processing stages (Camblin et al., 2007). The existing literature thus paints an inconclusive picture of the influences of expectancy and lexical association on ERPs: On the one hand, studies have found that lexical association effects are attenuated for incongruent target words, on the other hand, studies found that association is relevant even for these incongruent target words.

To assess how expectancy and lexical association affect retrieval and integration, we created an experimental design that crosses these stimulus properties, while aiming to minimise the confounding of expectancy and lexical association. To achieve this, we maximise the orthogonality of the two manipulations in a context manipulation design (Table 3.1) that manipulates strong (A+) and weak (A-) lexical association differentially by means of an intervening adverbial clause, for both expected and unexpected target words by using main verbs that either do (“sharpened”) or do not (“ate”) take the target word (“axe”) as a semantically fitting and expected direct object. While this manipulation of expectancy necessarily covaries with lexical association (analogous to Kutas and Hillyard, 1980), the additional – independent – manipulation of lexical association is achieved by using an intervening adverbial clause (“before he the wood stacked”/“before he the movie watched”). This adverbial clause contains words that either are or are not related to the target word, without changing the overall expectancy of the target word that is established by the main clause.

A: A+E+
Gestern schärfte der Holzfäller, bevor er das Holz stapelte, die Axt...
(*Yesterday sharpened the lumberjack, before he the wood stacked, the axe...*)

B: A-E+
Gestern schärfte der Holzfäller, bevor er den Film schaute, die Axt...
(*Yesterday sharpened the lumberjack, before he the movie watched, the axe...*)

C: A+E-
Gestern aß der Holzfäller, bevor er das Holz stapelte, die Axt...
(*Yesterday ate the lumberjack, before he the wood stacked, the axe...*)

D: A-E-
Gestern aß der Holzfäller, bevor er den Film schaute, die Axt...
(*Yesterday ate the lumberjack, before he the movie watched, the axe...*)

TABLE 3.1: Example item crossing the factors expectancy (E+-) and lexical association (A+-). Literal translations preserving the original word order are given in italics.

Importantly, and unlike previous studies, the association manipulation is completely independent of the expectancy manipulation, such that there is no dependency between the manipulated adverbial clause and the target word. Further, we choose a particularly strong expectancy manipulation in the form of a selectional restriction violation. This allows us to assess if expected target words that are less associated to the context, nonetheless produce an increase in N400 amplitude relative to associated and expected ones, and conversely, whether unexpected but associated targets have attenuated N400 amplitude relative to unexpected and unassociated ones. Furthermore, this strong expectation violation is intended to maximise the observability of both N400 and P600 effects in the face of spatiotemporal component overlap. That is, as demonstrated by Brouwer, Delogu, and Crocker (2021) and Delogu et al. (2019, 2021), because of spatiotemporal component overlap (Luck, 2005) – the summation of, and potential cancellation of the scalp-recorded activity from different neural generators – expected integration effects on P600 amplitude may sometimes be attenuated by a large, preceding N400, thereby not yielding a reliable effect in the average waveforms (see Brouwer & Crocker, 2017, for discussion). In order to maximise inferences about P600 modulation it is therefore important to address such spatiotemporal component overlap in both analyses (Brouwer, Delogu, & Crocker, 2021) and experimental designs (Delogu et al., 2021). The strong expectation violation is thus intended to attenuate the effects of spatiotemporal component overlap, in which the large predicted N400 amplitude for unexpected targets might otherwise obscure the effect of our manipulation with regard to P600 amplitude.

The materials were presented in two experiments: an ERP study and a web-based self-paced reading (SPR) study. RI theory, as an integrated theory of both the N400 and the P600, predicts N400 effects of retrieval facilitation due to both lexical

association (Condition A relative to B, and C to D) and expectancy (Condition A relative to C, and B to D). Crucially, for the P600, RI theory predicts only an effect of expectancy (again, Conditions A/B compared to C/D). The self-paced reading study was conducted to obtain behavioural correlates for the same items. Based on surprisal theory, we predict clear effects of expectancy, which – under the RI account – should pattern with the P600. Additionally, we can assess whether there is any additional influence of association on reading times, and compare the relative influence of the two factors in the critical and Spillover regions. We will elaborate on the results based on the integrated predictions of RI theory for the N400, the P600, and reading times, and based on the individual predictions of other theories.

3.2 Experiment 1: Event-Related Potentials

3.2.1 Method

Participants

Forty-nine participants from Saarland University took part in the ERP experiment, nine of which were excluded due to excessive artefacts or technical problems during recording. The final forty participants (mean age 23; SD: 2.96; age range 19-29; 6 male) were all right-handed, native speakers of German (12 early bilinguals). All participants had normal or corrected-to-normal vision and none of them reported any form of color blindness. They gave informed, written consent and were paid 20€ for taking part in the experiment.

Materials

The full list of final materials is available in Appendix B.1. We initially created 140 sentence quadruplets following the context-manipulation design exemplified in Table 3.1. To manipulate lexical association independently of expectancy, the target word ("axe") was preceded by an adverbial clause containing lexical material that either was ("before he *the wood stacked*" in A & C) or was not ("before he *the movie watched*" in B & D) lexically associated to the target. In order to rule out an interpretation of the resulting ERPs in terms of shallow processing (Rabovsky & McClelland, 2020) or good-enough representation (Ferreira, 2003; Ferreira et al., 2002), adverbial clauses were created such that no structural or thematic dependency of the target word with the adverbial clause was supported. Further, the adverbial clause did not allow for a role-reversal reading, i.e., there was no ambiguity about the correct assignment of agent and patient roles, thus avoiding so-called semantic illusion effects (see Bornkessel-Schlesewsky and Schlesewsky, 2008; Brouwer et al., 2012; Kuperberg, 2007, for overviews). Further, unambiguous readings were ensured by the use of definitive articles marked uniquely as nominative and accusative, respectively.

Expectancy, the second experimental factor, was manipulated by using a main clause verb that renders the target word either an expected ("*sharpened* the lumberjack ... the axe" in A & B) or an unexpected direct object continuation ("*ate* the lumberjack ... the axe" in C & D), given its selectional restrictions. To rule out any explanation of the observed ERP modulations in terms of syntactic processing difficulty, the target word and the main verb matched grammatically and in the preferred subcategorisation frame of the verb. Further, we avoided verbs with a preference for object-drop. The resulting match or mismatch between the main clause verb (*sharpened/ate*) and the target (axe) was thus purely selectional. We also avoided animacy violations, which have previously led to stronger P600 effects than other types of semantic violations (Szewczyk & Schriefers, 2011). Finally, to rule out interpretations of potentially observed P600 effects as reflecting prediction errors in unexpected targets (DeLong et al., 2011; Federmeier et al., 2007; Kuperberg et al., 2020; Otten & van Berkum, 2008; Vissers et al., 2006), we selected main clause verbs that did not create high expectations for a specific object noun (as validated in the cloze norming study reported below).

Each item ended with additional material following the target word (e.g., "and chopped the logs" for our archetypal item) to avoid sentence-final wrap-up effects on the target (even though their importance has been discussed as largely overstated by Stowe et al., 2018). More importantly, this additional material allows us to detect potential spillover effects in the follow-up self-paced reading experiments reported in Section 3.3. We also included 120 filler sentences, part of which were adapted from another study (Delogu et al., 2019). Half of the fillers were plausible and half implausible, matching the proportion of expected and unexpected target words in the experimental sentences. A portion of the fillers included adverbial clauses with unexpected words that made the described scenario implausible to increase attention to the (always plausible) adverbial clause of the experimental items.

Cloze Norming In order to validate the expectancy manipulation achieved through our pre-selected *main verb – target word* pairs, we collected cloze data for the experimental sentences in a web-based experiment. The experiment was implemented using the experimental software Ibex (Drummond, 2012). Forty-eight native speakers of German were recruited through Prolific Academic Ltd. (Prolific, 2021) and compensated with 8€ per hour. Participants gave informed consent by agreeing to the written study conditions. They were instructed to complete the sentence fragment that was presented up to, but not including the article of the target noun and, hence, we did not provide grammatical cues constraining for potential target nouns. The sentences were divided into four lists according to a Latin square design, such that each participant was presented with an equal amount of sentences in each of the four conditions, totalling 140 trials per person. Participants could enter as many words

Condition	Mean	SD	Range	Mean	SD	Range
	Cloze Probability			Noun–target Association		
A	0.67	0.23	0.17 - 1.00	6.29	0.82	1.90 - 7.00
B	0.64	0.23	0.17 - 1.00	2.09	1.01	1.00 - 5.70
C	0.01	0.03	0.00 - 0.17	6.29	0.82	1.90 - 7.00
D	0.01	0.03	0.00 - 0.17	2.09	1.01	1.00 - 5.70
	Main verb–target Association			Verb–target Association		
A	6.25	0.81	2.27 - 7.00	3.23	1.59	1.00 - 7.00
B	6.25	0.81	2.27 - 7.00	1.87	0.94	1.00 - 5.40
C	1.65	0.84	1.00 - 5.00	3.23	1.59	1.00 - 7.00
D	1.65	0.84	1.00 - 5.00	1.87	0.94	1.00 - 5.40

TABLE 3.2: Descriptive statistics of the results of the cloze probability (scale 0-1) and the association rating (scale 1-7) norming studies.

as they wished but were shown example items with simple *article+target* and *preposition+article+target* completions. The 140 experimental items were randomly interleaved with 70 filler sentences. For 12 items, the two unexpected conditions (C/D) produced high-Cloze completions (different from targets), indicating that these sentence fragments were highly constraining towards predicting a specific lexical item. We changed the main clause verbs of these sentences to achieve a more uniform cloze profile, i.e., we avoided contexts for implausible items that raise expectations for a specific plausible word. These modified sentences were presented in a Cloze test with new participants. Based on the results of the Cloze studies, we selected the final 120 experimental items in such a way that the difference in Cloze probability between expected and unexpected targets (i.e., A&B vs. C&D conditions) was maximised and the variability within high- and low-cloze targets was reduced (i.e., A vs. B and C vs. D conditions). The cloze probabilities of the target for the final set of items in the four conditions are presented in Table 3.2 and Figure 3.1. The non-zero Cloze for unexpected targets resulted from a very conservative approach in which the target word was counted even if it occurred as part of a compound noun or was produced in sentential positions other than the object of the main verb.

Association Norming In a second, web-based validation study, we aimed to quantify the lexical association of the target words with the lexical material appearing in the preceding adverbial clause.[1] To this end, we presented participants with word pairs and asked them to rate how associated they were on a 1-7 scale (7 meaning *highly associated*). The experiment was conducted using Ibex (Drummond, 2012). We presented participants with each content word in the adverbial clause (e.g., the noun

[1]We also computed two corpus-based word similarity metrics, GloVe (Pennington et al., 2014) and Word2Vec (Mikolov et al., 2013). However, inspection of the results yielded many items on which the association values were strongly at odds with our intuitions, suggesting that the corpus-based metrics were unreliable for the task at hand. We therefore considered human association ratings as the gold standard, which is common practice in many studies evaluating corpus-based metrics.

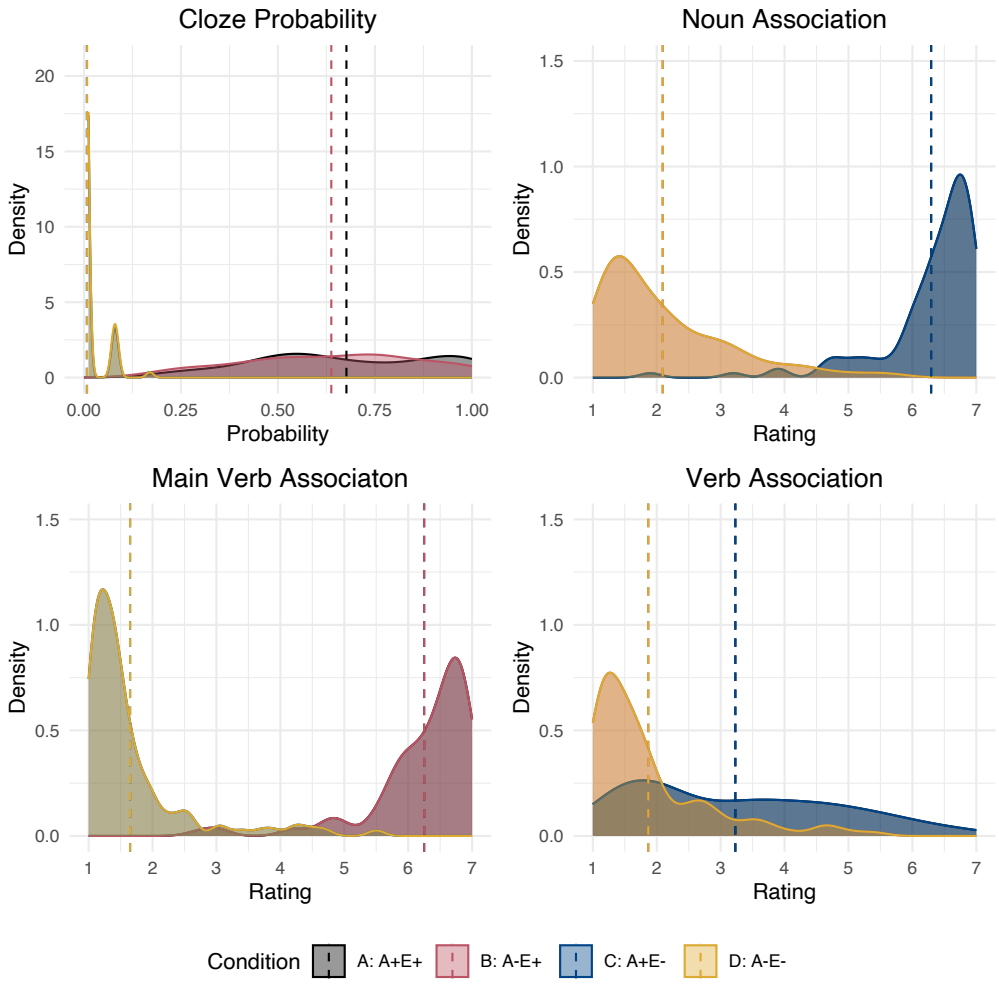


FIGURE 3.1: Density plots showing the per-condition distributions of cloze probability and association ratings collected in the norming studies. Vertical lines indicate per-condition averages. Two conditions with identical word pairs are overlapping in the association graphs.

and the verb in "before he the wood stacked"/"before he the movie watched") and the target ("axe"). Since the expectancy manipulation is achieved by using a different main clause verb ("sharpen" vs. "eat"), we collected association ratings also for these verbs and the target. Note that participants only rated word pairs, but never saw their source sentences, nor did they know that the words would be appearing in a sentence together. Sixty native speakers of German recruited through Prolific Academic Ltd. took part in the study. They did not take part in any other experiment reported in this Chapter and were compensated 11.50€ per hour. Participants gave informed consent by agreeing to the written study conditions. Stimuli were divided into six lists such that each participant saw one and only one of the *context word-target word* pairs for each item, resulting in 120 trials per participant. Association ratings for the three word pairs in each condition are summarised in Table

	Cloze	Main Verb Assoc.	Noun Assoc.	Verb Assoc.
Cloze	1.000			
Main Verb Assoc.	0.851	1.000		
Noun Assoc.	0.029	0.008	1.000	
Verb Assoc.	-0.005	0.001	0.467	1.000

TABLE 3.3: Correlations between stimulus properties.

3.2 and Figure 3.1. Words in the adverbial clause were more associated to the target in conditions A & C than in conditions B & D. The difference was stronger for the nouns than for the verbs of the adverbial clause. Association scores for the two main clause verbs also differed, such that expected targets were highly associated to the main verb compared to unexpected targets. *Main verb – target* association was strongly correlated with cloze probability (see Table 3.3). To avoid multicollinearity problems in our statistical models, we did not include main verb association in our analyses.

Procedure

The electroencephalogram (EEG) was recorded while participants were seated in a soundproof, electromagnetically shielded, and dimly lit chamber. Sentences were presented to the participants using rapid serial visual presentation (RSVP) in E-prime 2 (Schneider et al., 2002). Participants first practised with six items, half of which contained unexpected words. After the practice session, the experiment was conducted in three blocks of 80 sentences each, in which we presented the items in pseudorandomised order, and with breaks between the blocks. Participants pressed a button to start the trial and a fixation cross appeared in the centre of the screen for 750 ms. Next, each word of the sentence was presented centrally for 350 ms with a 150 ms inter-stimulus interval. Participants were then asked to judge the plausibility of the sentence by pressing one of two buttons (mapping to *yes*/*no*). The position of the *correct* and *incorrect* buttons varied randomly in order to avoid motor preparation effects. The position of the *correct*/*incorrect* buttons was indicated by the position on the screen of the words *Yes* and *No*, which were highlighted in green and red, respectively, to make them more salient.

Electrophysiological Recording and Processing

The EEG was recorded by 26 active Ag/AgCl scalp electrodes, using the standard 10-20 system. During recording, FCz was used as online reference and AFz as ground. Data were digitised at a sampling rate of 500 Hz. Eye movement artefacts were monitored through the electrooculogram of two electrodes placed horizontally at the outer canthi of each eye and two electrodes placed vertically above and below the left eye. Impedances were kept below 5 $k\Omega$ on scalp electrodes and below 10 $k\Omega$ on eye electrodes. No online filtering was applied. The EEG was re-referenced offline to

the average of the left and right mastoid electrodes and band-pass filtered between 0.01 and 30 Hz. Epochs starting 200 ms preceding the onset of the target word and lasting until 1200 ms following target onset were extracted from the EEG signal. We excluded 759 out of 4800 trials (15.81%) with ocular and muscular artefacts using a semi-automatic procedure. Baseline correction was performed on the 200 ms pre-stimulus interval.

Analysis

We analysed the data using a regression-based ERP estimation technique (rERPs Smith & Kutas, 2015a). This technique allows us to replace each individual scalp-recorded voltage with a voltage estimate from a regression model that optimally combines the manipulated variables (e.g., Cloze probability and association) to explain the variance in the signal (see also Brouwer, Delogu, & Crocker, 2021). Thus, applying this technique results in the decomposition of each observed scalp-recorded voltage into the contribution made by different experimentally manipulated factors. In the traditional rERP framework, one regression model is fitted for each time point, electrode, and subject. We apply a variation of this technique by replacing the `n` models fitted for `n` subjects at each electrode and time point with a single linear mixed effects regression (LMER) model at each electrode and time point (see Brouwer, Delogu, and Crocker, 2021, for discussion and Frank and Willems, 2017; Troyer et al., 2020; Urbach et al., 2020, for prior work using this method).

That is, rather than fitting one model for each subject, we fit only a single linear mixed model that captures per-subject variability as a random effect. As an extension, per-item variability can straightforwardly be modelled in the same regression equation, by introducing per-item random effects. Thus, the general model specification becomes

$$y_{et} = \beta_{0et} + S_{0s} + I_{0i} + \sum_{j=1}^{N} (\beta_{jet} + S_{js} + I_{ji}) x_j + \epsilon_{et} \tag{3.3}$$

where separate models are fitted for each electrode e and time sample t, and where S and I refer to random effects for subjects and items, respectively. Random intercepts are represented by S_{0s} and I_{0i}. For each predictor X_j, random slopes S_{js} and I_{ji} will be computed. The ϵ term represents the residual error, i.e., the unexplained variance in the data, for each electrode and time sample. This approach effectively distributes the multi-dimensionality of the dependent variable (in space and time) across separate statistical models, while the intra-experimental variability (across subjects and items) is modelled within each model. To distinguish this approach from the rERP technique described in Smith and Kutas (2015a), we label it lmerERP. In a nutshell, this approach allows us to (1) generate model-estimated ERP waveforms for each electrode and time sample and inspect them visually, (2)

quantify the fit of the models to the data by inspecting the residual error, i.e., the difference between observed and estimated voltages between conditions (the closer this difference is to 0, the better the fit of the estimates to the observed voltages), (3) inspect model coefficients for each time sample and electrode, and (4) inspect effect sizes (z-values) and assess statistical significance on each time sample and electrode. Data analysis was conducted using the `MixedModels` package for `Julia` (Bezanson et al., 2017). The analyses were performed on data from the three midline electrodes Fz, Cz, and Pz and on the time samples between 200 ms prior to stimulus onset and 1200 ms following it. Continuous predictors were the Cloze probabilities and association ratings (both noun-target and verb-target association, for nouns and verbs appearing in the adverbial clause) collected during pre-testing. Predictors were always included as fixed effects and as per-subject and per-item random slopes. Since predictors were z-standardized, the model coefficients represent the change in voltage associated with 1 standard deviation increase in the predictor, for each time sample and electrode. To make model interpretation more intuitive, we inverted the predictors, by multiplying each predictor with -1. This results in the coefficients' sign matching the sign of the predicted ERP deflection. Data analysis proceeded as follows. First, we aimed to maximize the fit of the two manipulated factors individually. To do so, we assessed the residuals on contrasts that differ only with respect to the predictor of interest. More specifically, Conditions A and C were used for isolating the effect of Cloze probability, as the adverbial clause is the same in these conditions and association scores are therefore constant. Conditions C and D were used to isolate the effect of association, as most items in these conditions resulted in zero Cloze probability. The data from each of these pairs of conditions was then analysed in regression models including an intercept and the single predictor of interest (as well as a random intercept and slope for this same predictor). At this stage, the effect of different data predictor transformations (e.g., logarithmic transformation) on model fit can be investigated. Finally, the data from all trials in the four conditions were re-estimated in regression models including all selected predictors. We report coefficients and corresponding z-values from this set of models. We also report the p-values for two time windows of interest: 350-450 ms (N400 time window) and 600-800 ms (P600 time window). We corrected for the inflated false-discovery rate, by controlling for multiple comparisons using the method illustrated by Benjamini and Hochberg (1995). We applied correction separately for the two time windows of interest, but across all three electrodes and time samples within each time window.

3.2.2 Results

Task: Plausibility Judgement

Participants judged the plausibility of the sentences in the four conditions as expected based on our experimental design. We considered rating Conditions A and

	Accuracy			Reaction Time		
Cond.	Mean	SD	Range	Mean	SD	Range
A	90.3%	8.1%	65.4% - 100.0%	598 ms	296 ms	223 ms - 1584 ms
B	86.2%	9.9%	65.2% - 100.0%	639 ms	267 ms	141 ms - 1378 ms
C	80.4%	14.3%	43.5% - 100.0%	611 ms	285 ms	232 ms - 1273 ms
D	85.5%	12.5%	53.6% - 100.0%	628 ms	285 ms	205 ms - 1371 ms

TABLE 3.4: Task performance on the binary plausibility ratings in the event-related potential experiment. Accuracy and reaction times were computed across subjects.

B plausible and C and D implausible as correct. Average accuracy was 85.6% (SD = 6.7%, range = 72.3 - 96.8%) with an average reaction time of 620 ms (SD = 253 ms, range = 202 - 1223 ms; both metrics computed across subjects). Means, standard deviations, and ranges of accuracy and reaction time in the four conditions are reported in Table 3.4.

ERPs

Grand-average waveforms for the four experimental conditions are displayed on all non-reference, non-eye electrodes (Figure 3.2) and on three midline electrodes. Visual inspection suggests larger negativities in response to both less associated targets (Condition B/D relative to A/C) and unexpected targets (Condition C/D relative to A/B) in the N400 time window. In the P600 time window, approximately 600 ms post stimulus onset, a positivity emerges in response to unexpected relative to expected targets on parietal electrodes. As the ERPs do not suggest differences in the ERPs across hemispheres or across laterality/mediality, our analyses focused on three midline electrodes (Figure 3.3).

Figure 3.4 shows the topographic distributions of the effects for each contrast of interest in the N400 and P600 time windows. In the N400 time window, unexpected target words elicited a larger negativity compared to the baseline, Condition A. A smaller N400 effect was also elicited by unassociated targets, within both the expected and unexpected conditions. The largest N400 amplitude is observed for targets that were both unexpected and unassociated. N400s were broadly distributed. Between 600 and 800 ms, we observed a posteriorly distributed positivity, peaking over parietal electrodes, for unexpected targets relative to expected targets. For unexpected-unassociated compared to unexpected-associated targets (Condition D relative to C), a small negativity remains, seemingly extending from the preceding N400 time window into the P600 time window.

To perform the lmerERP analyses, we first considered the single predictors individually (i.e., cloze probability, noun-target association, and verb-target association) and assessed their fit to the data as shown by the residual error (see Analysis section 3.2.1). To evaluate the fit of cloze probability, we considered the data from Condition

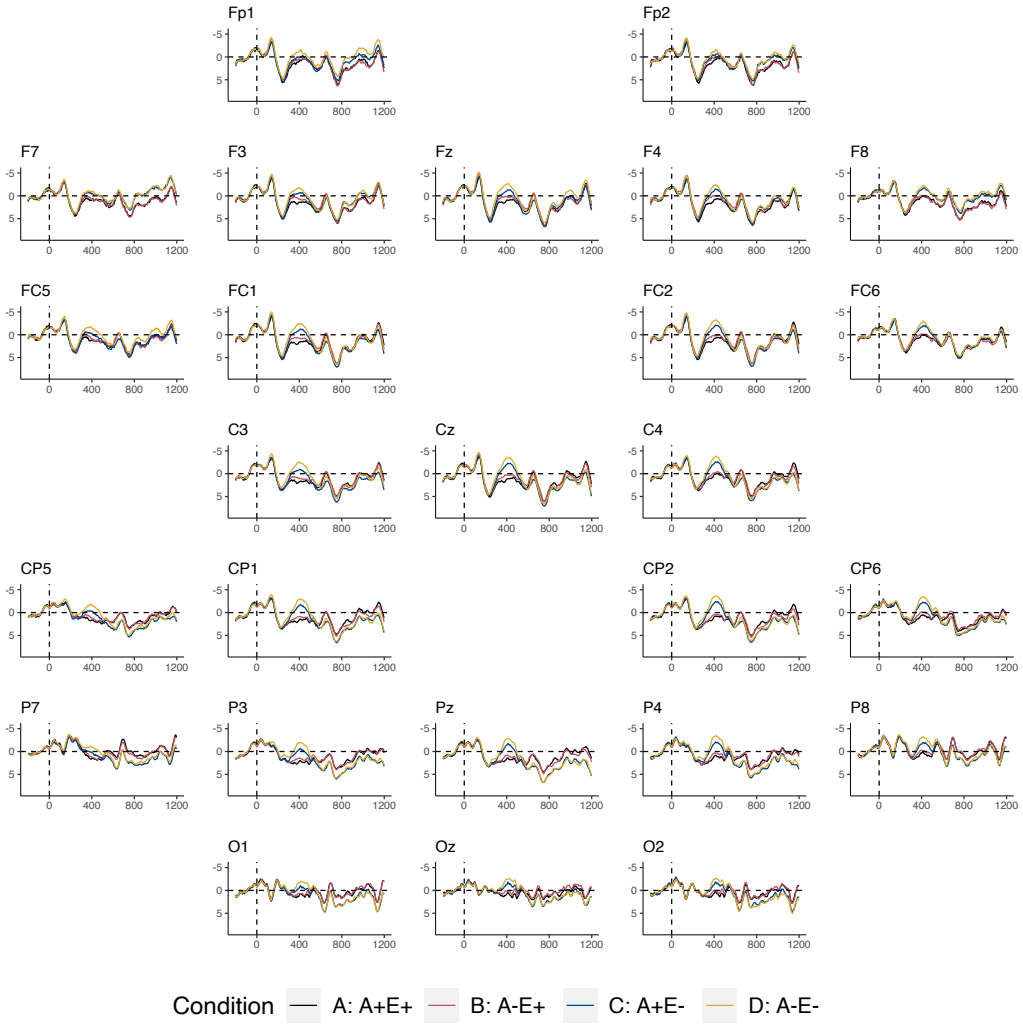


FIGURE 3.2: Grand-average ERPs in the four conditions crossing adverbial clause association and expectancy. Negative voltages are plotted upwards.

A and C. The residuals for the models including raw Cloze probability are shown in Figure 3.5 (left). Figure 3.5 (right) shows the residuals for log-transformed cloze probability (after smoothing cloze by adding 0.01 to the cloze values). We observed that log-transforming cloze probability visibly improves the fit to the data, compared to raw cloze probability.

To assess the fit of the association metrics, we considered data from conditions C and D, in which variability in cloze is minimised, as most items resulted in zero cloze probability. For these metrics, no standard non-linear transformation improved the fit compared to raw association values when inspecting the residuals visually. The residuals for the noun-target association and the verb-target association predictors are shown in Figure 3.6. Noun-target association explains most of the variability in conditions C and D, nearly predicting their averages perfectly. We observed that adding verb-target association to models already including noun-target association does not improve overall fit. We validated this finding by computing the mean of

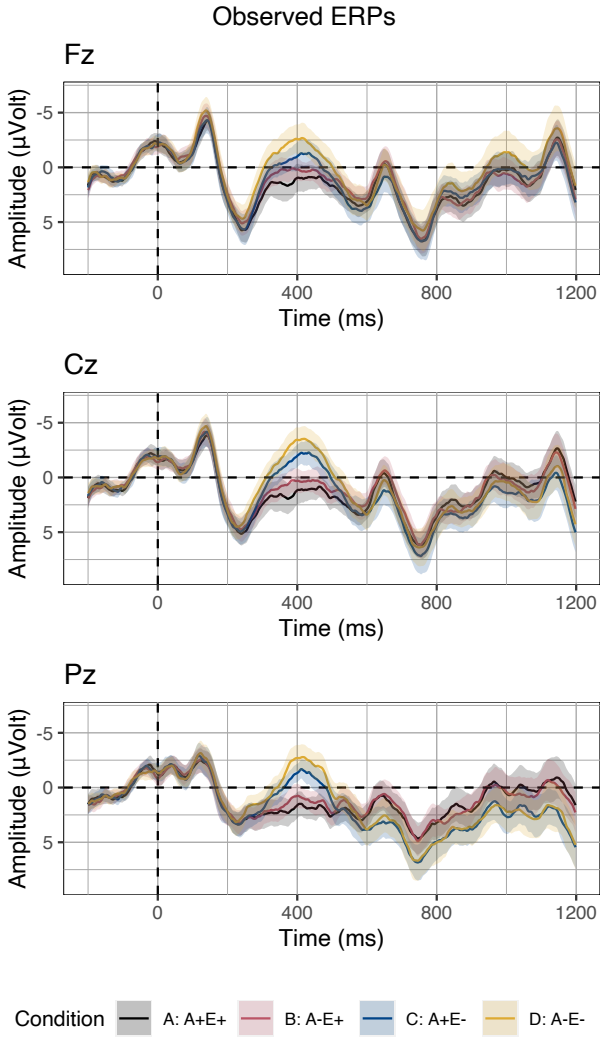


FIGURE 3.3: Grand-average ERPs on three midline electrodes in the four conditions crossing adverbial clause association and expectancy. Negative voltages are plotted upwards. Ribbons indicate standard errors computed from the per-subject, per-condition averages.

Akaike's Information Criterion (AIC) and the mean of the Bayesian Information Criterion (BIC) across models. These criteria of model quality take into account the model degrees of freedom, effectively penalising models with a larger number of predictors (including random effects). Both BIC and the less strongly penalising AIC were lower – indicating better model quality – for models including only noun-target association compared to models including both noun-target and verb-target association values (AIC: 15816 < 15826; BIC: 15866 < 15916).

Based on the results of the assessment of the individual predictors, we re-estimated the entire data set using log(cloze) probability and noun-target association as predictors for the data from all conditions. Estimated ERPs and residual error relative to the observed data are displayed in Figure 3.7. The re-estimated waveforms exhibit the same patterns as the observed data, i.e., a modulation of N400 amplitude

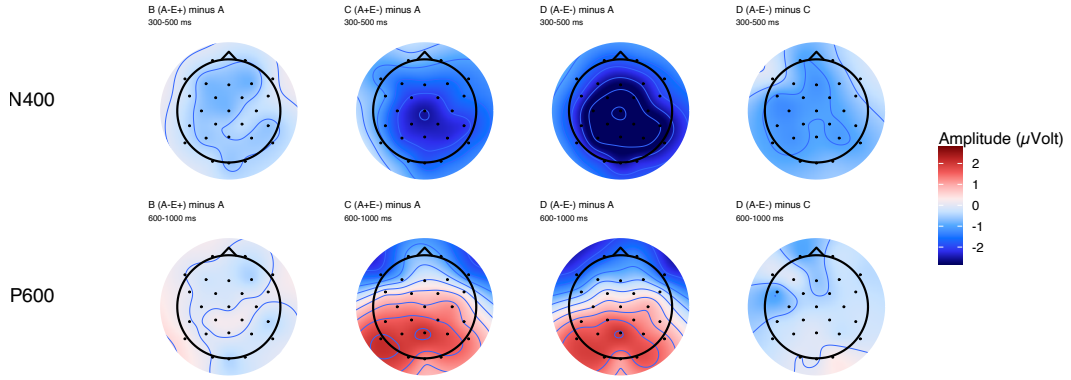


FIGURE 3.4: Topographic distributions of the average potentials in the N400 (row 1) and P600 time windows (row 2), relative to the baseline condition (columns 1-3) or relative to the unexpected-associated condition (column 4). Topographies computed from all non-reference, non-eye electrodes.

for both association and expectancy and a P600 effect in response to unexpected relative to expected targets. The residual error graph suggests that, on average, the N400 is underestimated for Condition D on electrode Pz. Furthermore, larger error is present in the very late portion of the epoch (approximately between 900 and 1200 ms). In general, however, the residuals appear low, indicating a successful approximation of the original data by the estimated data.

The coefficients from the final set of models built using log(Cloze) probability and noun-target association as predictors confirmed the aforementioned observations: Both log(Cloze) and noun-target association contribute to predicting N400 amplitude while the posterior positivity on electrode Pz is explained by log(Cloze) alone (Figure 3.8, left). Figure 3.8 (right) displays the corresponding z-values, with bars underneath the graph indicating statistically significant samples after multiple comparisons correction according to the Benjamini and Hochberg (1995) procedure. In the N400 time window, significant contributions of log(Cloze) and noun-target association were found on the three midline electrodes. The effect of noun-target association appears stronger on the frontal electrode Fz. In the P600 time window, there were significant contributions of log(Cloze) on the posterior electrode Pz, and a smaller effect on the central electrode Cz. Beyond significance, the lmerERP analysis clearly demonstrates that the predictors log(Cloze) and noun-target association can recover the observed N400-P600 complex from the observed data.

3.2.3 Discussion

In Experiment 1, we investigated the effects of lexical association and expectancy on the N400 and P600 components of the ERP signal. Specifically, we examined whether it is possible to identify a specific locus of expectancy effects, insensitive to lexical

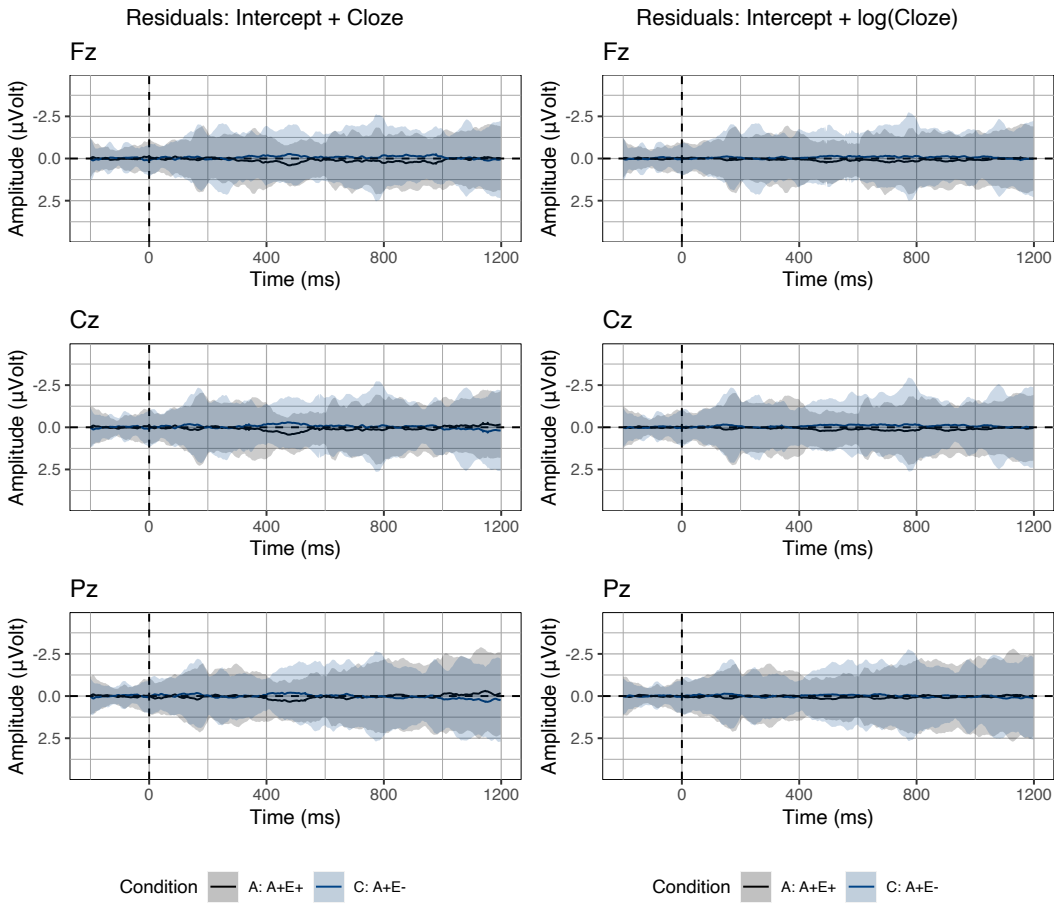


FIGURE 3.5: Residual error between observed voltages and estimated voltages in Conditions A and C using raw cloze (left) or log(Cloze) (right) as predictor. Larger deviations from zero indicate larger model error. Ribbons indicate standard errors computed from the per-subject, per-condition averages.

association. We found that while both association and expectancy contribute to modulating the amplitude of the N400, the P600 was sensitive to expectancy alone.

In the N400 time window, words that were unexpected given the selectional restrictions of the main clause verb elicited larger N400 amplitudes than more expected targets, replicating previous findings (Frank et al., 2015; Kutas & Hillyard, 1980; Kutas et al., 1984). This effect was attenuated when the critical word was semantically related to the lexical material appearing in the preceding adverbial clause, again replicating previous findings (e.g., Federmeier & Kutas, 1999; Kutas & Hillyard, 1984; Metusalem et al., 2012; Van Petten, 1993; Van Petten et al., 1997). Interestingly, the influence of association on the amplitude of the N400 was not limited to anomalous targets, but was also present for congruent ones, with larger N400 amplitude for unassociated but expected targets relative to associated and expected ones (see Frank & Willems, 2017, for the influences of expectancy and association on the N400 in naturalistic comprehension).

In the P600 time window, unexpected targets elicited a larger P600 than expected

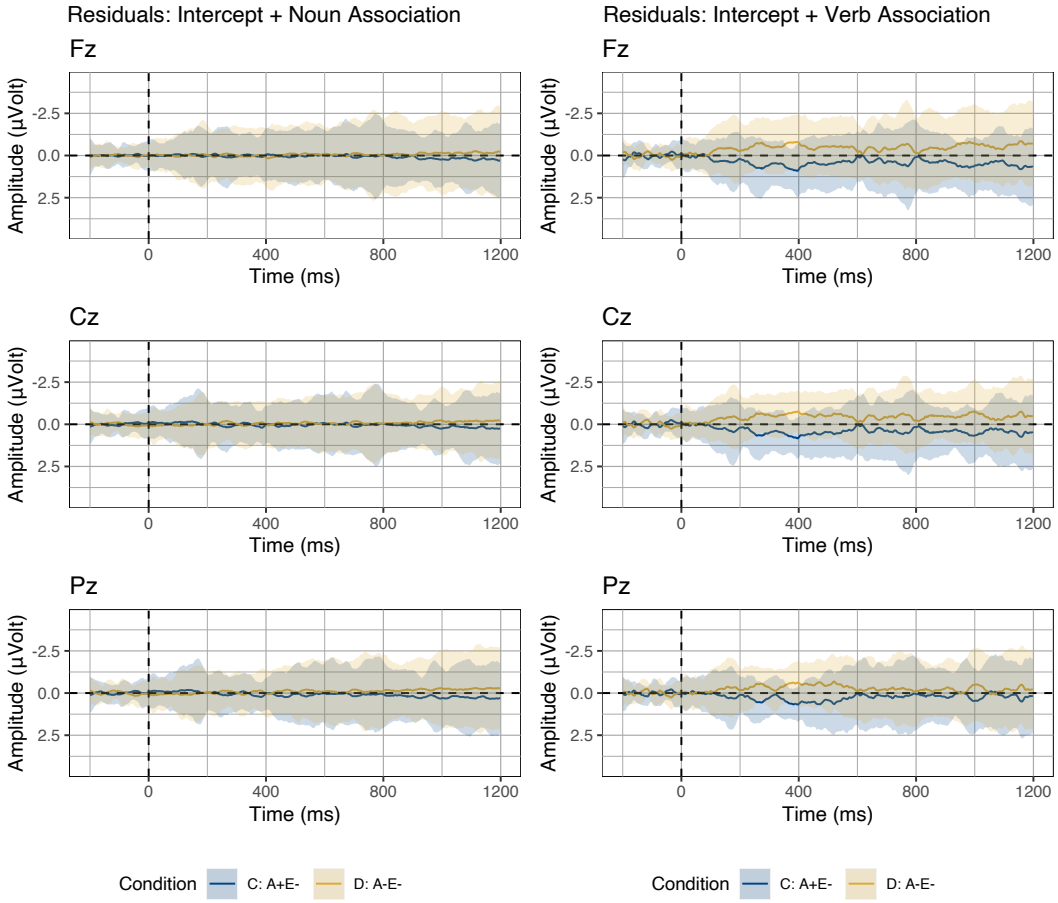


FIGURE 3.6: Residual error between observed voltages and estimated voltages in Conditions C and D using noun-target (left) or verb-target association (right) as predictor. Larger deviations from zero indicate larger model error. Ribbons indicate standard errors computed from the per-subject, per-condition averages.

targets on centro-parietal electrodes, while association had no effect. This finding is consistent with previous studies showing P600 effects elicited by semantic and world knowledge violations (e.g., Delogu et al., 2019; Hoeks et al., 2004; A. Kim & Osterhout, 2005; Nieuwland & van Berkum, 2005; Troyer & Kutas, 2020; Van Petten & Luka, 2012; van Herten et al., 2005). Since in most of those studies, as well as in ours, expectancy was manipulated via a violation of a verb's selectional restrictions, it is unclear if the observed P600 reflects expectancy or rather the detection of a semantic anomaly. To address this question, we subjected the ERP data to an additional exploratory analysis, in which lmerERPs were fitted to the EEG data recorded for Condition A only. This condition presented expected, non-anomalous targets that nonetheless exhibit variation in cloze probability (ranging from 0.17 to 1). The main goal of this analysis was to assess whether cloze probability in non-violating items predicts graded P600 amplitude on a by-item basis. This would provide evidence that the P600 is not sensitive only to categorical violations of expectancy, but rather a continuous correlate of word expectancy.

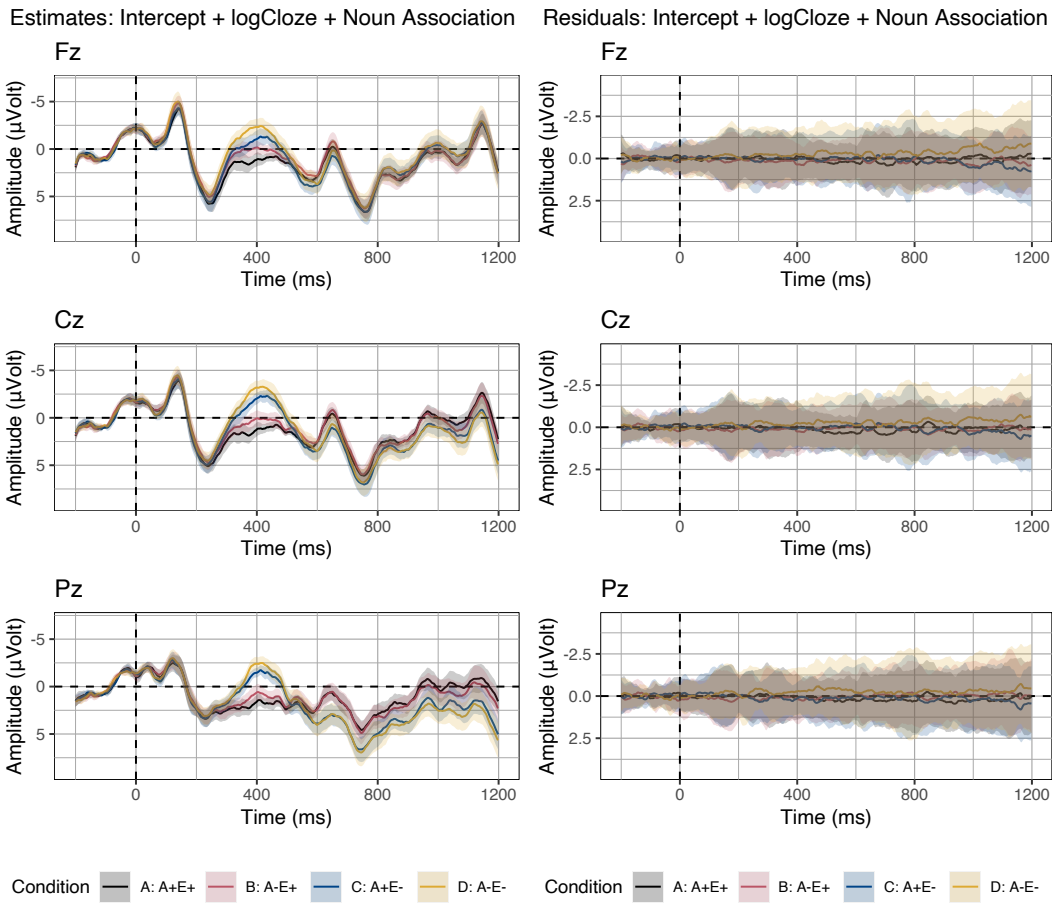


FIGURE 3.7: Estimated ERP waveforms (left) and residual error (right) computed from lmerERP models with log(Cloze) and noun-target association as predictor. Ribbons indicate standard error computed from the per-subject, per-condition averages.

As this analysis was conducted post-hoc and the stimuli were not explicitly designed to investigate graded effects of cloze probability, the results are to be interpreted with appropriate caution. Based on the aforementioned procedure, we excluded 199 out of 1200 trials (16.58%) within the baseline condition. We focus our analyses on the coefficients to assess when (in which time samples), where (on which electrodes), and to what extent (amplitude) log(cloze) probability predicts voltage deviations from the intercept. As displayed in Figure 3.9 (left), the coefficients suggest a biphasic N400-P600 modulation pattern for the baseline condition on electrode Pz. Since we used z-standardised predictors, the coefficients are mathematically equivalent to the estimated waveforms at average log(Cloze) probability (intercept) and at 1 standard deviation below average log(Cloze) probability (see also Troyer et al., 2020, for a similar approach). Accordingly, Figure 3.9 (right) displays the estimated waveforms for the entire range of log(Cloze) probabilities for Condition A, i.e., including the minimum and maximum values (cf. Table 3.2). None of the corresponding p-values reached significance in this subset of only one-fourth of the

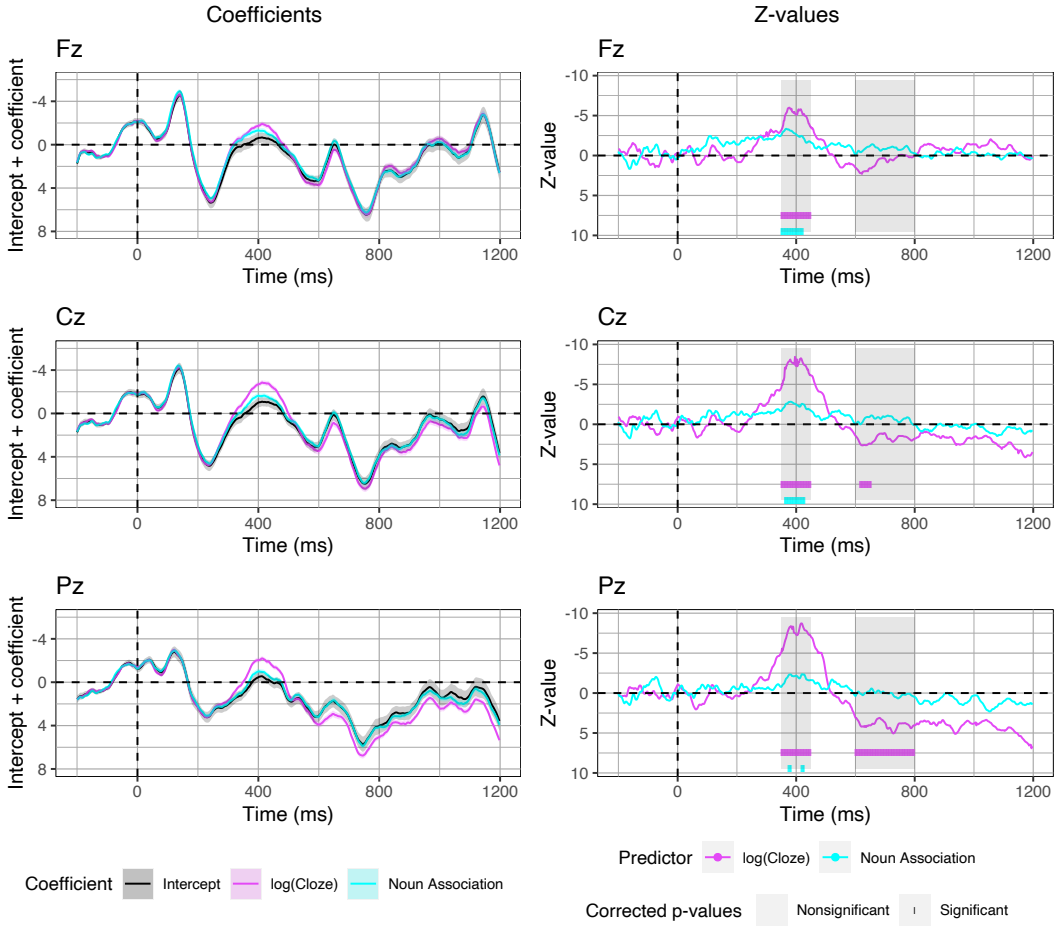


FIGURE 3.8: Coefficients (left; added to their intercept), effect sizes (z-values) from lmerERP models with log(Cloze) and noun-target association as predictors. Ribbons indicate the standard error on the coefficients from the statistical models. P-values that were significant after multiple comparisons correction are indicated by dots underneath the z-values.

original data.

3.3 Experiments 2 and 3: Self-Paced Reading

Experiment 1 provided evidence that the P600 is specifically sensitive to expectancy and insensitive to association, while both expectancy and semantic association contributed to modulation of the amplitude of the N400. In Experiments 2 and 3, we examined the relationship between these effects and behavioural processing measures. Previous work has shown that surprisal, as estimated from language models, accounts for a wide spectrum of behavioural processing phenomena, including reading times (Boston et al., 2008; Delogu et al., 2017; Demberg & Keller, 2008; Hale, 2001; Levy, 2008; Smith & Levy, 2008, 2013). These studies were, however, not explicitly designed to examine the influence of both association and expectancy on online processing. Eye-tracking studies investigating how association and plausibility interact

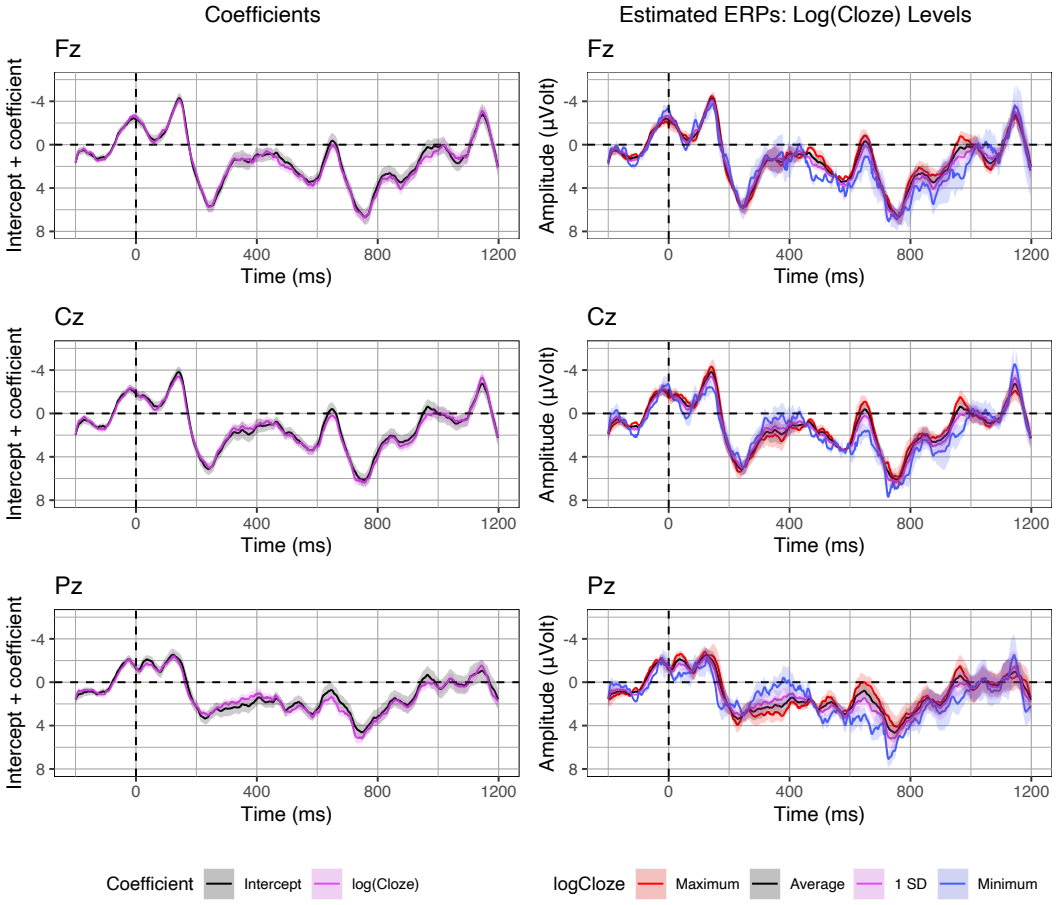


FIGURE 3.9: Coefficients (left; added to their intercept), and estimated ERPs (right) for exploratory LMER models fitted only on Condition A. Ribbons indicate the standard error on the coefficients from the statistical models (left) and standard errors computed from the per-subject, per-value averages (right).

in discourse found robust effects of plausibility, while the effect of lexical association was weaker and appeared to be modulated by the global context (Ledoux et al., 2006). For example, Camblin et al., 2007 showed robust effects of plausibility on eye movements, while lexical association had a smaller and more localised effect, and only on incongruent words. Similar results were found by Brouwer, Delogu, Venhuizen, and Crocker (2021) in a self-paced reading study showing a significant effect of plausibility, but not of association. Thus, it is not clear to what extent behavioural measures may capture the N400 effects of association that we observed in Experiment 1, beyond the effects of expectancy. Moreover, Frank (2017) has argued that any effect of semantic relatedness on reading times may be due to a confound with word predictability. Therefore, we conducted self-paced reading experiments using exactly the same stimuli as those used in Experiment 1. Reading times were then analysed using a similar regression-based estimation approach to assess if, how, and when expectancy and association contribute to modulations of behavioural processing indices.

3.3.1 Method

We presented the materials of Design 1 in two self-paced reading studies. We conducted the two experiments as web-based studies because a pandemic prohibited in-lab experiments.[2] The two experiments differed only in their task. In Experiment 2, we used the same plausibility judgement task that we applied in the ERP experiment, whereas in Experiment 3, we asked comprehension questions.

Participants

Participants were recruited through Prolific Academic Ltd. After the exclusion of individual participants due to inattentive reading (shown by short completion times and low accuracy), 48 participants were kept for each experiment. The remaining 48 participants were all native speakers of German and had not indicated any language-related disorders (such as reading difficulties). Demographics of the participants in the two studies were similar (Experiment 2: mean age 24.69; SD 4.47; age range 18-32; 17 female, 31 male; Experiment 3: mean age 24.21; SD 4.30; age range 18-32; 24 female, 24 male). All participants gave informed consent by agreeing to the written study conditions and were paid £6.25 for their participation.

Materials

The materials were exactly the same as those used in Experiment 1, following Design 1 (Table 3.1).

Procedure

The experiments were implemented using the software Ibex/PennController (Drummond, 2012; Zehr & Schwarz, 2018). On each trial, participants were prompted to press the Enter key to start reading, after which they were presented with a hash sign at the centre of the screen indicating the position of the words. From then on, each word was presented centrally and participants had to press the Space bar to proceed to the next word.

In Experiment 2, the task was the same as in the ERP experiment, i.e., participants judged the plausibility of the sentence using *Yes*/*No* on each trial. The answers were mapped to the D and K keys and key assignment changed randomly. In Experiment 3, participants were tasked with answering comprehension questions using *Yes* or *No*, again mapped to the D and K keys. There was a question on one third of the trials (experimental or filler) and the question could be about the content of any part of the experimental sentences to incentivise attentive reading of the entire sentence. Compared to a plausibility judgement task, we deem this a better-suited task for the web-based environment in which the experimenter can exert less control over

[2] In a lab-to-web replication (Keller et al., 2009), self-paced reading resulted in comparable reading time measures.

the environment and behaviour of the participant, as task engagement with a comprehension question should be larger than with binary plausibility judgements. For the comprehension questions, the *Yes*/*No* position did not vary randomly but was reversed for half of the participants. In both experiments, we recorded participants' response accuracies and reaction times to the task. After completion of ten practice trials, the materials were presented in three blocks of 80 trials each, half of which were fillers. In the version with comprehension questions (Experiment 3), we also provided coarse feedback on participants' response accuracy after the practice and after each block, in order to motivate attentive reading. Participants were encouraged to take a short break between blocks. Due to technical limitations, the self-paced reading experiments differed from the EEG experiment in that words were presented in black font on white background and not vice versa.

Analysis

The analysis of the reading time data was conducted similarly to that of the ERP data: Cloze probabilities and association ratings were used as numerical predictors in linear mixed effects models which were then used to re-estimate the observed data. We analysed reading times on the word preceding the target (the Pre-critical region), on the target word (the Critical region), and, to capture spillover effects, on the two words following it (the Spillover and Post-spillover region). The Spillover region always consisted of a closed class word (most commonly "und"/"and"), while the Post-spillover region consisted of both closed and open class words. We considered each region as pertaining to a separate family of hypotheses and, hence, we did not correct for multiple comparisons across regions. Reading times were log-transformed to normalise their right-skewed distribution. In the data of Experiment 3, the (Shapiro & Francia, 1972) test for normality, adequate for larger sample sizes, was however still significant on each region, suggesting non-normality.

3.3.2 Results

Experiment 2

We excluded data from all regions if reading time on any of the four regions was lower than 50 ms or higher than 2500 ms, or if response time on the task was lower than 50 ms or higher than 6000 ms. Based on these criteria, we excluded 83 out of 5760 trials (1.44%).

Plausibility Judgement Participants judged the sentences as expected based on the design, i.e., assuming *Yes* as correct answer for expected and *No* as correct answer for unexpected sentences. Average accuracy was 86.4 % (SD = 8.7%, range = 63.8 - 96.6%) with an average reaction time of 1218 ms (SD = 378 ms, range = 612 - 2497 ms;

	Accuracy			Reaction Time		
Cond.	Mean	SD	Range	Mean	SD	Range
A	91.0%	6.5%	73.3% - 100.0%	1164 ms	328 ms	604 ms - 1870 ms
B	86.2%	10.5%	56.7% - 100.0%	1284 ms	426 ms	645 ms - 2407 ms
C	81.7%	15.4%	46.7% - 100.0%	1203 ms	461 ms	584 ms - 3302 ms
D	86.9%	10.9%	57.1% - 100.0%	1222 ms	457 ms	608 ms - 3209 ms

TABLE 3.5: Task performance on the binary plausibility ratings in the first self-paced reading experiment. Accuracy and reaction times were computed across subjects.

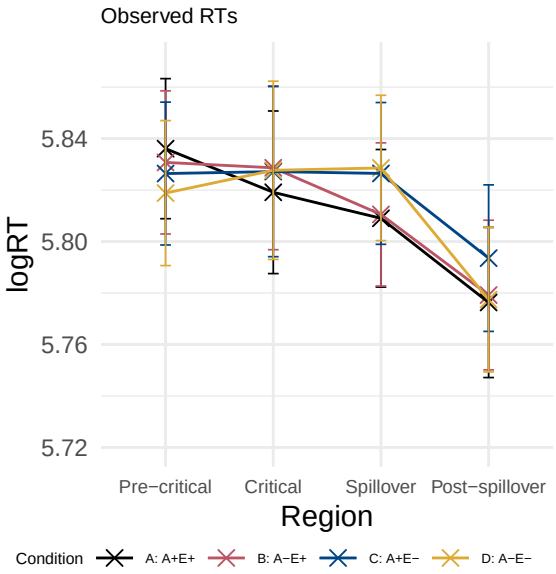


FIGURE 3.10: Log reading times per condition on the Pre-critical, Critical, Spillover, and Post-spillover regions. Error bars indicate standard errors computed from the per-subject, per-condition averages.

both metrics computed across subjects). Means, standard deviations, and ranges in the four conditions are reported in Table 3.5.

Reading Times Average reading times in the four conditions on the Pre-critical region (the article of the target word), the Critical region (the target word, axe), the Spillover region, and the Post-spillover region are displayed in Figure 3.10. Reading times at the Critical region were slowed for all manipulated conditions (B, C, and D). On the Spillover region, only expectancy had an effect, with slower reading times in the unexpected conditions (C and D) relative to the expected conditions (A and B). On the Post-spillover region, reading was slowed only in the unexpected-associated Condition C, whereas the unexpected-unassociated Condition D did not differ from baseline reading time.

Analogous to the analysis of the ERPs, we modelled log-transformed reading times as a linear function of noun association and log(Cloze). We calculated the

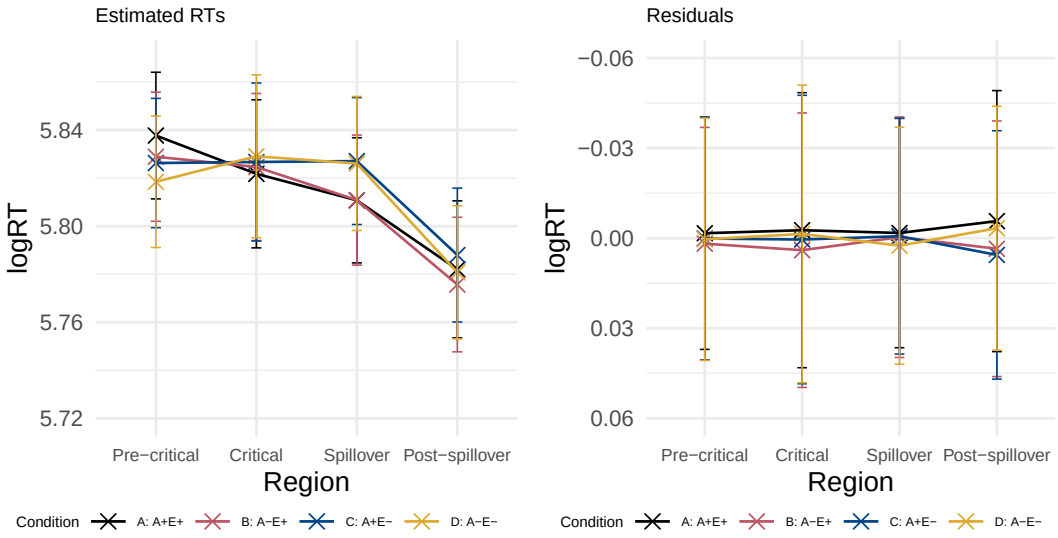


FIGURE 3.11: Estimated log reading times (left) and residual error (right), split per condition, on the Pre-critical, Critical, Spillover, and Post-spillover regions. Error bars indicate standard errors computed from the per-subject, per-condition averages.

forward estimates of the fitted models and provide estimated reading time data and the residual error, i.e., the difference between estimated and observed data, in Figure 3.11. Overall, the estimated data bear the same qualitative patterns as the observed data, with the exception of the Post-Spillover region. Because the models are not provided with an interaction term between association and log(Cloze), they are unable to arrive at a solution in which only Condition C is slowed down. However, since we did not have hypotheses about possible interactions, we decided not to include this interaction term ex-post.

The coefficients and effect sizes of the models (Figure 3.12) suggest minor and insignificant contributions of association and expectancy on the Critical region. On the Spillover region, the effect of expectancy is largest and becomes significant. No significant contributions are observed on the Post-spillover region, but we again note the limitation of these models that lack an interaction term.

Experiment 3

For the self-paced reading experiment with comprehension questions, we excluded data from all regions if any reading time on any of the four regions was lower than 50 ms or higher than 2500 ms, or if response time on the task – if there was one on this trial – was lower than 50 ms or higher than 6000 ms. Based on these criteria, we excluded 73 out of 5760 trials (1.27%).

Comprehension Questions Average accuracy on the comprehension questions was 87.4 % (SD = 6.9%, range = 69.0 - 97.6%) with an average reaction time of 2461

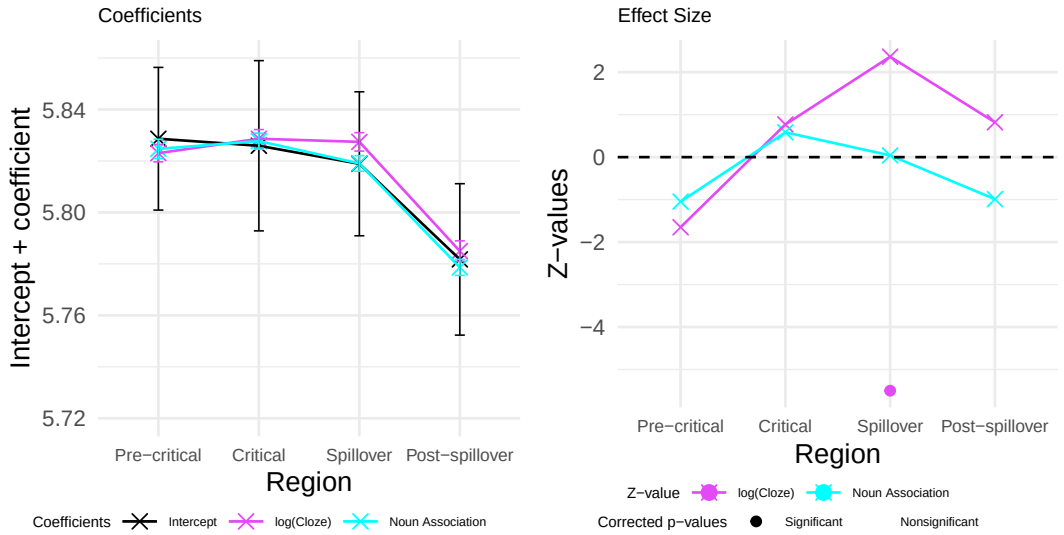


FIGURE 3.12: Coefficients (left; added to their intercept), effect sizes (z-values) and p-values (right) from lmerSPR models with log(Cloze) and noun-target association as predictors. Error bars indicate the standard error on the coefficients from the statistical models.

	Accuracy			**Reaction Time**		
Cond.	Mean	SD	Range	Mean	SD	Range
A	90.7%	8.6%	70.0% - 100.0%	2379 ms	529 ms	1298 ms - 3579 ms
B	88.5%	11.1%	63.6% - 100.0%	2402 ms	540 ms	1494 ms - 3657 ms
C	85.5%	12.6%	45.5% - 100.0%	2522 ms	521 ms	1441 ms - 3776 ms
D	85.6%	12.9%	54.5% - 100.0%	2534 ms	542 ms	1466 ms - 3628 ms

TABLE 3.6: Task performance on the binary plausibility ratings in the second self-paced reading experiment. Accuracy and reaction times were computed across subjects.

ms (SD = 489 ms, range = 1503 - 3485 ms; both metrics computed across subjects). Means, standard deviations, and ranges in the four conditions are reported in Table 3.6.

Reading Times Figure 3.13 displays average reading times in the four conditions on the Pre-critical, Critical, Spillover, and Post-spillover regions. Reading times on the Critical region were slowed for conditions B, C, and D. On the Spillover region, association and expectancy had an additive effect, with slower reading times in the weakly associated (B and D) and unexpected (C and D) conditions relative to the baseline condition (A). Lastly, on the Post-spillover region, association effects were no longer observed, and only the unexpected conditions were read slower.

Since the reading times obtained from the experiment with comprehension questions visually appeared to exhibit more robust effects than those obtained using the plausibility judgement task, we subjected the data from Experiment 3 to a more in-depth analysis. We first assessed whether transformations applied to the predictors

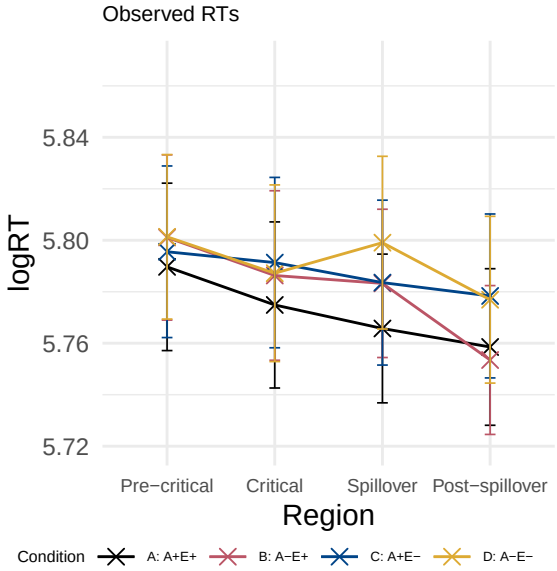


FIGURE 3.13: Log reading times, split per condition, on the Pre-critical, Critical, Spillover, and Post-spillover regions. Error bars indicate standard errors computed from the per-subject, per-condition averages.

led to improvements on the model residuals, but did not observe any large differences between log-transformed and untransformed cloze probability. With respect to the association ratings, we found that verb-target association did not account for log-transformed reading times over and above noun-target association (as was the case with the ERPs). In line with these findings, and in order to maximise comparability with the previous ERP and SPR (self-paced reading) results, we modelled log-transformed reading times as a linear function of log(Cloze) probability and noun-target association.

The estimated reading times adequately modelled the observed effect structure on the Pre-critical, Spillover, and Post-spillover regions (as shown in Figure 3.14, left). This is not the case in the Critical region, where Condition D is overestimated and Condition B is underestimated. Again, without an interaction term, the models are unable to arrive at a solution in which the estimated reading times of Condition B can be slowed without increasing the reading times of Condition D as well. Hence, residual error for these two conditions is larger in the Critical region (Figure 3.14, right). Again, we decided not to include an interaction term in our models, as we did not have hypotheses about possible interactions.

The model coefficients and effect sizes in Figure 3.15 confirm the visual inspection of the reading times in each condition, as laid out above. There are no large contributions of the predictors in the Pre-critical region. Indeed, log(Cloze) alone accounts for increased reading times on the Critical region (but note the limitation due to the lacking interaction term), whereas, on the Spillover region, log(Cloze) and noun-target association have an additive effect. On the Post-spillover region log(Cloze) alone predicts reading times departing from the intercept.

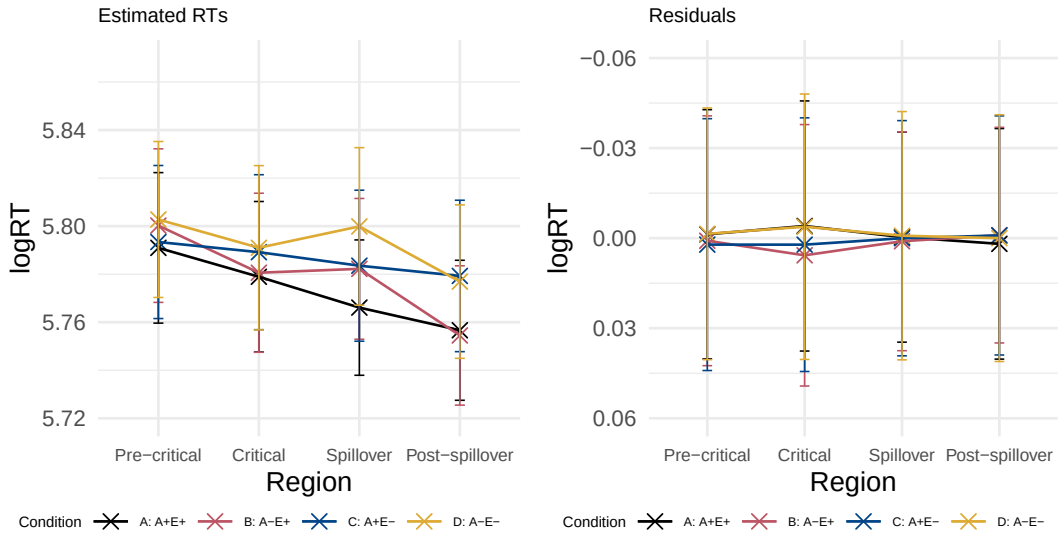


FIGURE 3.14: Estimated log-transformed reading times (left) and residual error (right), split per condition, on the Pre-critical, Critical, Spillover, and Post-spillover regions. Error bars indicate standard errors computed from the per-subject, per-condition averages.

As Experiment 3 resulted in a clearer expectancy effect than Experiment 2, we also applied the same exploratory analysis approach that was used in Experiment 1 – where we assessed the graded effects of expectancy on the N400 and the P600 (see Section 3.2.2) – to the data from Experiment 3. That is, we considered reading times from trials in the baseline condition only. Based on the previously mentioned criteria, we excluded 15 out of 1440 trials (1.04%). The results of this analysis are shown in Figure 3.16. Similarly to what we observed for the N400 and the P600, log(Cloze) probability appears to have a graded effect on reading times, with increased reading times for lower cloze probability trials.

3.3.3 Discussion

In the self-paced reading studies, we recorded reading times on exactly the same stimuli as in the ERP experiment in order to assess the relationship of association and expectancy in the behavioural domain and compare these results to the electrophysiological domain. The two self-paced reading studies differed only in their task: In the first version (Experiment 2), participants judged the plausibility of each sentence, and in the second version (Experiment 3), they replied to binary comprehension questions on approximately one third of trials. In Experiment 2, we observed slowed reading times on the Critical region for all manipulated conditions. On the Spillover region, the unexpected conditions were read slower than the expected ones. On the Post-Spillover region, however, Condition D (A-E-) was read as fast as the baseline condition, whereas Condition C (A+E-) was still slowed relative to baseline. In Experiment 3, reading times were also slowed already on the critical

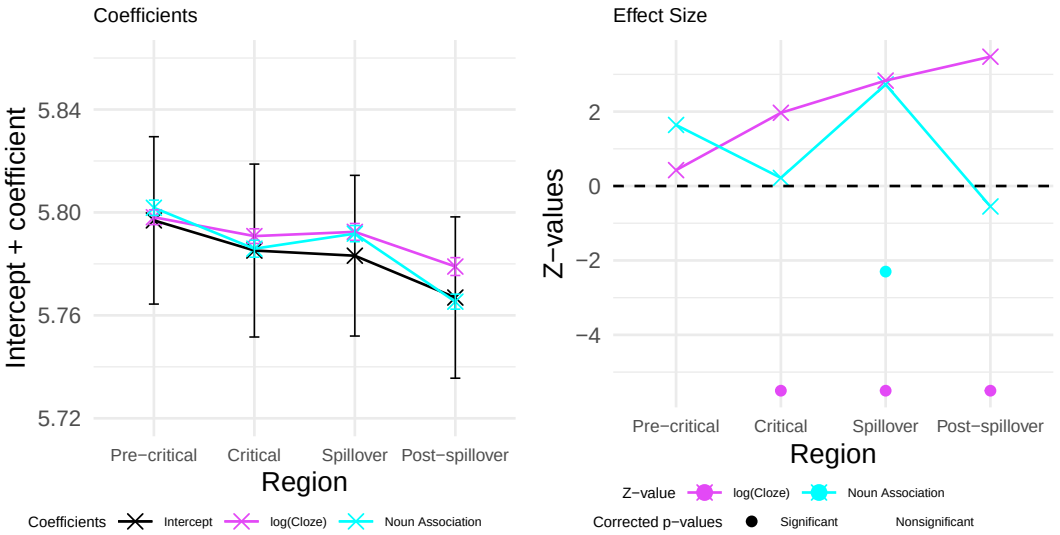


FIGURE 3.15: Coefficients (left; added to their intercept), effect sizes (z-values) and corrected p-values (right) from lmerSPR models with log(Cloze) and noun-target association as predictors. Error bars indicate the standard error on the coefficients from the statistical model.

region, for all manipulated conditions relative to the baseline. Different from Experiment 2 where only expectancy appeared to influence reading time, we observed that both expectancy and association influenced reading times on the Spillover region, and that both unexpected conditions were slowed on the Post-spillover region. The difference between the effect structure in the two experiments could be explained in terms of task demands: Whereas the plausibility rating puts the expected-unexpected contrast into focus, the comprehension questions also queried about the intervening adverbial clause, hence making the association manipulation relevant for the readers which may have resulted in slowed target word reading in the unassociated contexts on the Spillover region. The task environment may also lead readers to speed up their reading in Condition D (A-E-) of Experiment 2: In this condition, readers encounter both unassociated intervening material and an unexpected main verb-target word combination. This could lead readers to decide about their plausibility rating already early and result in a speed up in reading in order to proceed to the plausibility rating, whereas a similar strategy would not help to answer comprehension questions. A significant contribution of association was observed only in the experiment using comprehension questions, but not in the version with the plausibility judgement task. Potentially, the strength of association effects in reading time data could thus be task dependent and be less pronounced when the associated/unassociated material is not relevant for the task, as was the case in our plausibility judgement task, where implausibilities were induced by the main verb-target word relationship. Thus, the results of the two self-paced reading studies in combination are consistent with previous eye-tracking findings showing robust effects of plausibility but short-lived effects of association (e.g., Camblin et al., 2007).

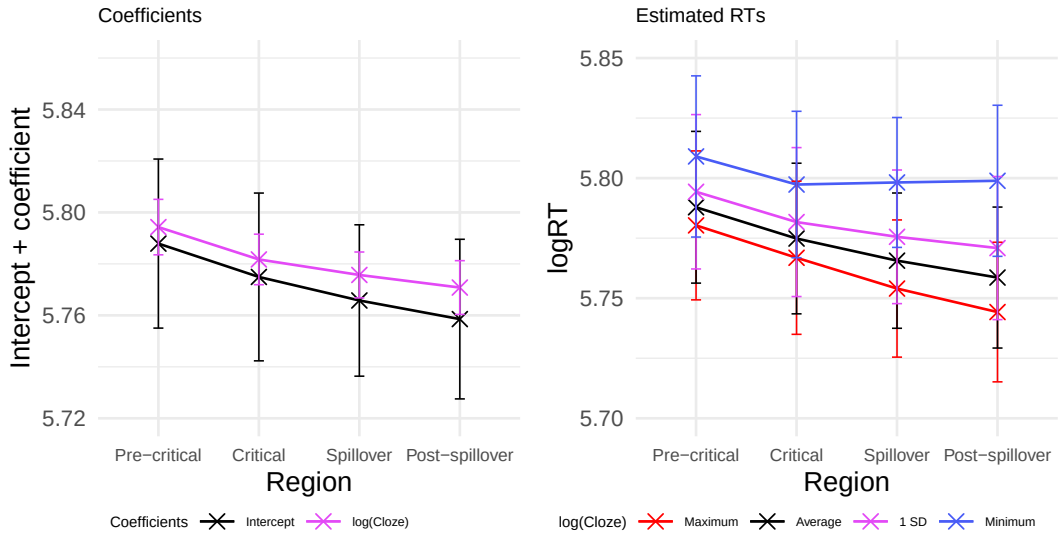


FIGURE 3.16: Coefficients (left; added to their intercept) and estimated log-RTs (right) for exploratory LMER models fitted only on Condition A. Error bars indicate standard errors on the coefficients from the statistical models (left) and standard errors computed from the per-subject, per-value averages (right).

Interestingly, the temporal distribution of the effects in Experiment 3 seemed to align with the ERP patterns observed in Experiment 1. Both association and expectancy had an impact on earlier processing stages, the N400 time window in Experiment 1 and the Spillover region in Experiment 3. Expectancy alone had a later effect, corresponding to the P600 time window in Experiment 1 and the Post-spillover region in Experiment 3. We return to possible interpretations of this pattern in the General Discussion (Section 3.4). Lastly, we also conducted a post-hoc analysis investigating expectancy-related reading time modulations in the baseline condition of Experiment 3, which closely matched the effect structure of the ERP experiment. Here, we replicated the graded effect of expectancy in the non-violating trials of the baseline condition that we observed in the ERP experiment. These findings provide evidence that the processing effort observed in the present experiments does not merely index the detection of an anomaly, but rather reflects the degree to which a word – whether anomalous or not – is expected given the prior context.

3.4 General Discussion

We conducted three experiments aimed at disentangling the effects of expectancy and lexical association on electrophysiological (Experiment 1) and reading time (Experiment 2 & 3) measures of online processing and examined if it is possible to identify a specific locus of expectancy effects in the ERP signal. The experimental design tested sentences in which a direct object noun was either expected or unexpected,

given the selectional restrictions of the main verb (as measured through cloze probability). Furthermore, the target was either strongly or weakly associated with the content words of an adverbial clause preceding the target (as measured through lexical association norms). Critically, this adverbial clause was completely independent of the expectancy manipulation, avoiding any dependence between these often confounded factors. In sum, our design crossed the factors expectancy and association using a context manipulation.

The results of Experiment 1 revealed that the N400 component is sensitive to both expectancy and lexical association. Unexpected targets elicited a larger N400 amplitude than expected targets, and this effect was modulated by lexical association, with highly associated targets eliciting lower N400 amplitude than weakly associated ones. The P600, on the other hand, was sensitive to expectancy alone, with unexpected targets eliciting a larger P600 than expected ones. Experiment 2 replicated the ERP study as a web-based self-paced reading experiment, using the same plausibility task. The results revealed slowed reading of unexpected words on one Spillover region and no association-related reading time modulations. Experiment 3 repeated the self-paced reading experiment, but replaced the plausibility judgement task with comprehension questions, which were deemed a better task for the web-based environment of the reading time studies. This experiment demonstrated that while both expectancy and lexical association significantly influenced reading times soon after the critical word, only expectancy had an effect downstream. Furthermore, the exploratory analyses conducted for both Experiment 1 and Experiment 3 provided preliminary evidence that the effect of expectancy on the P600 is graded and does not depend on the presence of a semantic violation. In what follows, we discuss the main findings and their implications for neurocognitive accounts of language comprehension and the notion of surprisal.

3.4.1 The N400 is Sensitive to Both Expectancy and Lexical Association

Both expectancy and lexical association contribute to predicting the amplitude of the N400. This finding is consistent with a substantial body of evidence showing N400 effects of cloze probability (Kutas et al., 1984), word surprisal (Delogu et al., 2017; Frank et al., 2015), and semantic similarity (Federmeier & Kutas, 1999; Frank & Willems, 2017). It is also consistent with several studies showing that N400 effects elicited by semantic violations or implausibility are attenuated or even overridden when the eliciting word is semantically related to the context (e.g., Delogu et al., 2019; Kutas & Hillyard, 1984; Metusalem et al., 2012; Nieuwland & van Berkum, 2005). Interestingly, our design allowed us to establish that lexical association modulates the ERP signal also within semantically congruent items, as evidenced by the small N400 effect elicited by unrelated but expected targets relative to their related and expected counterparts. Taken together, these findings indicate that the N400 is sensitive to lexical association over and above expectancy. An important question is

therefore to what extent the two effects hinge upon the same underlying cognitive mechanism as opposed to being qualitatively different.

We argue that the additive influences of these two properties can be naturally and parsimoniously accommodated within the memory-retrieval view of the N400 (Brouwer et al., 2012; Kutas & Federmeier, 2000, 2011; Lau et al., 2009; Lau et al., 2008; van Berkum, 2009, 2010). On this view, the amplitude of the N400 reflects the ease with which the meaning of a word is accessed in long-term memory. We define lexical access or retrieval as the cognitive process that maps perceived word forms onto their corresponding word meaning, taking context into account. This process is facilitated when this meaning is *associated* with conceptual knowledge activated by previous words in the context and/or when it can be *expected* given the unfolding utterance interpretation (among other factors; see Brouwer et al., 2017, for discussion). As a consequence, the retrieval account offers a parsimonious explanation of why both factors influence the N400, without needing to resort to a hybrid view on which expectancy effects would be explained in terms of integration effort.

3.4.2 The P600 is Sensitive to Expectancy Alone

In the P600 time window, we found that expectancy alone accounts for the positivity observed on centro-parietal sites. This effect can neither be explained in terms of syntactic processing difficulty, as our stimuli were syntactically well-formed and unambiguous, nor merely as a response to semantic violations (see Van Petten and Luka, 2012), as an exploratory analysis performed on a subset of data varying in cloze probability suggested a continuous sensitivity of the P600 to expectancy in congruent trials. This result is thus consistent with a growing body of evidence indicating that the P600 is a general index of integration difficulty at different levels of analysis (e.g., Burkhardt, 2006, 2007; Delogu et al., 2019, 2021; Delogu et al., 2018; Hoeks et al., 2013; Regel et al., 2010; Spotorno et al., 2013; see Brouwer et al., 2012, for discussion). We define integration as the cognitive process that maps retrieved word meanings into the utterance meaning representation of the sentence so far, taking context into account. Under this interpretation, the effort involved in updating the unfolding utterance meaning is greater the more unexpected the utterance meaning resulting from integrating the meaning of the incoming word.

Moreover, our data provide initial evidence that a negative correlation may exist between cloze and the amplitude of the P600, similar to to the established negative correlation between N400 amplitude and a word's cloze probability. The gradedness of this link between the P600 and integration effort should be corroborated further in a dedicated experimental study. Critically, however, studies aimed at assessing this relationship should control for spatiotemporal component overlap with the graded N400, resulting from retrieval see Brouwer and Crocker, 2017, for discussion. That is, in order to obtain a clear view of the gradedness of the P600, overlap with the graded N400 should be factored out. Experimentally, this can effectively be achieved

by strongly priming the target word while still varying its plausibility (Delogu et al., 2021; Nieuwland & van Berkum, 2005). We apply such an experimental design in the following Chapter 4. Overall, the present findings provide compelling evidence that the P600 is a specific locus of expectancy effects, and not sensitive to lexical association, consistent with the Retrieval-Integration account (Brouwer et al., 2017; Brouwer et al., 2012).

3.4.3 An Integrated Theory of the N400 and the P600

The functional interpretation of the N400 and P600 has been subject to debate for a long time. Based on the attenuation in N400 amplitude that we observed in response to associated adverbial clauses, we exclude the "pure" integration view of the N400 (Brown & Hagoort, 1993, 2000; Hagoort et al., 2004), which would predict an effect of expectancy alone. Similarly, it is our understanding that the computational model put forward by Rabovsky et al. (2018), while capturing the expectancy effects, would not predict the association effect arising from the preceding adverbial clauses. These clauses were constructed so as to rule out any structural, or even semantically attractive (thematic) dependency with the target word, which is typically prerequisite for "good-enough" processing effects (Ferreira & Patson, 2007; Rabovsky & McClelland, 2020). The "hybrid" view of the N400 (Baggio, 2018; Baggio & Hagoort, 2011; Lau et al., 2016; Nieuwland et al., 2020), however, can explain the observed N400 findings by assuming that both retrieval and integration processes are indexed by the N400. Nonetheless, the results are completely aligned with a pure retrieval view of the N400 (Brouwer et al., 2012; Kutas & Federmeier, 2000, 2011; Lau et al., 2009; Lau et al., 2008; van Berkum, 2009, 2010) as well, under which both association and expectancy facilitate word retrieval.

The P600 in our data resulted from a violation of the main verb's selectional restriction on its object, i.e., the target word. The resulting items were, however, syntactically well-formed, ruling out the view that the P600 serves as an index of morpho-syntactical processing (Friederici, 1995; Hagoort et al., 1999; Osterhout & Holcomb, 1992) or syntactic integration (Kaan et al., 2000; Kaan & Swaab, 2003) alone. While conflict monitoring/resolution theories (Bornkessel-Schlesewsky & Schlesewsky, 2008; A. Kim & Osterhout, 2005; Kos et al., 2010; Kuperberg, 2007; van Herten et al., 2005) could predict a P600 in response to the selectional restriction violation, such accounts generally have difficulty explaining biphasic N400-P600 patterns that are also present in our data (see Brouwer et al., 2012, for discussion). Lastly, the integration view of the P600 (Brouwer et al., 2017; Brouwer et al., 2012) is completely in line with our results. Further, only the integration view would predict a graded sensitivity of the P600 to expectancy, as suggested by our post-hoc analysis.

In contrast to the Retrieval-Integration account, which combines the retrieval view of the N400 and the integration view of the P600, other theories typically focus

on either the N400 or the P600 and therefore offer no account for their interdependence. Two notable exceptions, however, are the recent computational models by Rabovsky et al. (2018) and Fitz and Chang (2019). While the former offers a computational instantiation of the N400 as integration, Rabovsky and McClelland (2020) verbally theorise that the P600 may reflect an attention-dependent revision process that can re-assess wrong interpretations generated by an automatic interpretation process indexed by the N400. Our design specifically avoids creating a semantic illusion that can be resolved by revision (see the Materials in Section 3.2.1) and, as such, the violation of expectancy should be reflected only in the N400 and not in the P600. Further, it is unclear how such attention-dependent revision processes would explain the graded P600 response to word expectancy suggested by our data. The model proposed by Fitz and Chang (2019) successfully captures data from several ERP studies and characterises the N400 and the P600 as epiphenomena of error-based learning. This model accounts for the expectancy effect on the N400 as well as on the P600, with the latter being interpreted as a result of the selectional restriction violation. It is unclear, however, whether the model would predict the association effect from the adverbial clauses on the target word. Moreover, while this model predicts a graded link of the N400 to cloze probability, presumably, their model would also not predict a graded link of expectancy to the P600. Further enquiry into this latter point could thus provide a strong test to dissociate between the RI model and those of Fitz and Chang (2019) and Rabovsky and McClelland (2020).

In sum, one would have to invoke several theories, explaining both ERP components individually in order to account for the entire ERP complex in our data. A key strength of the Retrieval-Integration account, on the other hand, is that it explains the entire ERP complex within one integrated theory, making predictions for both components: The N400, as an index of lexical retrieval, is sensitive to both lexical association and expectancy, whereas the P600, as an index of integration, is sensitive to expectancy only.

3.4.4 Dissociating Retrieval and Integration in Behavioural Measures

The results of Experiments 2 and 3 replicated the well-known effect of expectancy (or surprisal) on reading times (e.g., Boston et al., 2008; Brouwer, Delogu, Venhuizen, et al., 2021; Delogu et al., 2017; Demberg & Keller, 2008; Frank, 2009; Smith & Levy, 2008, 2013). An interesting question, however, is to what extent this behavioural cost reflects retrieval or integration processes or both, as self-paced reading time is presumably the summation of the effort involved in several underlying processes. In Experiment 3, which used comprehension questions to increase participant engagement, we observed that association and expectancy significantly predicted reading times on the Spillover region, while the influence of expectancy remained up until the Post-spillover region, suggesting that expectancy influences both early (retrieval)

and later (integration) processes, similarly to what we observed with ERPs. An interesting open question is therefore how reading time effects in the time domain relate to ERP effects in the amplitude domain. The temporal dynamics of association and expectancy effects in reading times appear to echo the temporal pattern of the corresponding modulations in ERP components, with the N400 *effect* of association and expectancy preceding the P600 *effect* of expectancy alone (although the actual processes underlying the respective components do temporally overlap; see Delogu et al., 2021). We could therefore hypothesise that the reading time increases in the Spillover region capture facilitation related to memory-retrieval for associated words, which also modulates the amplitude of the N400, while the cost in the Post-spillover region reflects more demanding integrative processing, which in the electrophysiological domain is associated with increased P600 amplitude. Clearly, this is only speculative and would need to be examined in studies designed for this purpose. An experimental paradigm well-suited to address this issue could be one in which ERPs and self-paced reading times are recorded simultaneously (see Bulkes et al., 2020; Ditman et al., 2007; Payne & Federmeier, 2017, for examples). Further, it remains to be seen what reading time signature is elicited by an experimental design that elicits only P600 modulations in response to differential integration effort (see Chapter 4).

3.4.5 The P600 is an Index of Comprehension-Centric Surprisal

All contemporary models of language comprehension acknowledge the important role of expectancy in determining word processing difficulty. Among them, surprisal theory (Hale, 2001; Levy, 2008) posits that the cognitive effort incurred by each word in a sentence is proportional to its surprisal, defined as the negative log-probability of a word given the prior context. Surprisal has been estimated using various language models (i.e., n-gram models, phrase-structure grammars, and recurrent neural networks), and has been shown to correlate with reading time (Boston et al., 2008; Brouwer et al., 2010; Demberg & Keller, 2008) as well as N400 amplitude (Frank et al., 2015). Interestingly, Frank et al. (2015) interpret the N400 effect of surprisal as supporting the memory-retrieval rather than the integration account of the N400, since retrieving lexical information associated with a word is predicted to be easier when the word is more predictable. The integration account was excluded based on the observation that surprisal was estimated by language models that are only minimally (if at all) sensitive to semantics. Crucially, this may be the reason why Frank et al. (2015) failed to find surprisal effects on the P600 component (spatiotemporal component overlap being another possible explanation).

Rather than using language models, in the present study, expectancy was estimated using log-transformed cloze probability (see also Smith & Levy, 2013), which arguably more closely approximates a comprehension-centric, semantic notion of

surprisal that incorporates both linguistic experience and world knowledge (Venhuizen et al., 2019). The RI account predicts this notion of expectancy/surprisal to influence both the N400 and the P600 component. First, surprisal (and lexical association, among other factors) influences the ease with which the current word form is mapped to its word meaning (N400). Second, surprisal influences the ease with which the current word meaning is integrated (P600) into the new, updated utterance meaning representation. Crucially, this integration view subsumes syntactically-, semantically-, and pragmatically-induced processing difficulties, as these may all hamper the construction of a coherent utterance meaning representation (see Brouwer et al., 2012, for discussion).

In sum, in the neurocomputational model of incremental language comprehension proposed by Brouwer, Delogu, Venhuizen, and Crocker (2021), the comprehension-centric metric of surprisal reflects the likelihood of an updated interpretation given the interpretation prior to integrating the meaning of the current word. Surprisal is thus predicted to be indexed by the P600 component, which reflects the effort involved in integrating the retrieved word meaning into the unfolding utterance interpretation: The more unexpected, unclear, or implausible the resulting utterance interpretation, the higher the amplitude of the P600.

3.5 Conclusion

In this study, we investigated the contribution of expectancy and lexical association to ERP modulations and reading times, and whether a specific locus of expectancy-related effects can be established in the ERP signal. An ERP experiment revealed that the N400 is sensitive to both expectancy and lexical association while the P600 is sensitive only to expectancy. A post-hoc, exploratory, analysis suggests that the P600 is not only evoked in response to completely unexpected (zero cloze) target words but is also modulated by the degree of expectancy in non-zero cloze targets. In two self-paced reading experiments, we found evidence for the influence of expectancy on reading times across spillover regions, whereas the presence of association effects appeared to be task dependent. Specifically, only in the experiment with comprehension questions did both expectancy and lexical association influence reading times on the Spillover region, while the effect of expectancy extended into the Post-spillover region.

Based on the Retrieval-Integration account of the electrophysiology of language comprehension, we interpret the N400 and the P600 components to index two fundamental mechanisms involved in language comprehension, namely lexical retrieval and semantic integration, respectively. We further argue that word expectancy modulates neural and behavioural processing indices by facilitating/taxing both of these cognitive mechanisms. On the one hand, the meaning of expected words as well as words that are strongly associated with the prior context is easier to retrieve from

long-term memory. On the other hand, unexpected words increase the effort involved in updating the unfolding utterance meaning representation with the retrieved word meaning. Thus, while word expectancy influences both processes – retrieval and integration – they are qualitatively different processes that map different inputs to different outputs. Retrieval maps word forms into word meaning representations, while integration takes these word meanings and maps them into an updated utterance meaning representation. This view stresses that word expectancy effects are to be interpreted in terms of their consequences for cognitive processes, rather than as a process (e.g., that of anticipation) in and of itself. As the P600 was responsive to expectancy only, we argue that this component is the primary index of comprehension-centric surprisal, quantifying the difficulty incurred by integrating an incoming word's meaning into the unfolding interpretation.

Chapter 4

The P600 as a Continuous Index of Integration Effort

The contents of this chapter were published in a peer-reviewed journal article (Aurnhammer, Delogu, et al., 2023).

4.1 Introduction

In electrophysiological studies of language comprehension, the two most salient components of the event-related brain potential (ERP) signal are the N400 and the P600. It is still under debate, however, which of these two components indexes semantic integration – the core operation of compositionally updating an unfolding utterance meaning representation with incoming information – during online language comprehension. Traditionally, semantic integration has been attributed to the N400 component (Brown & Hagoort, 1993, 2000; Hagoort et al., 2004), such that its amplitude is continuously related to integration effort, a mapping that underpins several contemporary neurocomputational models of comprehension (for a review, see Eddine et al., 2022). The P600 has traditionally been discussed in relation to syntactic and structural processing (Hagoort et al., 1993; Osterhout & Holcomb, 1992). This linkage of the N400 to semantic integration and the P600 to purely structural processing is challenged, however, by studies employing semantic role violations, such as "the hearty meal was devouring/devoured" (A. Kim and Osterhout, 2005, see also Hoeks et al., 2004; Kolk et al., 2003; Kuperberg, 2007; Kuperberg et al., 2003; van Herten et al., 2006; van Herten et al., 2005), which lead to P600 rather than N400 effects relative to baseline. To reconcile these "semantic P600" findings with the traditional functional roles of the N400 and the P600, multi-stream models have been proposed that postulate distinct cognitive mechanisms that trigger either an N400 increase or a P600 increase, but typically not both (see Bornkessel-Schlesewsky & Schlesewsky, 2008; Brouwer et al., 2012; Kuperberg, 2007, for reviews). Motivated by several limitations of these multi-stream models, Retrieval-Integration (RI) theory (Brouwer et al., 2017; Brouwer et al., 2012) offers an alternative, single-stream account which explains semantic P600 findings by interpreting the N400 as reflecting lexical retrieval (Kutas & Federmeier, 2000, 2011; Lau et al., 2009; Lau et al., 2008;

van Berkum, 2009, 2010) and reinterpreting the P600 as a *continuous* index of integration effort. We here employ an experimental design that tests the graded nature of the P600 as an index of integration effort, while also teasing apart the different predictions made by RI theory and multi-stream models about which ERP component should be modulated.

4.1.1 Multi-Stream Models

Multi-stream models typically consist of two processing streams (but see Kuperberg, 2007): a *semantic* stream, linked to the N400, and an *algorithmic* stream linked (indirectly) to the P600. The precise mechanisms thought to underlie the streams vary: For instance, the Semantic Attraction account (SA, A. Kim & Osterhout, 2005), Monitoring Theory (MT, van Herten et al., 2006; van Herten et al., 2005), and the extended Argument Dependency Model (eADM, Bornkessel-Schlesewsky & Schlesewsky, 2008) characterise the semantic stream as assigning thematic roles based on plausibility heuristics and world knowledge, independent of morpho-syntactic cues (see also the Processing Competition account, Kos et al., 2010). In the Continued Combinatory Analysis model (CCA, Kuperberg, 2007), the *semantic memory-based stream* computes semantic features and categorical relationships between words and compares them with pre-existing relations stored in semantic memory. Finally, in a more recent model proposed by Michalon and Baggio (2019), the semantic stream constructs an interpretation of the input by assigning grammatical roles based on lexical-semantic information. While the precise conceptualisation of this stream varies across multi-stream models, the absence of an N400 effect in semantic P600 studies is explained by these accounts in a similar manner: The semantic processing stream is agnostic to the syntactic constraints of the input and thus fails to detect a semantic anomaly whenever a semantically plausible (but syntactically unlicensed) alternative interpretation can be constructed from the content words encountered thus far. In sum, multi-stream accounts typically explain the absence of an N400 effect in semantic P600 findings by positing the presence of a form of semantic attraction (for example, for the more plausible "the hearty meal was devoured" upon encountering "devouring"; see Li and Ettinger, 2023; Rabovsky et al., 2018; Ryskin et al., 2021, for more recent instantiations of a similar line of reasoning).

The other stream, called *algorithmic stream* (van Herten et al., 2006), *syntactic stream* (A. Kim & Osterhout, 2005; Kos et al., 2010), or *combinatorial stream* (Kuperberg, 2007), has been described as constructing an interpretation of the input by taking into account morpho-syntactic cues. Again, the conceptualisation of this stream changes depending on the specific model. For example, in the eADM model, this stream assigns *thematic* roles based on syntactic "prominence" information (Bornkessel-Schlesewsky & Schlesewsky, 2008). In CCA, the combinatorial stream combines words based on morpho-syntactic constraints and is complemented with a stream sensitive to semantic-thematic constraints such as animacy (Kuperberg,

2007). In the model proposed by Michalon and Baggio (2019), the syntactic stream assigns grammatical roles based on word position and parts-of-speech.

Crucially, on these multi-stream models, semantic P600 effects do not directly result from variations in processing cost within the algorithmic stream but rather from situations in which the interpretations generated by the semantic and the algorithmic streams disagree. For example, at the word "devouring', the algorithmic stream assigns the syntactically cued role of *agent* to "meal", which conflicts with the interpretation generated by the semantic stream in which "meal" is the *theme* for "devour". It is this conflict that is posited to result in a P600 effect relative to baseline. Crucially, the absence of an N400 effect together with the presence of a P600 effect for semantic anomalies such as those induced by implausible thematic roles depends on the availability of a semantically attractive alternative interpretation, for instance, one in which the thematic roles are reversed. If such an alternative is not present, multi-stream models predict an N400 increase indexing integration difficulty for the anomalous word in the semantic stream, but no P600 increase, as the outputs of the streams should not conflict.

4.1.2 Retrieval-Integration Theory

Retrieval-Integration theory proposes an alternative, single-stream account in which the N400 is taken to reflect retrieval of word meaning and the P600 is taken to index semantic integration effort (Brouwer et al., 2017; Brouwer et al., 2012).

Conceptually, RI theory relies on a notion of retrieval that is grounded in the semantic access/retrieval view of the N400 (Kutas & Federmeier, 2000, 2011; Lau et al., 2009; Lau et al., 2008; van Berkum, 2009, 2010), on which semantic/conceptual knowledge associated with a word form – that is, its meaning – is accessed in long-term memory. This retrieval process is cued both by association and by expectation and, indeed, associative and expectation-based influences on retrieval facilitation have been shown to manifest in additive N400 modulations (Chapter 3). Critically, while associative and expectation-based influences join in facilitating retrieval of word meaning for the current word form, RI theory assumes this process to be non-combinatorial and non-compositional in nature. That is, while the utterance meaning representation influences retrieval of word meaning, the retrieval process itself, as reflected in the N400, does not entail any form of compositional update of the utterance meaning representation.[1] Integrative processes are instead manifest in the P600 component. Conceptually, integration is the updating in working memory of

[1]This perspective on retrieval separates RI theory from the hybrid view of the N400. On RI theory, retrieval is taken to include both what has, on the hybrid view, been called pre-activation – the process by which "the semantics of the context activates lexical features of an incoming word" (Baggio & Hagoort, 2011, p. 1348) and the process by which "different sources of information converge on a common memory representation" (Baggio & Hagoort, 2011, p. 1347, the hybrid view calls the latter notion "integration" and does not posit this process to be reflected in the N400). RI theory diverges from the hybrid view, in that the latter additionally posits unification – the "integration of word meaning into an unfolding representation of the preceding context" (Hagoort et al., 2009, p. 1) – to be indexed by the N400. This update is what RI theory calls integration and attributes to the P600.

the incrementally constructed utterance meaning representation with the retrieved word meaning. On the RI account, this notion of integration implies a combinatorial process that relies not only on semantic, but, critically, also on pragmatic and morpho-syntactic information.

More explicitly, RI theory posits that the word-by-word processing of a sentence is defined by the *process* function (Brouwer, Delogu, Venhuizen, & Crocker, 2021):

$$
\begin{aligned}
&\textit{process(word form, utterance context)} && \mapsto \textit{utterance representation} && \\
&\quad \textit{retrieve(word form, utterance context)} && \mapsto \textit{word meaning} && [\sim \text{N400}] \\
&\quad \textit{integrate(word meaning, utterance context)} && \mapsto \textit{utterance meaning} && [\sim \text{P600}]
\end{aligned}
$$

Incoming word forms are mapped onto an utterance representation while taking utterance context, i.e., the utterance representation constructed so far, into account. The process function is, however, divided into two sub-processes – *retrieve* and *integrate* – which are linked to the N400 and the P600 component, respectively. The *retrieve* function maps incoming word forms onto a representation of word meaning while taking utterance context into account. In the neurocomputational model instantiation of the theory (Figure 4.1), the N400 is taken to be proportional to the distance of the **retrieval** layer at the previous processing step to that at the current processing step. The retrieval process is facilitated – and N400 amplitude attenuated – when the meaning of an incoming word is primed associatively or contextually. The absence of an N400 effect for "the hearty meal was devouring/devoured" is explained by the similar associative priming that both target words receive from the context. Thus, the process underlying the N400 is restricted to accessing word meaning in long-term memory and mapping it into working memory and extends neither to quasi-compositional integration – as proposed by several multi-stream models – nor to compositional integration of word meaning with the utterance meaning representation constructed up to that point, as proposed by the integration view of the N400. The output of the *retrieve* function serves as an input to the *integrate* function, which maps the retrieved word meaning onto an updated utterance meaning representation while taking previous utterance context into account. The P600 is taken to proportionally reflect the distance in activation between the **integration** layer at the previous processing step and that at the current processing step. The P600 increase for "devouring" compared to "devoured" thus results from a more difficult integration process due to the implausibility of *meal* fulfilling the agent role.

The interpretation of the P600 as an index of integration effort is, however, not limited to role-reversal manipulations but naturally extends to those semantic P600 findings induced not only by semantic and pragmatic factors (Burkhardt, 2006, 2007; Cohn & Kutas, 2015; Delogu et al., 2019; Dimitrova et al., 2012; Hoeks et al., 2013; Regel et al., 2010; Schumacher, 2011; Spotorno et al., 2013; Xu & Zhou, 2016) but also those induced by manipulations of syntax (Gouvea et al., 2010; see Brouwer

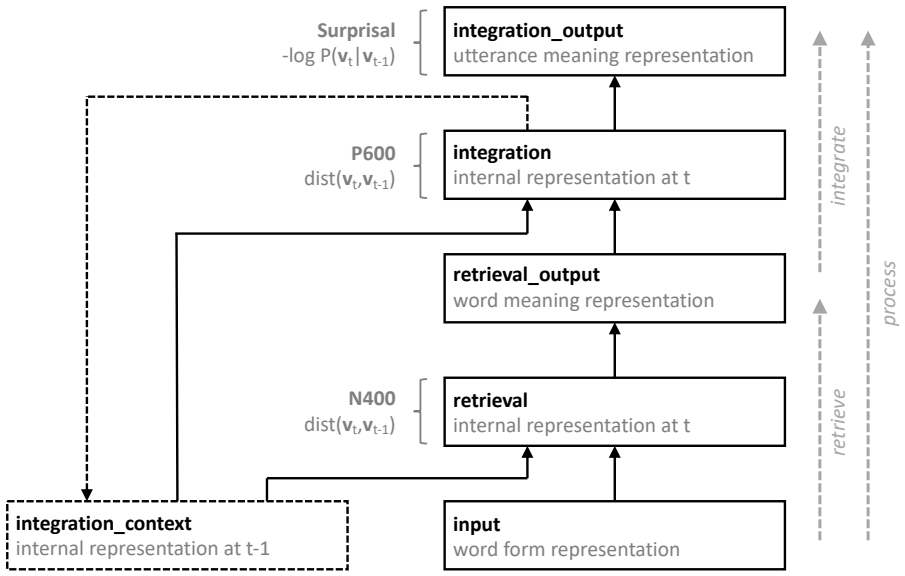


FIGURE 4.1: Schematic architecture of the neurocomputational instantiation of Retrieval-Integration theory, implementing word-by-word language processing through the *retrieve* and *integrate* functions. For full detail on model implementation see Brouwer, Delogu, Venhuizen, and Crocker (2021).

et al., 2012; Delogu et al., 2019, for discussion) and syntax-driven semantic composition (Fritz & Baggio, 2020, 2022). Importantly, on the RI account, the amplitude of the P600 should not be a binary response to violating stimuli, but should rather be sensitive to integration effort on a continuous scale (Brouwer et al., 2012), reflecting comprehension-centric surprisal (Brouwer, Delogu, Venhuizen, & Crocker, 2021). Preliminary evidence for this prediction has been presented in a post-hoc analysis in Chapter 3, where we demonstrated a graded response of both the N400 and the P600 to congruous sentences that varied in target word expectancy.

Crucially, the notion of integration assumed by RI theory is not coextensive with the aspects of integration proposed for the semantic stream by multi-stream models. Rather, integration in the RI model is closer to the algorithmic stream, in that integration is posited as morpho-syntactically constrained utterance meaning composition. Importantly, however, while most multi-stream models do not directly attribute any electrophysiological processing correlate to the algorithmic stream, RI theory takes the P600 to be directly proportional to the change in utterance meaning representation induced by the current word meaning.

4.1.3 Disentangling Multi-Stream Models and RI Theory

While both multi-stream models and RI theory can account for semantic P600 effects elicited in the presence of semantic attraction (for example, caused by role reversals), they differ in predicting which component should reveal integrative effort

Introduction A tourist wanted to bring his huge suitcase onto the airplane. However, because the suitcase was so heavy, the woman behind the check-in counter decided to charge the tourist extra. In response, the tourist opened his suitcase and threw some stuff out. So now, the suitcase of the resourceful tourist weighed less than the maximum twenty kilos.	
Coherent continuation Next, the woman told the *tourist* that she thought he looked really trendy. The tourist grabbed the woman's hand and eagerly asked her for a date. But the woman reprimanded the tourist for being pushy and told him to just get on the plane right away.	*Incoherent continuation* Next, the woman told the *suitcase* that she thought he looked really trendy. The suitcase grabbed the woman's hand and eagerly asked her for a date. But the woman reprimanded the suitcase for being pushy and told him to just get on the plane right away.

TABLE 4.1: Experimental stimulus from the design of Nieuwland and van Berkum (2005), translated from Dutch. Underlines added by the author of this dissertation.

in the absence of a semantically attractive alternative interpretation. As previously discussed, multi-stream models predict an N400 effect reflecting an unrepairable semantic anomaly and no P600 effect, as no conflict should arise between the semantic and the algorithmic stream, relative to a plausible baseline. By contrast, the RI account predicts the N400 to be modulated by the degree to which the meaning of the implausible word is associatively primed and contextually expected, and a P600 effect reflecting continuous semantic integration effort, relative to a plausible baseline.

Semantic P600 Effects in a Wider Discourse

Here, we present an experimental design that directly tests the predictions of multi-stream models against those of RI theory. To this end, we build on the design employed by Nieuwland and van Berkum (2005) in which a context paragraph is followed by a critical region including either a plausible (coherent: "the woman told the tourist") or an implausible (incoherent: "the woman told the suitcase") target word (Table 4.1). Crucially, both target words, "tourist" and "suitcase", are mentioned several times in the preceding context paragraph. Stimuli were presented in spoken form and without a task. The contrast of the implausible (incoherent) "suitcase" to the plausible (incoherent) "tourist" elicited a broadly distributed P600 effect, but no N400 effect.

This result seems inconsistent with multi-stream accounts: When encountering the implausible target word "suitcase", there is no *locally* available semantically

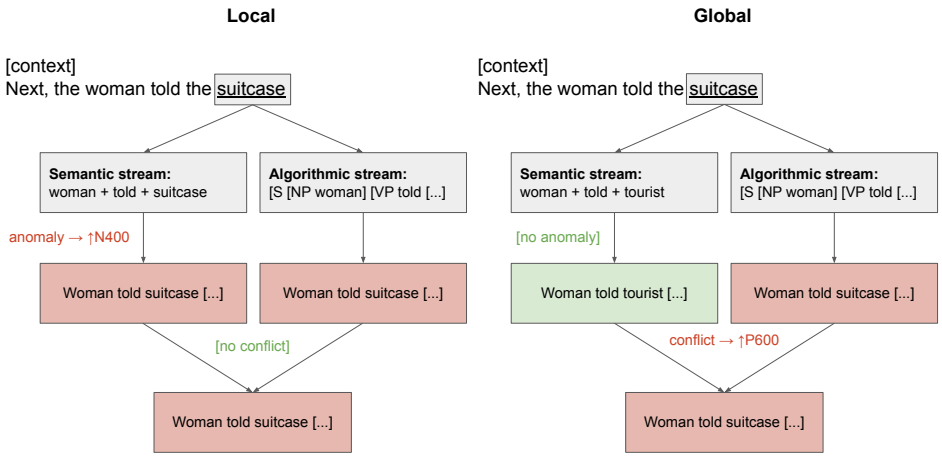


FIGURE 4.2: Schematic overview of multi-stream explanations assuming either a *local* or a *global* revision mechanism.

attractive alternative – for example, through sentence-internal permutation of thematic roles and/or morphological inflexion – that would yield a plausible interpretation of the sentence. As a result, multi-stream models predict an N400 effect, reflecting the difficulty in arriving at a semantically plausible analysis when compared to a plausible sentence, but no P600 effect, as there is no disagreement between the independent semantic stream and the algorithmic stream (see Brouwer et al., 2012, for discussion; a schematic multi-stream analysis is given in Figure 4.2, left). It has been argued, however, that a semantically attractive alternative may be *globally* available in the larger discourse (see Bornkessel-Schlesewsky & Schlesewsky, 2008; Kuperberg, 2007, for discussion). That is, as both "tourist" and "suitcase" are salient entities in the discourse, which have been mentioned numerous times, the interpretation of the coherent condition ("the woman told the tourist") may actually be a strong attractor in the incongruent condition. In other words, the salience of the plausible noun phrase "the tourist" may *distract* the system away from the actual noun phrase "the suitcase". If this is the case, a multi-stream account of this result would entail the independent semantic stream encountering no difficulty in producing a plausible analysis, which should lead to no N400 modulation, thereby yielding a conflict with the algorithmic processing stream (which arrives at the analysis "the woman told the suitcase"), thereby triggering a P600 effect relative to baseline (see Figure 4.2, right).

Retrieval-Integration theory attributes the absence of an N400 effect in Nieuwland and van Berkum (2005) to facilitated retrieval. That is, the lexical repetition of both the congruent and incongruent target words leads to maximal priming of their meaning. Indeed, in line with this interpretation, the N400 effect resurfaced for similar stimuli presented in story-initial position, that is, without any preceding context mentioning the target words (see Figure 4 in Nieuwland & van Berkum, 2005),

due to the absence of equal priming for "suitcase" and "tourist".[2] The presence of a P600 effect, in turn, reflects the difficulty in integrating "suitcase" versus "tourist" in "the woman told [...]", as the former yields an interpretation that goes against world knowledge. If we accept the independent processing streams of multi-stream models to be able to compute a *globally* available semantically attractive alternative interpretation, then multi-stream models and RI theory make the same N400 and P600 predictions, and both account for the Nieuwland and van Berkum (2005) data. Crucially, however, if no such alternative interpretation is available, the accounts make diverging predictions: multi-stream models predict an N400 effect and no P600 effect, while RI theory predicts no N400 effect and a P600 effect relative to baseline. Further, while previous studies observing semantic P600 effects typically employed binary designs, RI theory makes the specific prediction that P600 amplitude should be a function of graded integration effort (see Chapter 3 for preliminary support). To test these diverging predictions, we here present an adapted version of the Nieuwland and van Berkum (2005) design.

Global Attraction Versus Continuous Integration

Specifically, the new design implements several manipulations (see Table 4.2). First, we created a baseline condition, in which the target word is expected and plausible and no processing difficulties should ensue (Condition A). In order to test the prediction of multi-stream models that it is the availability of a semantically attractive alternative that explains the absence of an N400 effect and the presence of a P600 effect, we constructed one condition such that an alternative is made *globally* available by a distractor word in the context (Condition B). In another condition, no such alternative is available (Condition C) and we compare both conditions to the unmanipulated baseline (Condition A). Furthermore, to test for the gradedness of integration effort, the target word in Condition B has intermediate plausibility, in that it renders the interpretation semantically unlikely yet possible, while Condition C is implausible, yielding a semantic anomaly (see Table 4.4 for more examples). Finally, to maximise comparability of target word processing across conditions, our design employs a context rather than a target manipulation design and we harness lexical repetition to maximally and equally prime the target word meaning in the three conditions.

In the adapted design, multi-stream models predict a P600 effect and no N400 effect for Condition B relative to Condition A (see Table 4.3). This is because the anomaly is repairable by replacing the anomalous interpretation resulting from the observed word with the *globally* available alternative interpretation that derives from the distractor word, similar to the original study. In Condition C, however, no such alternative interpretation is licensed by the context and, hence, multi-stream models

[2] Visual inspection suggests that this N400 effect co-occurs with an increase in P600 amplitude.

Context
Ein Tourist wollte seinen riesigen **Koffer** mit in das Flugzeug nehmen. Der **Koffer** war allerdings so schwer, dass die Dame am Check-in entschied, dem Touristen eine extra Gebühr zu berechnen. Daraufhin öffnete der Tourist seinen **Koffer** und warf einige Sachen hinaus. Somit wog der **Koffer** des einfallsreichen Touristen weniger als das Maximum von 30 Kilogramm.

*A tourist wanted to take his huge **suitcase** onto the airplane. The **suitcase** was however so heavy that the woman at the check-in decided to charge the tourist an extra fee. After that, the tourist opened his **suitcase** and threw several things out. Now, the **suitcase** of the ingenious tourist weighed less than the maximum of 30 kilograms.*

Condition A: Plausible, no attraction
Dann verabschiedete die Dame den Touristen und danach ging er zum Gate.
Then dismissed the lady the tourist and afterwards he went to the gate.

Condition B: Less plausible, attraction
Dann wog die Dame den Touristen und danach ging er zum Gate.
Then weighed the lady the tourist and afterwards he went to the gate.

Condition C: Implausible, no attraction
Dann unterschrieb die Dame den Touristen und danach ging er zum Gate.
Then signed the lady the tourist and afterwards he went to the gate.

TABLE 4.2: Experimental design of the present study. German word order is preserved for the English transliterations of the final sentences. Target words are underlined and distractor words are highlighted in boldface.

predict an N400 effect and, critically, no P600 effect relative to the baseline condition. RI theory predicts that no N400 differences should be produced between conditions due to the lexical repetition of the target word in the context paragraph maximally facilitating lexical retrieval of its meaning. Under the hypothesis that P600 amplitude continuously indexes the effort of integrating word meaning with the utterance meaning representation constructed so far, the P600 is predicted to be graded for plausibility with increasing amplitude for conditions $A < B < C$. In sum, while multi-stream models predict a P600 effect for Condition B and an N400 effect for Condition C relative to the baseline Condition A, RI theory predicts the absence of N400 effects and graded P600 amplitude differences across conditions.

On the assumption that reading times provide an index of overall word-by-word processing effort, we first collected self-paced reading time data for our novel design. We expect that reading times should be graded for target word plausibility, reflecting graded integration effort. Subsequently, we recorded event-related potentials for the same stimuli, allowing for a direct comparison between behavioural and neurophysiological indices of integrative processing effort (see Brouwer, Delogu, Venhuizen, & Crocker, 2021, for discussion).

	Multi-stream		Retrieval-Integration	
	N400	P600	N400	P600
A: Plausible, no attraction	-	-	-	-
B: Less plausible, attraction	-	+	-	+
C: Implausible, no attraction	+	-	-	++

TABLE 4.3: Predictions of multi-stream models and Retrieval-Integration theory for the N400 and the P600 component in the current design.

4.2 Experiment 4: Self-Paced Reading

4.2.1 Method

Materials

The materials were optimised to be used in the same form in the self-paced reading study and the electroencephalography (EEG) study (see Appendix B.2 for the full list of German stimuli). In the creation of the stimuli, we translated and adapted items from Nieuwland and van Berkum (2005) where possible, and otherwise developed new items. In total, we developed 96 items for which we changed the original target manipulation to a context manipulation design. Employing a context manipulation design in which the target word is the same in every condition is intended to reduce effects due to differences in word length, frequency, and so forth. Every item had the same context paragraph in each condition.

The context paragraph repeatedly mentioned both the target word as well as a distractor word. The target word and the distractor word were mentioned the same amount of times within item (three or four times). Presenting the target word several times in the context paragraph should maximally prime the target word's meaning when presented in target position. Under RI theory, we thus expect no N400 (retrieval) effect between conditions (see Brouwer & Crocker, 2017; Brouwer et al., 2012). Which of the two words – target or distractor – was last mentioned in the context was approximately balanced across items.

The context paragraph was followed by a manipulated final sentence. Conditions differed only in the main verb of the final sentence, rendering the target word of the sentence – that is, the direct object – plausible (Condition A, "the lady *dismissed* the tourist"), less plausible (Condition B, "the lady *weighed* the tourist"), or implausible (Condition C, "the lady *signed* the tourist"). Indeed, Condition C creates a standard semantic anomaly by violating the selectional restrictions of the main verb. The only important difference to a standard semantic anomaly is that the target word has been presented several times before appearing in target position.[3] Taken together,

[3]Furthermore, most of these semantic anomalies render reference transfer to a related entity unlikely. For instance, while it is conceivable that reference may be transferred from "tourist" to the "tourist's ticket" in the example stimulus, for most of our stimuli no such reference transfer is licensed (for example, "the apprentice ate the hammer").

Item 2
A teacher saw an old world map in the showcase of an antique shop. Such an authentic artefact appeared suitable for his classroom and he approached the **saleswoman**...
A: Then bought the teacher the map ...
B: Then kissed the teacher the map ...
C: Then filled the teacher the map ...

Item 4
While building a table, a **carpenter** broke his nice hammer into pieces...
A: Then took the apprentice the hammer ...
B: Then sneered-at the apprentice the hammer ...
C: Then ate the apprentice the hammer...

Item 11
In a foreign city, a vacationer booked a guided tour. The **guide** was happy that the vacationer was interested and gifted him a flyer...
A: After the tour folded the vacationer the flyer ...
B: After the tour commended the vacationer the flyer ...
C: After the tour cooked the vacationer the flyer ...

Item 18
A young lady wanted to have a **jewel** evaluated by a jeweller...
A: Delighted remunerated the lady the jeweller ...
B: Delighted marveled-at the lady the jeweller ...
C: Delighted seasoned the lady the jeweller ...

TABLE 4.4: Four example items, transliterated from German. Target words are underlined and distractor words are highlighted in boldface.

this allows us to assess whether differences in plausibility result in graded modulations of both RTs and P600s. Additionally, the distractor word, which was never presented in target position, was either expected (Condition B, "the lady *weighed*" attracting "suitcase") or not expected (conditions A and C), allowing us to investigate whether the presence of a semantically attractive alternative interpretation modulates the presence of P600 effects (Condition B; semantic attraction) or N400 effects (Condition C; no semantic attraction) in the ERP experiment. The final sentence of each item ended with an additional clause following the target word ("[...] and afterwards he went to the gate"), which avoids placement of the target in sentence-final position and allows us to capture spillover effects in reading times. Table 4.4 shows four more transliterated items.

Cloze We collected cloze probabilities to validate the differential expectancy of both the target and distractor word across conditions. Sentence completions were collected in a web-based experiment using the software PCIbex (Zehr & Schwarz, 2018), which we also used for all other web-based norming studies and experiments

		Cloze			Plausibility		
	Cond.	Mean	SD	Range	Mean	SD	Range
Target	A	0.80	0.20	0.33-1.00	5.84	0.93	3.60-7.00
	B	0.09	0.11	0.00-0.40	3.69	1.33	1.50-6.30
	C	0.02	0.04	0.00-0.20	1.42	0.33	1.00-2.40
Distractor	A	0.05	0.90	0.00-0.33	2.53	1.34	1.10-6.30
	B	0.78	0.17	0.33-1.00	5.94	1.05	2.40-7.00
	C	0.03	0.06	0.00-0.20	1.66	0.69	1.00-4.80

TABLE 4.5: Averages, standard deviations, and ranges for the results of two norming studies that collected cloze probabilities and seven-point scale plausibility ratings for the target and the distractor word.

reported here. We did not use filler items, since the materials up to the target word do not contain any anomalies. Participants were presented with the entire context paragraph and the final sentence up to – but not including – the determiner of the target word. That is, we did not provide a determiner as the grammatical gender of German would constrain the set of possible completions. Cloze probabilities were obtained in two rounds, both optimising the contrast of high expectation for the target (Condition A) or the distractor word (Condition B). Sentence contexts for implausible words were created such that they do not raise strong expectations for any specific word (Condition C). In total, we collected responses from 90 participants who were recruited through Prolific Academic Ltd. (Prolific, 2021) and were each paid £7.50. We selected the 60 best items based on the results of the cloze task. Alternative cloze probabilities for any other word in Condition C were kept below 0.27 (mean = 0.20; SD = 0.07). The resulting cloze probabilities for the target and distractor word across the three conditions are presented in Table 4.5 and Figure 4.3 (left). Target word cloze probability is high in Condition A, indicating high expectancy of the target word in the baseline condition, which should therefore induce only low integrative effort. In Condition B, participants actively produced the distractor word rather than the target word, indicating that the distractor word indeed makes a semantically attractive alternative interpretation available in this condition. In Condition C, expectancy of both the target word and the distractor word was low. The latter suggests that the alternative interpretation available for Condition B is removed in Condition C. In sum, the cloze probabilities suggest that the availability of the semantically attractive alternative interpretation has been manipulated successfully (Condition A: baseline; Condition B: semantic attraction; Condition C: no semantic attraction). We turn to a second norming study in which we collect plausibility ratings to discern whether the target words of conditions B and C – which were similarly unexpected – indeed differ in their plausibility.

Plausibility In a second norming study, we collected plausibility ratings for the target and distractor words on a seven-point Likert scale, with 7 indicating "very

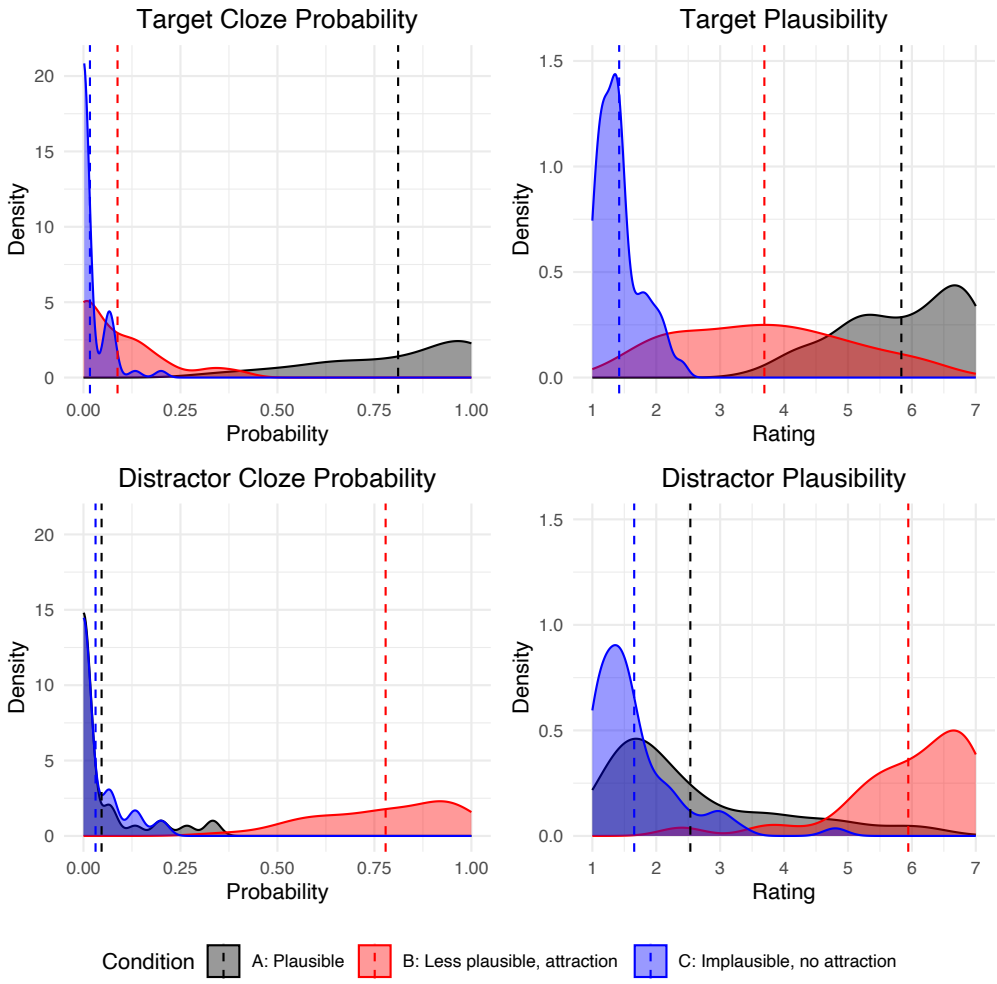


FIGURE 4.3: Densities for the results of two norming studies that collected cloze probabilities and seven-point scale plausibility ratings for the target and the distractor words. Vertical lines indicate per-condition averages.

plausible" and 1 indicating "not plausible". In total, 60 participants were recruited through Prolific Academic Ltd. and were each paid £7.50. For the rating task, the final sentence was presented in one paragraph together with the context material, to ensure reading of the entire paragraph and not only the final sentence. Participants were instructed to rate the plausibility of the final sentence in light of the context. We excluded the final sentence continuation ("and afterwards he went to the gate") to maximise rating the target word rather than another part of the final sentence. During the rating task, there were 10 items with attention checks which presented mid-paragraph instructions to rate this trial with a given number (either 1 or 7). On average, participants completed 98% of attention checks successfully (mean = 98.19%; SD = 4.09%; range 83.33 - 100.00%). The resulting plausibility ratings are reported in Table 4.5 and Figure 4.3 (right). Target word plausibility is stepped across conditions (A > B > C), which should thus result in a similarly graded effect of integration effort on the target in the three contexts. Distractor word plausibility is high

		Cloze		Plausibility	
		Target	Distractor	Target	Distractor
Cloze	Target	1.00			
	Distractor	-0.40	1.00		
Plausibility	Target	0.79	0.01	1.00	
	Distractor	-0.24	0.88	0.22	1.00

TABLE 4.6: Correlations between cloze probabilities and plausibility ratings of the target and distractor words.

in Condition B while in Condition A and C, distractor word plausibility is low, again supporting the availability of a semantically attractive alternative interpretation in Condition B.

Correlations between target and distractor word cloze probability and plausibility are reported in Table 4.6. Our analyses will focus on target word plausibility to investigate graded effects of plausibility and on distractor cloze to investigate additional effects of semantic attraction. As the correlations show, these predictors are effectively independent ($r = 0.01$).

Participants

Forty-three participants were recruited through Prolific Academic Ltd., to take part in a web-based self-paced reading experiment. One participant was excluded due to inattentive reading, as shown by low accuracy on the task (60% correct; see below for specifics of the task). The remaining 42 participants (mean age 24.43; SD 3.7; age range 18-32; 15 male, 27 female) were all native speakers of German (two early bilinguals) and had not indicated any language-related disorder or literacy difficulty. They did not participate in any other studies reported in this article. All participants gave their consent by agreeing to a consent form and were paid £7.50 for their participation.

Procedure

We conducted the self-paced reading experiment as a web-based study. On each trial, participants were prompted to press the Enter key to start, after which they were presented with a context paragraph. Upon pressing the Enter key again, a hash sign was presented centrally, indicating the position of the words of the final sentence. From here on, participants pressed the Space bar to proceed to the next word, each presented centrally. After three practice items, stimuli were presented in three blocks with 35 items each, summing to a total of 105 items, 45 of which were fillers. For half of the participants, the blocks and the items within them were presented in reverse order. On 46% of trials – half of the experimental trials and on two fifths of the fillers – participants were presented with a comprehension question to which they had to answer using either *Yes* or *No* (mapped to the *D* and *K*

keys). Comprehension questions had *Yes* and *No* as correct answers on 50% of the questions and they could concern the context paragraph or the final sentence, within which they could focus on the manipulated region or the final sentence completion. To encourage attentive reading, we provided coarse feedback on participants' response accuracy after the practice session and after each block. Participants were encouraged to take a short break between blocks.

Analysis

We excluded trials if reading time on any critical region was lower than 50 ms or higher than 2500 ms and if reaction time on the task (if there was one on that trial) was lower than 50 ms or higher than 6,000 ms. Based on these criteria, 47 of 2520 trials were excluded (1.87%). All results and analyses reported below are computed after exclusion.

Log-transformed reading times were analysed with a linear mixed effects regression re-estimation technique (cf. Chapter 3), using the `MixedModels` package for `Julia` (Bezanson et al., 2017). This technique fits statistical models separately on each region of interest, allowing us to trace across regions the relative influence and significance of each predictor in the regression equation as well as the residual error, i.e., the difference between the observed data and the forward estimates computed by the models. As predictors of interest, we focus on target word plausibility and distractor cloze probability. Plausibility ratings will serve as a continuous predictor to operationalise integration difficulty of the target word. Distractor cloze probability serves as a predictor that will explain any additional effort incurred by the availability of a semantically attractive alternative interpretation. Random intercepts as well as random slopes for each predictor are estimated for both subjects and items. The full model specification is

$$y_t = \beta_{0t} + S_{0s} + I_{0i} + (\beta_{1t} + S_{1s} + I_{1i})Plaus + (\beta_{2t} + S_{2s} + I_{2i})Clozedist + \epsilon_t \quad (4.1)$$

in which β_0 represents the fixed-effect intercept and β_1 and β_2 refer to the fixed-effect coefficients of plausibility and distractor cloze probability for each region t. S and I terms represent random intercepts and slopes for subjects and items. The unexplained variance in the data is represented by the residual error term ϵ. All predictors were standardized, centring their average value on zero and expressing them on a scale of standard deviations. Standardising predictors additionally has the effect that the intercept will equal the mean of the data to which the model is fitted. Plausibility was also inverted, as we predict that higher reading times ensue for lower plausibility ratings. We run separate analyses for the different regions of interest, which we treat as separate families of hypotheses. Hence, we do not correct for multiple comparisons.

	Accuracy			Reaction Time		
Cond.	Mean	SD	Range	Mean	SD	Range
A	96.7%	6.1%	80.0% - 100.0%	2900 ms	560 ms	1566 ms - 3820 ms
B	95.3%	8.3%	70.0% - 100.0%	3032 ms	567 ms	1986 ms - 4106 ms
C	96.0%	7.0%	77.8% - 100.0%	3047 ms	586 ms	2086 ms - 4259 ms

TABLE 4.7: Task performance on the comprehension questions in the self-paced reading experiment. Accuracy and reaction times were computed across subjects.

4.2.2 Results

Comprehension Questions

Participants answered comprehension questions on half of the experimental items. Descriptive metrics for accuracy and reaction times were computed across subjects. Average accuracy was 95.8% (SD = 5.2%, range = 80.0 - 100.0%). Mean reaction time was 3098 ms (SD = 619 ms, range = 1907 - 4426 ms). Accuracies and reaction times split per condition are given in Table 4.7.

Reading Times

Figure 4.4 displays log-transformed reading times, split up per condition, on the Pre-critical region (the ambiguous article "den" / "the" of the target word), the Critical region (the target word "tourist"), the Spillover region ("and") and the Post-spillover region ("afterwards"). Visual inspection of the data suggests that already on the Pre-critical and Critical regions, Condition C is read slower than Condition A and B. On the Spillover region, Condition B and C are slowed down. Lastly, on the post-spillover region, reading times appear to pattern with the three levels of plausibility across conditions A, B, and C.

We modelled the reading times as a function of target word plausibility and distractor cloze probability separately on each region. Figure 4.5 displays the estimated reading times from these models as well as the residual error, that is, the difference between the observed and the estimated reading times. Visual inspection suggests that the models capture the effect structure in the observed data as evidenced by small residual error across regions and conditions.

Figure 4.6 (left) displays model coefficients, added to their intercept, for plausibility and distractor cloze probability together with their respective z- and p-values (right). The positive coefficients for plausibility indicate that lower plausibility predicts slower reading. The coefficient for distractor cloze probability is smaller and changes sign moving from the Critical to the Spillover and the Post-spillover region, indicating that this predictor estimates slower or faster reading time depending on the region of interest. The z- and p-values demonstrate that target word plausibility significantly predicts reading times across all regions, interestingly also including

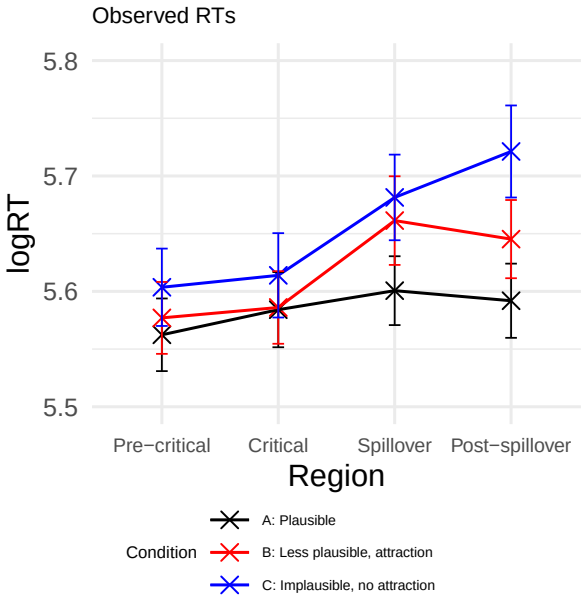


FIGURE 4.4: Log reading times, averaged per condition from the per subject averages, on the Pre-critical, Critical, Spillover, and Post-spillover regions. Error bars indicate standard errors computed from the per-subject, per-condition averages.

the Pre-critical region, while no significant contribution of distractor cloze probability is found.

Reading in the implausible Condition C is slowed already prior to the target word, presumably due to differences in processing of the main verbs preceding the targets. This raises the question to what extent reading time differences observed on and after the critical word are due to the plausibility of the target word itself, rather than due to the different contexts. To answer this question, we included the reading time on the Pre-critical region as a predictor in our statistical analyses, allowing the models to capture any pre-critical reading time offsets. We only z-scored but did not log-transform the Pre-critical RT predictor, in order to avoid identity of the dependent (logRT) and an independent variable (Pre-critical RT) on the Pre-critical region. The remaining predictors now explain any systematic variability in reading time over and above reading time offsets present at the Pre-critical region. The resulting coefficients and z-values indicate that target word plausibility significantly predicts slowed reading time at the Spillover and Post-spillover regions, over and above what is explained by Pre-critical reading time, whereas the predictor is no longer significant on the Pre-critical and Critical regions. Distractor cloze probability still does not significantly predict reading times on any region (Figure 4.7).

4.2.3 Discussion

The results of the self-paced reading experiment show that reading times scale gradually with plausibility, indicating that our manipulation of target plausibility indeed

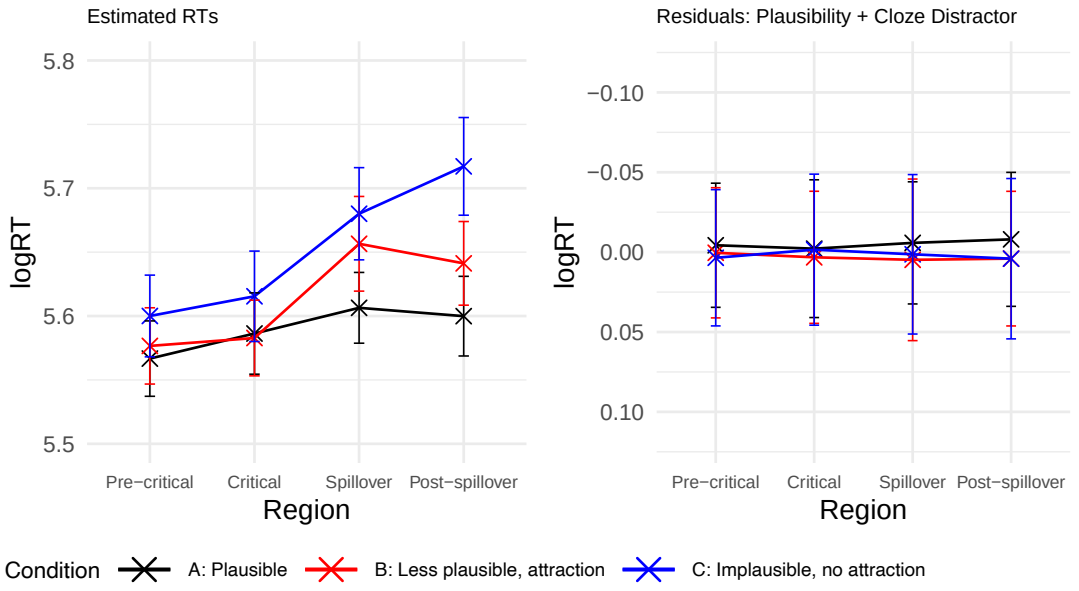


FIGURE 4.5: Estimated log-Reading Times (left) and residual error (right), averaged per condition, on the Pre-critical, Critical, Spillover, and Post-spillover regions. Error bars indicate standard errors computed from the per-subject, per-condition averages.

resulted in a graded modulation of integration effort. Furthermore, the regression-based analysis revealed that plausibility is a continuous predictor of reading time.

Based on the traditional surprisal literature (Fernandez Monsalve et al., 2012; Frank et al., 2015; Levy, 2008), it could be expected that the same items that show modulations in reading times would also elicit a graded N400 response. However, the hypothesis that the P600 reflects integration effort predicts a strong link between this component and late reading time measures (Brouwer & Crocker, 2017; Brouwer et al., 2012). Empirical evidence in support of this is provided by Brouwer, Delogu, Venhuizen, and Crocker (2021) as well as in Chapter 3, showing that reading time modulations pattern with P600 effects.[4] The obtained reading times thus offer an opportunity to investigate whether the experimental design will result in a graded N400 or P600 pattern.

The current results did not reveal significant reading time modulations due to distractor cloze probability. Hence, our results indicate no significant reading time modulation that can be attributed to the presence of a semantically attractive alternative interpretation in Condition B. However, multi-stream models typically do not make predictions for behavioural measures and hence we will not rely on this result to argue against these accounts. Our manipulation does, however, create a

[4] Additionally, effects of association, which were also reflected in N400 amplitude, modulated reading times on the first Spillover region of Experiment 3 in Chapter 3. As the current design maximally primes the targets across all conditions, no such association-related effects were expected in the current data.

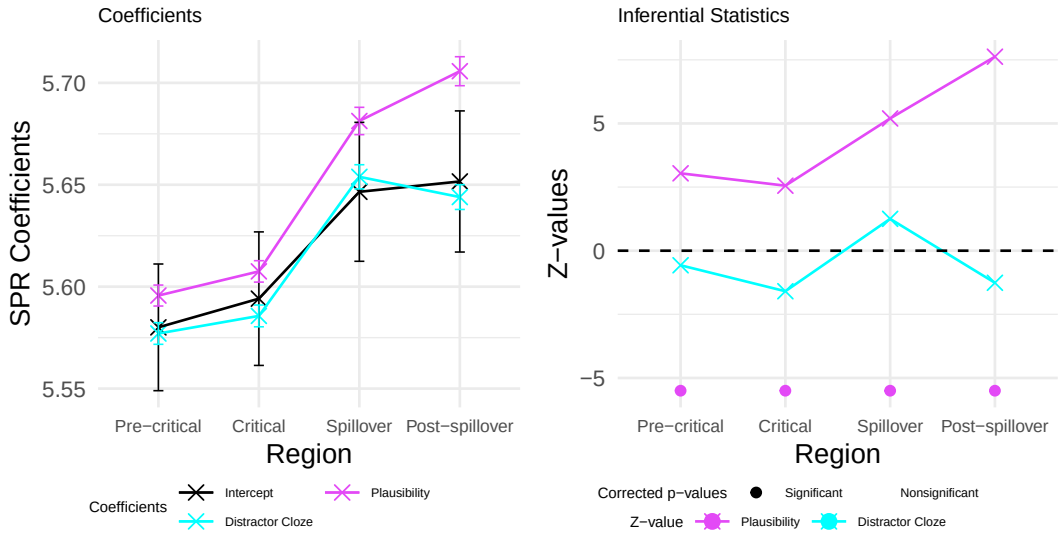


FIGURE 4.6: Coefficients (left; added to their intercept), effect sizes (z-values) and p-values (right). Error bars indicate the standard error of the coefficients in the fitted statistical model.

prediction disconfirmation, since in the context "Then weighed the lady", the expected word "suitcase" is not presented, while "tourist" is provided instead. Previous research on prediction error cost has not found disconfirmation effects in the behavioural domain using eye-tracking (Frisson et al., 2017; Luke & Christianson, 2016) or self-paced reading (Rich & Harris, 2021). In a self-paced reading experiment by van Berkum et al. (2005) a disconfirmation effect was observed - however its timing did not coincide with the ERP deflection found for the same stimuli. Similarly, lexical decision times did not exhibit facilitation effects for unrelated, unexpected words in high-constraint sentences relative to the same words in low-constraint sentences (Schwanenflugel & LaCount, 1988). In line with this previous research, our results suggest that reading times may not be sensitive to unfulfilled expectations. With regard to the comparison of multi-stream models and RI theory, the absence of a significant contribution of semantic attraction (distractor cloze probability) in behavioural measures raises the question of whether semantic attraction will modulate the presence of P600 and N400 effects in the ERP signal.

4.3 Experiment 5: Event-Related Potentials

4.3.1 Method

Materials

The materials were the same as in the self-paced reading experiment (see Section 4.2.1).

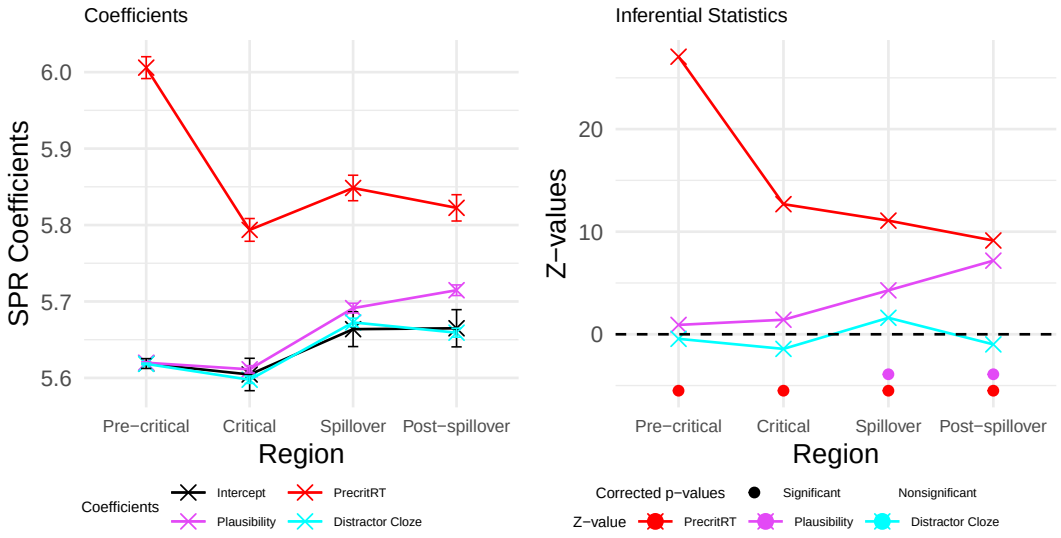


FIGURE 4.7: Coefficients (left; added to their intercept), effect sizes (z-values) and p-values (right) from models including Pre-critical reading time as a predictor. Error bars indicate the standard error of the coefficients in the fitted statistical models.

Participants

We recruited 33 participants at Saarland University to take part in the experiment. Three participants were excluded due to excessive eye movement artefacts. The final 30 participants (mean age 25; SD = 3.35; range 18-32; 25 female, 5 male) were right-handed, native speakers of German (six early bilinguals) and had normal or corrected-to-normal vision. No participant reported any form of colour blindness. Participants gave informed, written consent and were paid 25€.

Procedure

We recorded the EEG while the participants were seated in an electromagnetically shielded, sound-proof, and dimly lit chamber. The experiment was presented using E-prime 3 (Schneider et al., 2002). We first presented three practice items, two of which included a comprehension question. Practice items varied in their degree of plausibility. The practice session was followed by three blocks, each containing 35 items, including the same fillers that were used in the self-paced reading experiment. Participants took a break between blocks. Items were presented in pseudo-randomised order. For half of the participants, the blocks and the items within them were presented in reverse order. On each trial, participants used a button-box to start the item and were presented with the entire context paragraph which remained on the screen until the button was pressed again. Then, a fixation cross appeared in the centre of the screen for 750 ms. After that, the final sentence was presented using rapid serial visual presentation (RSVP). Each word of the final sentence was

presented centrally for 350 ms with a 150 ms inter-stimulus interval. If the item contained a comprehension question, the question appeared after the last word of the final sentence. Questions were answered using two buttons that mapped to *Yes*/*No*, highlighted on the screen in green and red colour, respectively. The position of the correct and incorrect buttons varied randomly in order to avoid motor preparation effects.

Electrophysiological Recording and Processing

The EEG was recorded using 26 active Ag/AgCl electrodes, positioned on the scalp following the standard 10-20 system. During recording, FCz was used as online reference and AFz as ground. Data were digitized at a sampling rate of 1000 Hz, leading to a temporal resolution at 1 ms increments. Eye-movement artefacts were monitored through the electrooculogram of two electrodes placed horizontally at the outer canthi of each eye and two electrodes placed vertically above and below the left eye. We aimed to keep impedances below $5k\Omega$ on scalp electrodes and below $10k\Omega$ on eye electrodes and did not apply online filtering. We re-referenced the EEG offline to the averages of the left and right mastoid electrodes and band-pass filtered the data between 0.01 Hz and 30 Hz. Epochs ranging from -200 to 1200 ms relative to target word onset were extracted from the EEG signal. We excluded 309 out of 1800 trials (17.17%) with ocular and muscular artefacts using a semi-automatic procedure. Baseline correction was performed based on the 200 ms pre-stimulus interval.

Analysis

To analyse the data, we apply rERPs (Smith & Kutas, 2015a), a regression-based ERP (re-)estimation technique (implemented in `Julia`; Bezanson et al., 2017), similar to the analysis used for the self-paced reading data. For this analysis, we apply linear regression, as opposed to linear mixed-effects regression, as the analytical solution of solving least-squares regression will provide stable models and faster computation speed. This will allow us to re-estimate the data on all electrodes and inspect topographic differences in the analyses. In particular, rERPs apply within-subjects regression and the models' parameters and forward estimates are averaged across subjects, analogous to the traditional ERP averaging procedure in which condition averages are computed from the means of individual subjects. The advantage of the rERP technique compared to traditional statistical analyses is that it allows us to gauge the relative explanatory power of target word plausibility and distractor cloze probability across time and electrodes: By computing a separate regression model for each subject on each electrode and time sample, we can trace predictor coefficients, inspect estimated waveforms and residual error, and obtain effect sizes across the temporal and spatial dimensions of the ERP signal. Crucially, this approach goes beyond simple condition contrasts, as we are interested in the continuous relationship between stimulus properties and ERPs. In fact, the rERP analyses

	Accuracy			Reaction Time		
Condition	Mean	SD	Range	Mean	SD	Range
A	95.1%	7.3%	75.0% - 100.0%	2144 ms	309 ms	1618 ms - 2781 ms
B	98.1%	6.1%	75.0% - 100.0%	2153 ms	316 ms	1459 ms - 3077 ms
C	95.5%	8.3%	62.5% - 100.0%	2182 ms	325 ms	1522 ms - 2841 ms

TABLE 4.8: Task performance on the comprehension questions in the EEG experiment. Accuracy and reaction times were computed across subjects.

themselves are only informed by the continuous by-trial stimulus properties and not by any explicit condition coding. That is, we only average by condition after fitting the models, to assess the extent to which our predictors capture the effect structure across conditions.

We will apply the same predictor combination that we used for the analysis of the reading times and model the ERP signal as a function of target word plausibility and distractor cloze probability. The model specification for the rERP models is

$$y_{ets} = \beta_{0ets} + \beta_{1ets} Plaus + \beta_{2ets} Clozedist + \epsilon_{ets} \tag{4.2}$$

For each electrode e, time sample t, and subject s, we compute a separate regression model. We report coefficients (β terms), estimates (the forward estimates $\hat{y}$), and residual error (ϵ, the difference between observed data y and $\hat{y}$), averaged across subjects (s). Additionally, we will compute the same model across subjects. This has the advantage that we obtain a single t-value and p-value for each electrode and time sample, rather than vectors of t-values and p-values (one value for each subject). However, there is still a multiple comparisons problem due to the multitude of time samples and electrodes and hence we correct p-values for the inflated false discovery rate using the method proposed by Benjamini and Hochberg (1995). We adjust p-values separately for the two time windows of interest but across nine analysed electrodes (F3, Fz, F4, C3, Cz, C4, P3, Pz, P4) and the time samples within a time window (N400: 300 - 500 ms; P600: 600 - 1000 ms).

4.3.2 Results

Comprehension Questions

Participants answered comprehension questions on half of the experimental items. Descriptive metrics for accuracy and reaction times were computed across subjects. Average Accuracy was 96.2% (SD = 3.9%, range = 87.0% - 100.0%). Mean reaction time was 2162 ms (SD = 254 ms, range = 1568 ms - 2841 ms). Accuracies and reaction times split per condition are given in Table 4.8.

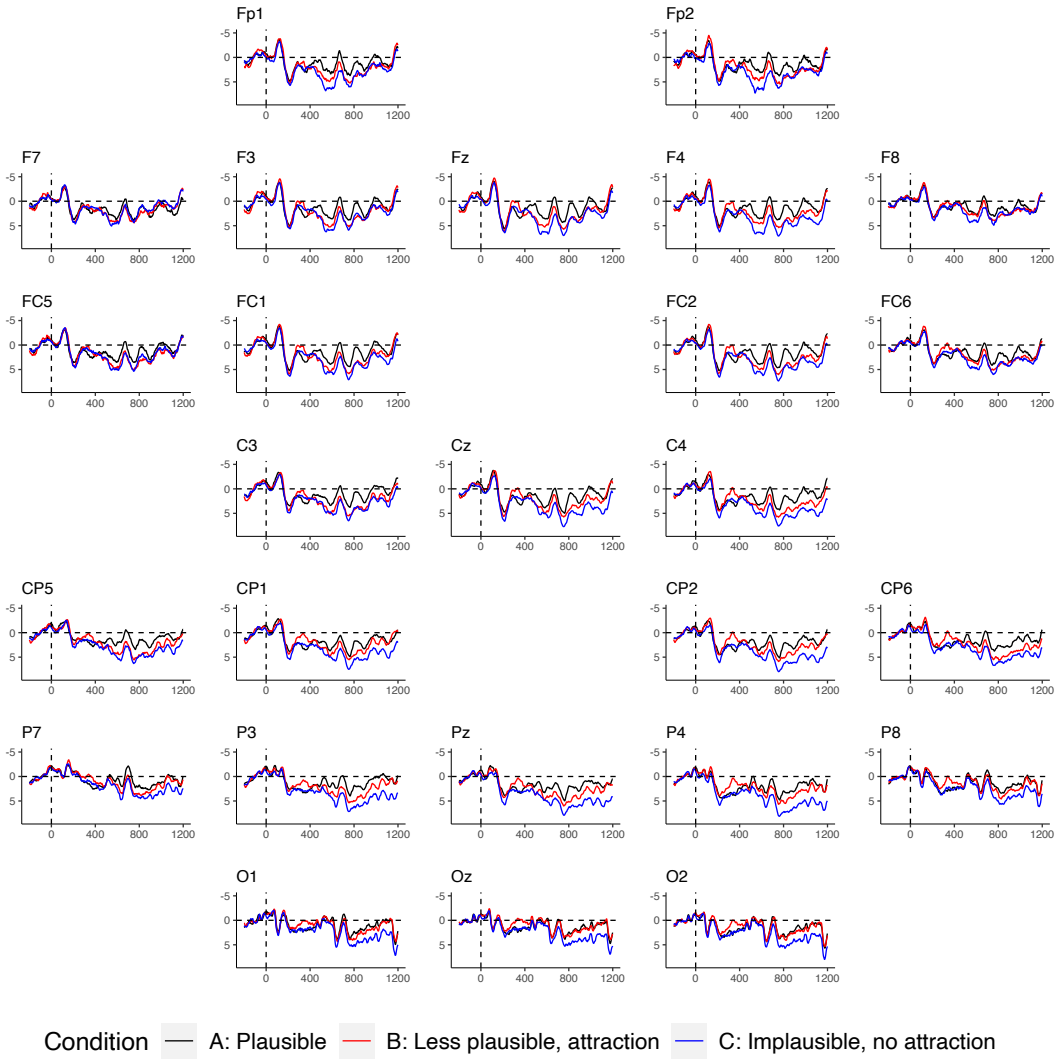


FIGURE 4.8: Grand-average ERPs in the three conditions manipulating plausibility and semantic attraction. Negative voltages are plotted upwards.

ERPs

Grand-averaged ERPs for the three conditions on all non-reference, non-eye electrodes are displayed in Figure 4.8. Visual inspection suggests a broadly distributed negativity, lasting approximately from 250 ms to 400 ms post stimulus onset in response to target words that are less plausible and for which a semantically attractive alternative interpretation is present (Condition B). A smaller, more frontally pronounced early negativity, lasting approximately from 250 to 400 ms post stimulus onset, is also evoked by implausible target words (Condition C) on frontal and central electrodes. Around the typical peak of the N400 component, no pattern of N400 amplitude with plausibility is observable by visual inspection. Furthermore, both Condition B (less plausible, semantic attraction) and Condition C (implausible, no semantic attraction) elicit broadly distributed positivities, emerging from 500 ms

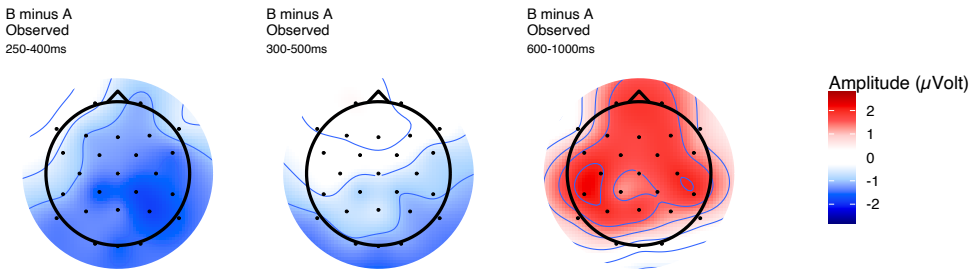


FIGURE 4.9: Topographic distributions of the average potentials of Condition B for the earlier negativity (250-400 ms), the canonical N400 (300-500 ms), and P600 (600-1000 ms) time windows, relative to the baseline condition. Topographies are computed from all non-reference, non-eye electrodes.

post stimulus onset. The positivity elicited by Condition C is stronger in amplitude than that elicited by Condition B on parietal electrodes. On left frontal electrodes, however, their amplitudes appear similar in parts of the epoch.

To further examine the topographies of the condition contrasts, we display topographic maps of the differences between the conditions in a time window matching visual inspection of the negativities (250 - 400 ms) and in the canonical N400 (300 - 400 ms) and P600 time windows (600 - 1000 ms). The topographic map of Condition B (less plausible; semantic attraction) relative to Condition A is presented in Figure 4.9. The early negativity is broadly distributed and peaks over right-parietal electrodes, whereas left-frontally, the difference is smaller. The temporal average of the N400 time window exhibits negativities over right-parietal and occipital electrodes. Inspection of the waveforms (Figure 4.8) strongly suggests that this negativity is driven by the temporally overlapping preceding negativity and that, additionally, the N400 time window also includes the onset of the P600 effect of Condition B relative to A. The late positivity has peaks both over central electrodes on both hemispheres with a trough between them.

In the topographic maps for Condition C (Figure 4.10), the early negativity appears much smaller than that of Condition B and peaks over left-frontal electrodes. The topography in the N400 time window does not contain the topography of a typical, centrally peaking N400, but more likely shows the early, emerging P600 effect. The late positivity clearly peaks over parietal electrodes.

Turning to the rERP analysis, we first inspect the estimated waveforms for a single electrode, Pz (Figure 4.11; left) as well as the residual error (right), i.e., the difference between the observed and the estimated data. The estimates were generated by a model with target word plausibility and distractor cloze probability as predictors. The estimates and residuals suggest that the models accurately capture the major trends in the data, as observable by visual inspection. That is, the models predict a negativity for Condition B between 250 and 400 ms, no negativity for Condition C

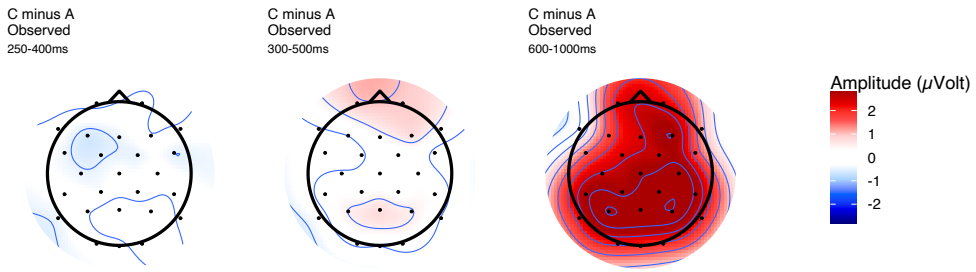


FIGURE 4.10: Topographic distributions of the average potentials of Condition C for the earlier negativity (250-400 ms), the canonical N400 (300-500 ms), and P600 (600-1000 ms) time windows, relative to the baseline condition. Topographies are computed from all non-reference, non-eye electrodes.

(on this electrode), and late positivities with increasing amplitudes for Condition B and C, respectively.

To assess which predictor captures the voltage deflections, we turn to the model coefficients, plotted over time (Figure 4.12; right). The coefficient for distractor cloze probability, which is high only in Condition B, predicts the negativity elicited by that condition. Plausibility, which is stepped across the three conditions, captures the graded late positivities. In order to assess whether distractor cloze probability also predicts a late positivity on another electrode site, we also inspect the coefficients on electrode C3 (Figure 4.12; left), on which the late positivities for Condition B and C appeared to match (Figure 4.8). Indeed, on this electrode, distractor cloze probability predicts additional positivity in parts of the P600 time window. On this electrode, plausibility also predicts a smaller earlier negativity.

Using these coefficients, we can now compute the ERPs estimated by a single predictor in isolation. To achieve this, we compute the forward estimates for the entire data set while factoring out the influence of the other predictor by fixing it to its average value, which is zero for z-scored predictors. The isolated estimates of distractor cloze on electrode Pz contain the negativity of Condition B (Figure 4.13, left). Isolating the estimates of plausibility on electrode Pz (right) reveals no modulation in the N400 time window but the three-step modulation in the P600 time window. These estimates suggest that the negativity is elicited by the expectancy of the distractor word and that plausibility predicts no N400 but P600 modulations.

As the single-electrode inspection of the coefficients suggests potential topographic differences between the contributions of the predictors, we visualise the estimated ERPs as topographic maps. This allows us to dissect how target word plausibility and distractor cloze probability interact in shaping the topographic map of the difference between Condition B and A (see Figure 4.9). Figure 4.14 displays the individual contributions of distractor cloze probability (left), target plausibility (middle left), and their sum (middle right) to the estimated data for Condition B,

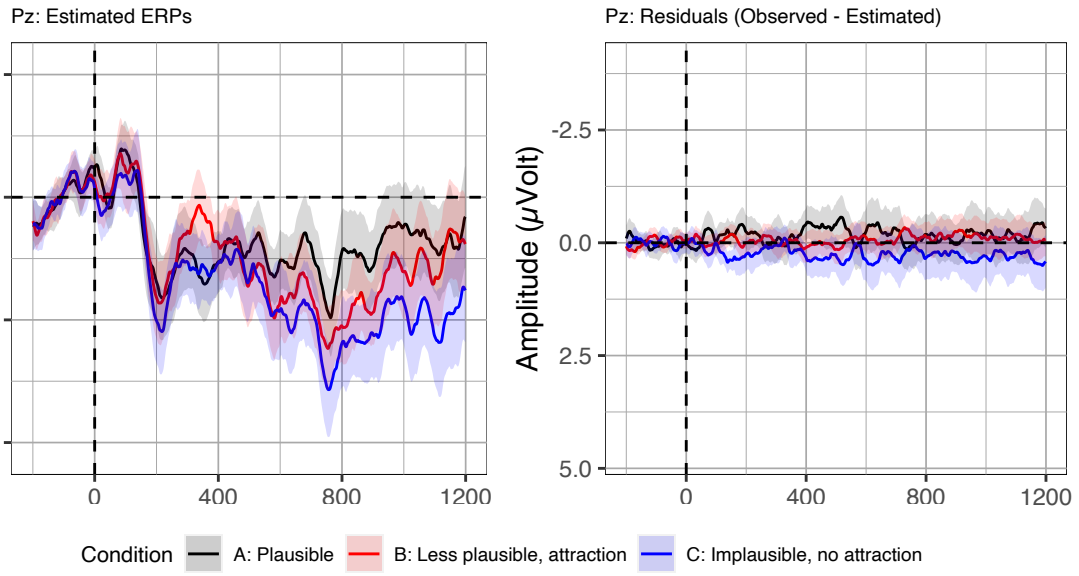


FIGURE 4.11: Estimated waveforms (left) and residual error (right) on electrode Pz based on regression models using target word plausibility and distractor cloze probability as predictors. Error ribbons indicate confidence intervals computed as 1.96 times the standard error across subjects.

which is similarly distributed to the observed data (right). The topographic maps suggest that while plausibility predicts a larger, parietally peaking positivity, there is an additional left-frontally peaking positivity, predicted by distractor cloze probability. This suggests that the overall topographic distribution observed for Condition B (Figure 4.9) is composed of a parietal and a left-frontocentral sub-component.

To assess the statistical significance of our two predictors, we computed models in which we determine the regression coefficients across all subjects, rather than fitting individual models per subject. We report the t-values for the two predictors on nine central electrodes (Figure 4.15). Furthermore, the bar below the t-values indicates time samples that were significant after correcting for multiple comparisons within the N400 and the P600 time window and across electrodes and time samples. Our inferential statistics indicate that distractor cloze probability significantly predicts a negativity in the 300 ms to 400 ms range. While the t-values for plausibility are large on frontal electrodes in the pre-N400 time window, indicative of a negativity predicted by low plausibility items, this does not reach significance in the current selection of time window and electrodes. Plausibility significantly predicts a late positivity (600 ms - 1000 ms) with a peak over parietal electrodes. Distractor cloze probability, while generating a left-frontocentral late positivity in the forward estimates (Figure 4.14), does not reach significance in our selection of late time window and electrodes.

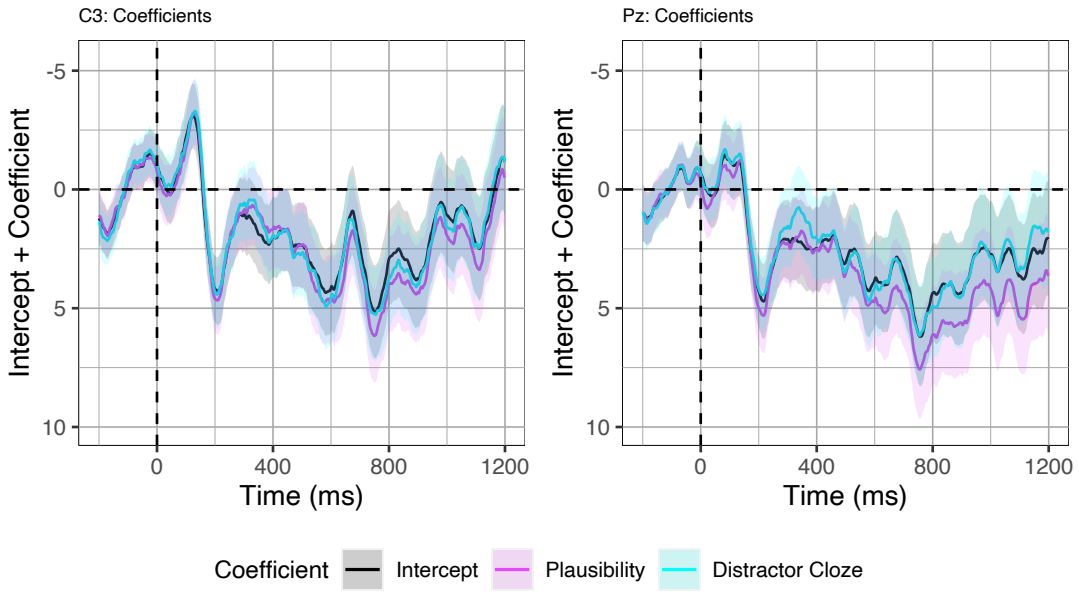


FIGURE 4.12: Regression model coefficients (added to their intercept) across time on electrodes C3 and Pz. Error ribbons indicate the standard error on the coefficients in the statistical models.

4.3.3 Discussion

Experiment 5 replicated the main findings of Nieuwland and van Berkum (2005) using visual rather than auditory language comprehension and employing an explicit task that incentivises reading for comprehension. In the original design, a context paragraph repeatedly mentioned the target words before those same words were presented either as plausible or implausible continuations. Rather than eliciting an N400 effect, a P600 effect relative to baseline was observed. This matches our data in the less plausible condition (B: "Then *weighed* the lady the tourist") compared to the baseline (A: "Then *dismissed* the lady the tourist"). Further, while a semantically attractive alternative interpretation is globally available in Condition B, it is unavailable in Condition C (C: "Then *signed* the lady the tourist"). Indeed, Condition C thus instantiates a classic semantic incongruency (see Van Petten & Luka, 2012, for a review). On multi-stream models, the absence of such semantic attraction (Condition C) should result in the emergence of an N400 effect compared to the baseline condition. However, no N400 effect but only a P600 effect was observed in Condition C relative to A. Further, our design manipulated plausibility on three levels (A: plausible $<$ B: less plausible $<$ C: implausible), showing that target words with intermediate plausibility ratings (B: "Then *weighed* the lady the tourist") also elicit a P600 effect, intermediate in amplitude, compared to the fully plausible and implausible conditions. Indeed, the plausibility ratings collected in a pre-test provided a continuous predictor which significantly predicted the P600 modulations observed across nine electrodes.

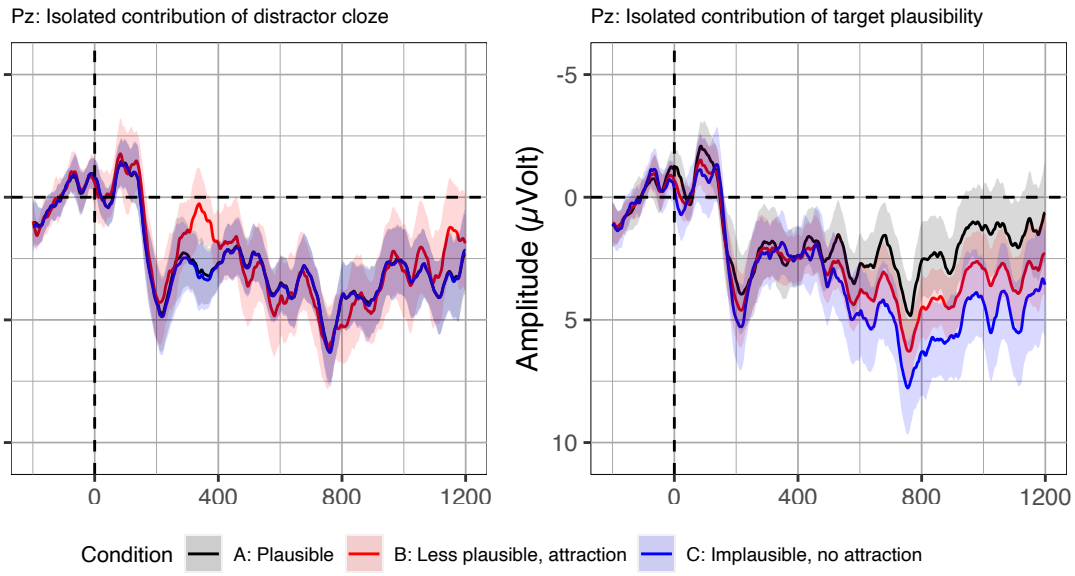


FIGURE 4.13: The isolated forward estimates of distractor cloze probability (left) and plausibility (right), based on coefficients that were fitted in models containing both predictors. Error ribbons indicate confidence intervals computed as 1.96 times the standard error across subjects.

While distractor *absence* did not elicit an N400 effect relative to baseline, the *presence* of a distractor elicited an earlier negativity, emerging from around 250 ms and lasting until 400 ms post-stimulus onset for Condition B. An interpretation of this earlier negativity as an N400 appears implausible given the temporal invariability of the N400's peak latency (Federmeier & Laszlo, 2009). Rather, we interpret this component to be elicited by the strong and unfulfilled expectation of the distractor word on a lexical level. Likely, this early component often overlaps with the N400 and it is the combination of lexical repetition and disconfirmation in our experiment that allows us to observe it in isolation. That is, even though the distractor word was strongly expected and not presented, lexical retrieval - indexed by the N400 - of the target word's meaning was still maximally facilitated. Interestingly, Nieuwland and van Berkum (2005) did not observe a similar negativity in their study, even though they relied on auditory presentation - a modality in which a component with a similar time course, the phonological mismatch negativity (PMN), is often observed (Connolly et al., 1990; Hagoort & Brown, 2000; Jachmann et al., 2019).

Further, our rERP analyses suggest that the presence of a strongly anticipated distractor word that is then *not* presented as target word (Condition B) leads to additional modulation in the late ERP signal with a positive left-frontal peak. While distractor cloze probability was not significant in the later time window, a frontal positivity could, in fact, be expected for our design, as the way in which our design makes a semantically attractive alternative interpretation available effectively

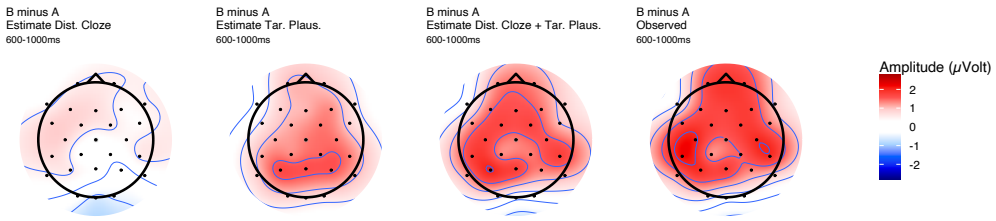


FIGURE 4.14: Topographic distributions of the potentials in the P600 time window estimated by distractor cloze probability (left), plausibility (middle left), and their summed estimated potential (middle right) as well as the observed potential (right) for Condition B between 600-1000, relative to the baseline condition. Topographies are computed from all non-reference, non-eye electrodes.

creates a prediction disconfirmation ("Then *weighed* the lady the tourist" where "suitcase" is expected), which has been linked to frontal positivities in previous research (Brothers et al., 2015; DeLong et al., 2014; DeLong et al., 2011; Federmeier et al., 2007; Kuperberg et al., 2020; Quante et al., 2018; see also earlier results by Kutas, 1993). Our rERP analysis suggests that the positivity observed for Condition B can be dissected into two sub-components: A P600 with a parietal peak, predicted by plausibility, and a disconfirmation-related positivity with a left-central peak, predicted by distractor cloze probability. In the design of Nieuwland and van Berkum (2005), a disconfirmation was also present, however, the replacement word was implausible. Their difference waves suggest no apparent deviation from a canonical, parietal P600. This is in line with the finding that the frontal positivity is produced by unexpected but *plausible* target words, whereas unexpected and implausible target words lead to a parietally distributed late positivity (Van Petten & Luka, 2012).

4.4 General Discussion

The goal of the present study was to test competing hypotheses about the functional interpretation of the N400 and P600 components. In particular, building on a previous study (Nieuwland & van Berkum, 2005), we tested the prediction of RI theory that the P600 is a continuous index of integration effort (Brouwer et al., 2017; Brouwer, Delogu, Venhuizen, et al., 2021) directly against the predictions made by multi-stream models (Bornkessel-Schlesewsky and Schlesewsky, 2008; A. Kim and Osterhout, 2005; Kos et al., 2010; Kuperberg, 2007; Michalon and Baggio, 2019; van Herten et al., 2005; and similarly Li and Ettinger, 2023; Rabovsky et al., 2018; Ryskin et al., 2021).

Multi-stream models maintain that the N400 indexes aspects of integrative or combinatorial processing of the input word with the prior context. On multi-stream accounts, no N400 modulation is generated if the processor initially does not detect an anomaly in the semantic stream because of the availability of a semantically

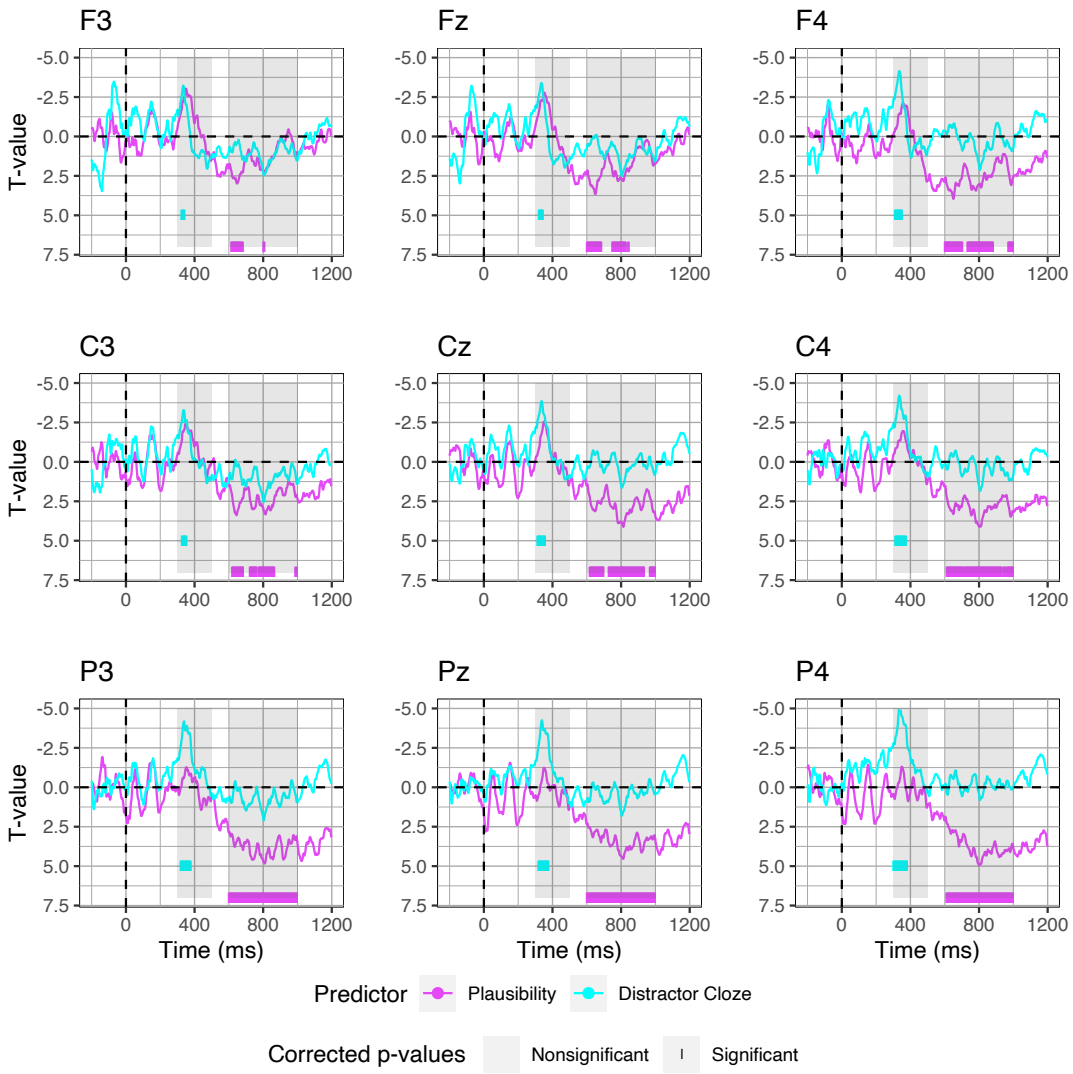


FIGURE 4.15: T-values for the predictors plausibility and distractor cloze probability on nine central electrodes based on across-subjects regression. Bars indicate time samples with significant p-values after multiple comparisons correction.

attractive alternative interpretation. The anomaly is then detected by a second, algorithmic stream, and it is the mismatch between the analyses of the semantic stream and the algorithmic stream which produces an increase in P600 amplitude. On RI theory, by contrast, the N400 is taken to index lexical retrieval. Critically, in our design (see Figure 4.2) – which employs a context manipulation, in which a semantically attractive alternative is either available or not (Condition B vs. C), and target word plausibility is varied across three levels (Condition A $<$ B $<$ C) – the target word is repeated several times in a preceding context paragraph. On the retrieval view of the N400 (Brouwer et al., 2012; Kutas & Federmeier, 2000, 2011; Lau et al., 2009; Lau et al., 2008; van Berkum, 2009, 2010), this is predicted to maximally facilitate retrieval of target word meaning and thus minimise N400 differences between conditions. In sum, RI theory predicts no N400 differences between conditions,

and increasing P600 amplitudes as a function of decreasing target word plausibility. Multi-stream models predict a P600 effect, but no N400 effect, if a semantically attractive alternative interpretation is available (Condition B relative to A) and an N400 effect, but no P600 effect, if no alternative interpretation is available (Condition C relative to A).

We validated the design in a self-paced reading experiment (Experiment 4) that revealed a graded sensitivity of reading times to plausibility, indicating that the stimuli indeed induce graded integration effort. Distractor cloze probability did not modulate reading speed significantly. The EEG experiment (Experiment 5), replicated the original findings of Nieuwland and van Berkum (2005), that is, the absence of an N400 effect and the presence of a P600 effect for less plausible relative to plausible target words when the target word is strongly primed by the context and in the presence of a semantically attractive alternative interpretation (our Condition B). Furthermore, our results revealed the *graded* sensitivity of a posterior late positivity to plausibility, as shown by stepped P600 amplitudes for plausible (A), less plausible (B), and implausible (C) items. The absence a plausibility-related N400 effect is inconsistent with an interpretation of the N400 as a graded index of integration difficulty. Additionally, the presence of an expected word which was then not presented elicited an early negativity (250-400 ms) – likely a correlate of lexical mismatch. Further, an rERP analysis revealed that the presence of a strongly expected distractor word – or rather its disconfirmation – resulted in an additional left-frontal positivity in a later time window, in line with previous research. However, in our analyses, the contribution of disconfirmations to late positivities was not statistically significant. In sum, as we discuss in more detail below, these findings reveal a critical novel dimension to the functional interpretation of the P600 that has important implications for existing and future neurocognitive experiments and theories, namely that the P600 is a *continuous* index of integration effort.

4.4.1 The Processing Cost of Disconfirmed Expectations

While the main goal of our design was to manipulate the availability of a semantically attractive alternative interpretation (The lady weighing the suitcase rather than the tourist), how we achieved this manipulation effectively created a prediction disconfirmation in Condition B. That is, when presenting the final sentence fragment "Then weighed the lady the ...", "suitcase" was expected – as shown by high distractor cloze probability – but "tourist" was encountered instead. While not the main focus of our hypotheses, the results are relevant to the literature on disconfirmed predictions.

For Condition B, we observed an early negativity relative to both Condition A and C, lasting approximately from 250 to 400 ms post-stimulus onset. This deflection may relate to the mismatch between the observed word form (target) and the anticipated word form (distractor). Critically, under the retrieval view of the N400

(Brouwer et al., 2012; Kutas & Federmeier, 2000, 2011; Lau et al., 2009; Lau et al., 2008; van Berkum, 2009, 2010), this mismatch does not appear to tax lexical retrieval, as no N400 modulation was observed: The difference between the waveforms disappeared by 400 ms, which would be the typical peak of the N400 component (Federmeier & Laszlo, 2009). This earlier negative component likely overlaps with the N400 in previous studies on disconfirmations and it is the absence of an N400 effect relative to baseline in our data that allows us to observe the earlier negativity in isolation. Results that are directly relevant to ours are presented by Brothers et al. (2015), who observed a centrally peaking N250 for the contrast between a medium-cloze unpredicted versus a medium-cloze predicted target word. Further, in their data, the earlier negativity was not observed for the contrast of a low-cloze unpredicted to a medium-cloze unpredicted target word, which only elicited an N400 effect. Similarly, the visual mismatch negativity has been reported for exactly the time window between 250 ms to 400 ms (Tales et al., 1999). Furthermore, negativities preceding the N400 time window have been found for expectation-incompatible relative to expectation-compatible stimuli (Bartholow et al., 2005), for expectation-based semantic priming (Franklin et al., 2007), and, using pictorial stimuli, for perceptual hypothesis testing which is argued to precede multimodal semantic memory access, as indexed by the following N400 (Kumar et al., 2021).

In the time window from 600 to 1000 ms, our rERP analysis suggests that target words that disconfirmed expected distractor words induced a left-frontal positivity. Distractor cloze probability did, however, not reach significance in the analyses, and hence these results warrant adequate caution. However, previous research has repeatedly reported frontal positivities elicited by prediction disconfirmations (Brothers et al., 2015; DeLong et al., 2014; DeLong et al., 2011; Federmeier et al., 2007; Kuperberg et al., 2020; Kutas, 1993; Quante et al., 2018), making our results relevant to this line of research. A prominent idea has been that if the target is unexpected but plausible, disconfirmations result in a frontally pronounced positivity, whereas implausible replacements result in a parietal positivity (Van Petten & Luka, 2012). We see, however, two open issues with regard to this strict functional segregation of frontal and parietal positivities. First, the apparent distinction between frontally and parietally distributed positivities could be an artefact of spatiotemporal component overlap with the N400 (Brouwer and Crocker, 2017, see also Brouwer, Delogu, and Crocker, 2021; Delogu et al., 2021), and secondly, frontally and parietally distributed positivities may not be mutually exclusive.

A relevant study by DeLong et al. (2014) included plausible, less plausible disconfirming, and implausible disconfirming target words. The design elicited a frontal positivity for less plausible disconfirming words, a parietal positivity for implausible disconfirming words and, critically, N400 effects in response to both less plausible and implausible words, relative to baseline. Our design does not elicit N400 differences and hence circumvents the issue of component overlap, thereby providing a clearer view of the distribution of the late positivities. The estimates

generated by our rERP models (Figure 4.14) suggest that even without a strong N400 overlapping with the late positivity, unfulfilled expectations create an additional positivity with a left-frontocentral distribution. Further, in the disconfirming condition (B), the context additionally made the target word less plausible compared to the baseline condition. Our rERP analysis revealed that for Condition B, plausibility induces a parietal P600 – which was not observed in the data of DeLong et al. (2014) – in addition to the frontal positivity elicited by the disconfirmation. In sum, our results and the rERP analysis suggest that disconfirmations indeed induce a frontal positivity, but that this frontal positivity can co-occur with a plausibility-related parietal positivity on less plausible, but ultimately possible target words.

4.4.2 Global Revision on the Multi-Stream Account

The main goal of this study was to test the hypotheses of multi-stream models against those of RI theory. Multi-stream models were originally proposed in response to studies eliciting semantic P600s, in which semantic anomalies did not elicit N400 effects but rather P600 effects, relative to baseline. Multi-stream accounts explain some of the original data points, by postulating that the semantic stream does not detect the anomaly because a semantically attractive alternative interpretation is available. For instance, in order to "repair" the sentence "the hearty meal was devouring", the inflexion of the verb could be changed to "devoured", yielding a plausible interpretation. However, the surface structure of the sentence does not match this interpretation, which is detected by the algorithmic stream and the conflict between the two streams leads to a P600 effect when compared to a congruous condition.

This explanation was based on a *locally* available alternative interpretation (see Figure 4.2). However, no such *local* availability is given in the design of Nieuwland and van Berkum (2005, "Next, the lady told the tourist/suitcase), and, accordingly, an N400 and no P600 effect relative to baseline would be predicted by multi-stream models. However, the reverse pattern was observed. To account for this, multi-stream may invoke a *globally* attractive alternative interpretation (see Bornkessel-Schlesewsky & Schlesewsky, 2008; Kuperberg, 2007, for discussion). That is, making use of the *globally* available information, the word "suitcase" could be replaced with the discourse-salient word "tourist" in order to arrive at a plausible interpretation in the semantic stream. Again, the analysis generated by the algorithmic stream conflicts with the analysis of the semantic stream, explaining the P600 increase found by Nieuwland and van Berkum (2005). Importantly, it follows that if neither a *locally* nor a *globally* available alternative interpretation is present, an N400 effect should be observed relative to baseline.

The current study extended the original design by Nieuwland and van Berkum (2005) to test this prediction. In the new context manipulation design, we made an

alternative interpretation available *globally* for a less plausible target word (Condition B: "Next, the lady weighed the tourist"), whereas no alternative interpretation was available for the fully implausible target word (Condition C: "Next, the lady signed the tourist"). Assuming the *globally* active plausibility heuristic described above, multi-stream models predict only a P600 effect for Condition B and only an N400 effect for Condition C relative to Condition A. Note that multi-stream models, in general, predict either an N400 or a P600, which makes biphasic N400-P600 results problematic for most multi-stream accounts (see Van Petten and Luka, 2012, for an overview, Brouwer et al., 2012, for discussion, and Bornkessel-Schlesewsky and Schlesewsky, 2008; Li and Ettinger, 2023, for exceptions).

In Condition B, for which only a P600 is predicted by multi-stream accounts, we found a P600 effect relative to Condition A. This condition replicates the results of Nieuwland and van Berkum (2005), and, accordingly, multi-stream models can only explain this P600 effect by invoking a *globally* available alternative interpretation. In Condition C, for which only an N400 effect is predicted by multi-stream accounts, we observed only a P600 effect, relative to Condition A. Critically, the absence of an N400 effect relative to baseline when any semantically attractive alternative interpretation is removed provides strong evidence against multi-stream accounts. One explanation of the absence of the N400 effect in Condition C relative to A would be to assume that the revision process changed the context of Condition C ("Then *signed* the lady the") to make the target word ("tourist") plausible. It is difficult, however, to imagine a mechanism that could revise the context in such a way, while at the same time predicting the presence of N400 effects in cases of canonical semantic incongruencies (see Van Petten & Luka, 2012). Another explanation would entail misunderstanding "tourist" for something contextually relevant, such as the "tourist's ticket". Many of our stimuli, however, contain strong selectional restriction violations, such as "the apprentice ate the hammer" (see Appendix B.2), where reference transfer to a thus far unnamed entity seems unlikely, and hence this explanation cannot account for the complete absence of an N400 effect of Condition C relative to A. Again, it is difficult to see how such an account would predict the absence of an N400 effect for the present stimuli, while at the same predicting the presence of an N400 effect for canonical semantic incongruencies. In sum, we do not see how the present data can be reconciled with the mechanisms assumed by multi-stream accounts.

4.4.3 Retrieval Facilitation under Repetition Priming

The current design had the goal of maximally priming the target word by mentioning it repeatedly in a context paragraph preceding the final sentence. RI theory predicted that maximal priming should maximally facilitate retrieval of the target word's meaning from long-term memory, thus leading to equal N400 amplitudes across conditions. Our results revealed that while an earlier negativity was present

in Condition B relative to A (see above), no difference in the canonical N400 time window was observed for any condition contrast - in line with the retrieval view of the N400 (Brouwer et al., 2012; Kutas & Federmeier, 2000, 2011; Lau et al., 2009; Lau et al., 2008; van Berkum, 2009, 2010). This study thus adds to several studies that elicited no N400 differences for target words that were equally strongly or weakly primed by the preceding context (Delogu et al., 2019, 2021; Hoeks et al., 2004; A. Kim & Osterhout, 2005; Kos et al., 2010; Kuperberg, 2007; Nieuwland & van Berkum, 2005; Otten & van Berkum, 2008; van Herten et al., 2005).

Critically, our results show that even when the target word is of intermediate plausibility (Condition B) or entirely implausible (Condition C), no N400 increase is produced – a result that is at odds with the traditional interpretation of the N400 as semantic integration (Brown & Hagoort, 1993, 2000; Hagoort et al., 2004). Further, also when assuming a hybrid view of the N400 that takes the N400 to index both retrieval and aspects of integrative processing (see Baggio and Hagoort, 2011, who refer to this as "unification", and Baggio, 2018, for an updated account), we would expect to find N400 modulations for the less plausible or implausible target words even when their word meaning is strongly and equally primed - a prediction which was not confirmed. That is, even though retrieval may be facilitated, these accounts should still predict increased integration effort to be reflected in the N400. Thus, for hybrid models to predict the absence of any N400 effect of implausibility, they must still assume that retrieval processes dominate integration/unification. While it may be possible to construct such a hybrid account, the data are more parsimoniously explained by a retrieval-only account, and we are unaware of any other findings that necessitate the inclusion of an integration mechanism. Moreover, it is difficult to see how such an account can explain the absence of an N400 effect of implausibility when target words are equally unassociated to the context (Delogu et al., 2021). Another proposal by Nieuwland et al. (2020) suggests that the earlier part of the N400 is sensitive to retrieval processes, while the later part indexes integration. Critically, however, we did not observe any N400 differences in either the earlier or later part of this component, thereby also ruling out this proposition. On a final note, the absence of N400 modulations by plausibility supports the view that the correlation between corpus-based word surprisal and the N400 may be best explained by expectation-based modulations of lexical retrieval rather than integration (see Frank et al., 2015, and Chapter 3 for discussion).

4.4.4 The P600 as a Graded Index of Integration Effort

Most strikingly, our ERP data revealed an important novel dimension of the P600 component: Our design manipulated plausibility on three levels (plausible, less plausible, implausible) and revealed that P600 amplitude patterns with plausibility. Going beyond the three discrete levels of plausibility, we successfully modelled

the ERP signal as a continuous function of numeric per-item plausibility ratings collected in a pre-test, indicating that the P600 may indeed be a continuous index of integration effort. We conclude that P600s are not only elicited by highly implausible, impossible, or violating target words (Bornkessel-Schlesewsky et al., 2011; Kuperberg, 2007) but rather, that P600 amplitude is modulated as a function of integration effort by every word.

Our proposition that the P600 is a continuous index of integration effort is indeed supported by numerous previous studies showing P600 effects for non-violating but semantically or pragmatically taxing continuations (Burkhardt, 2006, 2007; Cohn & Kutas, 2015; Delogu et al., 2019; Dimitrova et al., 2012; Hoeks et al., 2013; Regel et al., 2010; Schumacher, 2011; Spotorno et al., 2013; Xu & Zhou, 2016). For instance, a world knowledge implausibility without a violation of selectional restrictions induced a P600 effect relative to control (Delogu et al., 2019). The graded nature of the P600 was also suggested by a post-hoc analysis conducted in Chapter 3. By analysing the data of the baseline condition only ("Yesterday sharpened the lumberjack [...] the axe", translated from German), it was found that not only the N400 but also the P600 varied gradually as a function of target word expectancy. This observation was interpreted as indicating a gradual modulation of lexical retrieval (N400) and integration (P600) by expectancy. Hence, the current study directly supports this exploratory, post-hoc analysis with regard to the P600 component.

In Experiment 4, the observed reading times closely matched the P600 in that both were modulated by plausibility across the three levels of our manipulation. Taken together with the absence of N400 modulations by plausibility, this strengthens the proposed link between reading times and the P600 through comprehension-centric surprisal (Brouwer, Delogu, Venhuizen, & Crocker, 2021). To further test this idea, we conduct a post-hoc analysis, in which we apply the rERP technique to model the ERPs obtained in Experiment 5 by the reading times obtained on the Post-spillover region in Experiment 4 (averaged per item). The resulting coefficients (Figure 4.16) suggest that indeed, the observed positivities are correlated to the observed reading times, suggesting they may be closely associated indices of processing effort across pools of participants. This finding further corroborates the P600 as a continuous index of integration effort.

4.5 Conclusion

Event-related potentials provide a multi-dimensional window into the nature and time course of language comprehension. Critically, establishing the locus of specific sub-processes of comprehension in the ERP signal has direct consequences for our understanding of the temporal organisation and architecture of the comprehension system. The present study directly tested competing views on whether the N400 or the P600 component of the ERP signal indexes the integration of incoming word

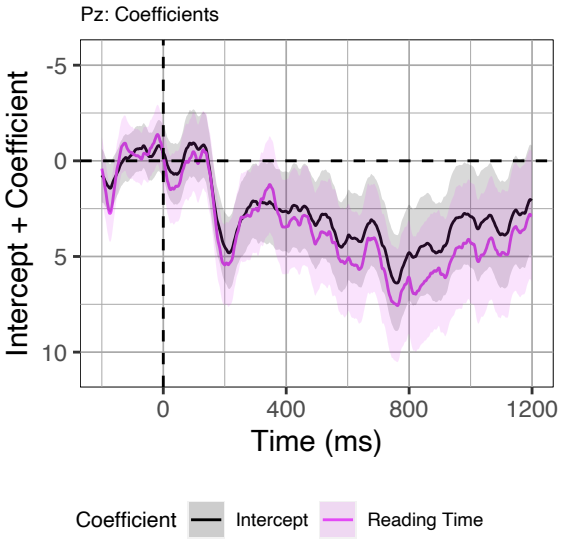


FIGURE 4.16: Regression model coefficients (added to their intercept) across time on electrode Pz from models predicting the ERPs as a function of the per-item reading times obtained on the Post-spillover region in Experiment 4. Error ribbons indicate the standard error on the coefficients in the statistical models.

meaning into an unfolding utterance representation. Crucially, the traditional view of the N400 as an index of integration relies on the presence of a semantically attractive alternative interpretation to explain the absence of an N400 effect in response to certain semantic anomalies. The more recent view of the P600 as an index of integration, in turn, predicts P600 amplitude to be a continuous index of integration effort, a prediction that had yet to be confirmed. We harnessed these predictions to decide between the competing views using a design in which a semantically attractive alternative is either available or not, and target word plausibility is varied across three levels. Further, to minimise lexical processing differences across conditions, target words were equally primed by the prior context.

An initial self-paced reading study revealed a gradual slow-down of reading times for gradual decreases in target word plausibility, suggesting differential integration effort. In the ERP study, the plausibility manipulation did not elicit any N400 differences between conditions. Indeed, the lack of an increased N400 for the implausible conditions – even when no semantically attractive alternative interpretation is available – is directly at odds with the prediction made by contemporary models that maintain the N400 as an index of semantics-driven, quasi-compositional integration. In fact, the plausibility manipulation rather revealed P600 amplitude to be graded for plausibility. Taken together, these results cannot be reconciled with the N400 as an index of integration, while they are consistent with the P600 as a continuous index of integrative effort. More generally, the results are consistent with

Retrieval-Integration theory, a single-stream account in which the N400 indexes lexical retrieval from long-term memory and the P600 indexes integration of incoming word meaning into an unfolding utterance representation. No N400 differences were found, as lexical retrieval was equally facilitated across conditions through repetition priming, and the link between plausibility, reading times, and P600 amplitude establishes the P600 as a direct index of semantic integration that – in line with a comprehension-centric notion of surprisal – is continuous in amplitude as a function of integration effort. This novel dimension of the P600 has important implications for existing and future experiments, as well as for theories and models of language comprehension.

4.6 Acknowledgements

We would like to thank Mante Nieuwland for providing the materials of the Nieuwland and van Berkum (2005) study. We also thank Lea Müller-Kirchen for her help in creating the materials and conducting the ERP study.

Chapter 5

Single-Trial Neurodynamics Reveal N400 and P600 Coupling in Language Comprehension

The contents of this chapter were published in a peer-reviewed journal article (Aurnhammer, Crocker, & Brouwer, 2023).

5.1 Introduction

The two most commonly observed components of the event-related potential (ERP) signature of language comprehension are the N400 and the P600. While the N400 has traditionally been interpreted as an index of integrative-semantic processing (Brown & Hagoort, 1993, 2000; Hagoort et al., 2004), the P600 was first discussed in relation to syntactic and structural processing (Hagoort et al., 1993; Osterhout & Holcomb, 1992). Later studies challenged this functional distinction by eliciting semantic P600s for manipulations in which thematic roles are reversed ("the javelin has the athletes thrown" relative to "the javelin was by the athletes thrown", Hoeks et al., 2004, translated from Dutch) or grammatical inflexions lead to implausible interpretations ("the hearty meal was devouring/devoured", A. Kim and Osterhout, 2005; see Bornkessel-Schlesewsky and Schlesewsky, 2008; Brouwer et al., 2012; Kuperberg, 2007, for reviews). Since then, theories of the electrophysiology of language processing are faced with the challenge to offer a unifying account of the mechanisms underlying the N400 and the P600 that can explain the sensitivities of both components.[1] Specifically, Semantic P600 data gave rise to two alternative views on the language comprehension architecture: Multi-stream models (Bornkessel-Schlesewsky & Schlesewsky, 2008; A. Kim & Osterhout, 2005; Kos et al., 2010; Kuperberg, 2007; Michalon & Baggio, 2019; van Herten et al., 2005) and Retrieval-Integration theory (Brouwer et al., 2017; Brouwer et al., 2012), a single-stream model. Importantly, these theories are mostly informed by the binary presence and absence of N400 and P600

[1] While there are numerous other theories about the N400 component and the P600 component in isolation (see Delogu et al., 2019, for discussion), we here seek to investigate the single-trial dynamics of the N400 and the P600, and therefore focus on *integrated* theories of the N400 and the P600 only.

effects which are typically assessed by comparing mean amplitude across trials in a predefined time window, such as 300 ms – 500 ms post-stimulus onset for the N400 and 600 ms – 1000 ms for the P600. A problem with this approach is that competing theoretical accounts may explain the same ERP data while assuming fundamentally different mechanisms. We here argue that important dissociations between competing effect-level explanations can be achieved by spelling out how different models envision the N400 and the P600 effect, observed between per-condition averages, to arise from language processing in single trials. Consequently, predictions derived from these single-trial level proposals can be investigated empirically in single-trial ERP data. In particular, we here demonstrate that by specifying predictions at the single-trial level, we can test two competing explanations of biphasic N400-P600 effects, offered by multi-stream models and Retrieval-Integration theory, respectively.

5.1.1 Explaining N400 and P600 Effects: Multi-Stream vs. Single-Stream Accounts

Multi-stream models were developed in order to reconcile the integration view of the N400 with the absence of N400 effects and the presence of P600 effects in Semantic P600 studies by postulating that language processing makes use of two processing streams (but see Kuperberg, 2007, for an account with three streams). While the precise conceptualisation of the different processing streams varies across multi-stream models, they share several critical elements: Typically, a *semantic* processing stream employs a plausibility heuristic that constructs an utterance meaning representation based on the content words of the input, while ignoring syntactic constraints (see Bornkessel-Schlesewsky and Schlesewsky, 2008; A. Kim and Osterhout, 2005; Kos, Vosse, Van Den Brink, and Hagoort, 2010; Kuperberg, 2007; van Herten, Kolk, and Chwilla, 2005, for a more detailed discussion, and see Li and Ettinger, 2023; Michalon and Baggio, 2019; Rabovsky et al., 2018; Ryskin et al., 2021, for more recent models with a similar processing mechanism). Critically, for some experimental conditions, no increase in N400 amplitude is taken to occur if the content words make a plausible alternative interpretation available, e.g., by ignoring word order in role-reversed input and assuming the most probable interpretation instead (e.g., interpreting "the javelin has the athletes thrown" as "the javelin was by the athletes thrown"; Hoeks et al., 2004). The *algorithmic* processing stream, however, does adhere to morphological, syntactic, as well as structural constraints and detects the anomaly in the input (Bornkessel-Schlesewsky and Schlesewsky, 2008; A. Kim and Osterhout, 2005; Kos et al., 2010; Kuperberg, 2007; van Herten et al., 2005; see also Li and Ettinger, 2023; Michalon and Baggio, 2019; Rabovsky and McClelland, 2020; Ryskin et al., 2021, for more recent examples). According to multi-stream models, it is the conflict between the analyses generated by the semantic (the athletes threw the javelin) and the algorithmic processing stream (the javelin threw the athletes) that gives rise to the increase in P600 amplitude (see Figure 5.1, right). However, if

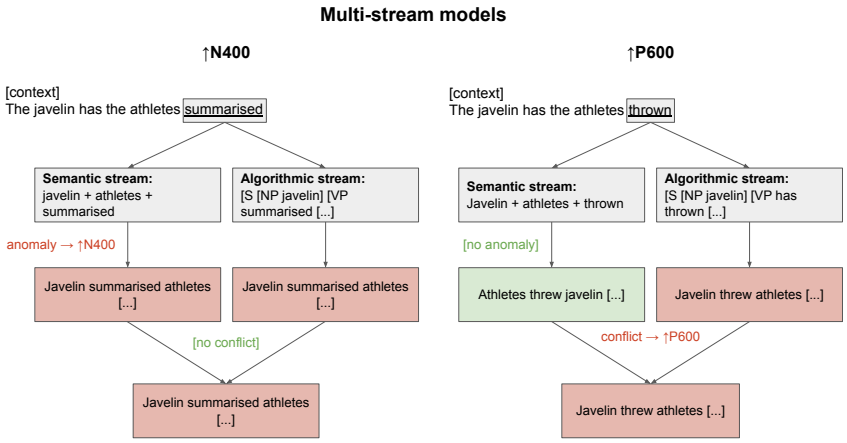


FIGURE 5.1: Schematic overview of the multi-stream explanation of N400 and P600 increases. Stimuli are examples from two conditions in Hoeks et al. (2004). Predicted N400 and P600 increases are specified relative to the baseline condition ("the javelin was by the athletes thrown"). Hoeks et al. (2004) found an N400 effect and a P600 effect for "The javelin has the athletes summarised" and a P600 effect for "The javelin has the athletes thrown", relative to baseline. All examples are transliterated from Dutch.

the anomalous condition does not make a semantically attractive alternative interpretation available ("The javelin has the athletes summarised"), an increase in N400 amplitude is predicted to be produced by the semantic stream, and the two streams agree in their analyses. Hence, there is no conflict and no increase in P600 amplitude is predicted – contra to the findings of Hoeks et al. (2004) who found a biphasic effect (see Figure 5.1, left).

An alternative, single-stream, account of the N400 and the P600 is Retrieval-Integration (RI) theory (Brouwer et al., 2017; Brouwer et al., 2012). On the RI account, the N400 is taken to index lexical retrieval (Kutas & Federmeier, 2000; Lau et al., 2009; Lau et al., 2008; van Berkum, 2009, 2010), i.e., the access of word meaning in long-term memory, and the P600 is posited to index integration, the updating of an utterance meaning representation with the meaning of the current word. RI theory posits that the N400 and the P600 are elicited by every word and that their amplitudes are continuous indices of retrieval effort (N400) and integration effort (P600), respectively. Figure 5.2 depicts a schematic of the computational instantiation of RI theory proposed by (Brouwer, Delogu, Venhuizen, & Crocker, 2021). In this model, the amplitudes of the N400 component and the P600 component are taken to be proportional to the word-by-word change in the **retrieval** and **integration** layers, respectively. According to this model, no N400 *effect* between conditions is observed, if conditions facilitate retrieval equally, and no P600 *effect* between conditions is observed if integration is equally effortful in the conditions. Indeed, this explanation is consistent with the absence of an N400 effect for the sentence "The javelin has the athletes thrown" relative to the baseline "the javelin was by the athletes thrown"

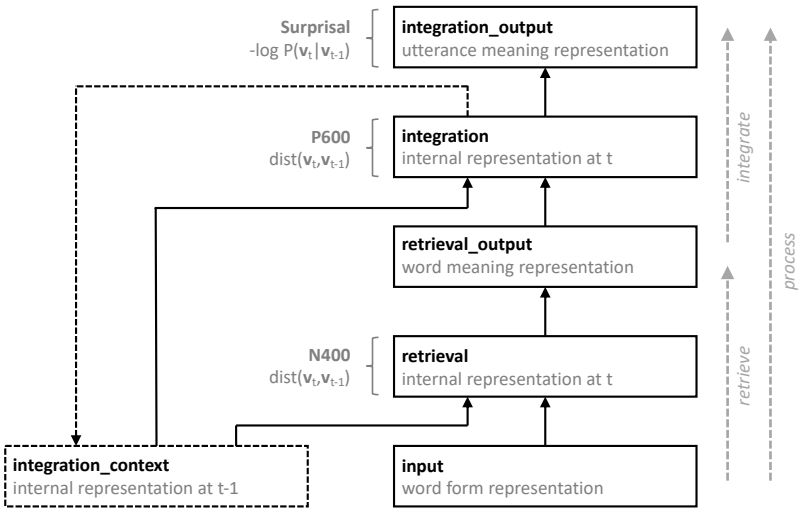


FIGURE 5.2: Schematic architecture of the neurocomputational instantiation of Retrieval-Integration theory, implementing word-by-word language processing and the linkage of retrieval to the N400 and integration to the P600. For full detail on model implementation see Brouwer, Delogu, Venhuizen, and Crocker (2021).

(Hoeks et al., 2004), as the target word is similarly associated to the context in both conditions. The P600 effect is explained by the implausibility of the role-reversed condition relative to the control condition.

5.1.2 Dissociating Effect-Level Explanations at the Single-Trial Level

Multi-stream models were strongly motivated by the monophasic P600 effects and monophasic N400 effects observed in Semantic P600 studies. However, several condition contrasts in these studies also elicited biphasic effects (e.g., Hoeks et al., 2004). Further, recent studies demonstrated that component overlap between the N400 and the P600 can result in the attenuation or absence of P600 effects (Brouwer, Delogu, & Crocker, 2021; Delogu et al., 2019, 2021). Indeed, consulting the empirical evidence, it is striking that language comprehension ERP experiments manipulating semantic congruency (e.g., "He spread the warm bread with socks/butter"; Kutas & Hillyard, 1980) often elicit *biphasic* ERP responses consisting of both an N400 effect and a P600 effect relative to baseline (see Van Petten & Luka, 2012, for an overview). For instance, the ERP experiment of Chapter 3 manipulated the expectancy of the target word ("Yesterday, sharpened the lumberjack [...] the axe" vs. "Yesterday ate the lumberjack [...] the axe"; transliterated from German, see Table 5.1). The Unexpected condition elicited both a more negative N400 amplitude and a more positive P600 amplitude, relative to the Expected baseline condition (Figure 5.3).

Expected	Yesterday sharpened the lumberjack [...] the axe and ...
Unexpected	Yesterday ate the lumberjack [...] the axe and ...

TABLE 5.1: Example item, showing the expectancy manipulation of Chapter 3, achieved by violating the selectional restrictions of the main verb ("sharpened/ate"). Stimuli are transliterated from German, preserving word order. Target words were underlined for this table. The original design also manipulated lexical association of the target word to the words in an adverbial clause preceding the target word ("before he the [wood stacked/movie watched]"). In the two conditions shown here, the adverbial clauses are identical, associated, and omitted in the table.

Semantic P600 studies employed experimental designs that maximised the presence/absence of semantic attraction, which on multi-stream accounts should determine the presence/absence of P600/N400 increases. Because of this, the biphasic effect patterns observed in some Semantic P600 studies (e.g., Hoeks et al., 2004) have been discussed as difficult to reconcile with multi-stream models (Brouwer et al., 2012). However, in cases of canonical semantic incongruities, one possible multi-stream explanation of biphasic effects could be that the N400 and P600 effects observed in the *averages* derive from trial-specific N400-only and P600-only elicitations. That is, in the case of canonical semantic incongruities, it is not always clear whether all experimental items exclude the presence of a semantically attractive alternative interpretation for the incongruent items, especially if a broad notion of global semantic attraction is adopted (see Bornkessel-Schlesewsky and Schlesewsky, 2008; Kuperberg, 2007, and Chapter 4, for discussion). Hence, it would be conceivable that for one subset of the unexpected trials, there was no semantically attractive alternative interpretation and the unexpected target word was detected in the semantic stream, which resulted in an N400 increase. However, for the remaining unexpected trials, there may have been semantic attraction, and the input may have been judged as plausible in the semantic stream, leading to no N400 increase. In the algorithmic stream, these trials would however result in an implausible analysis and the conflict between the analyses of the semantic and the algorithmic stream should induce a P600 increase (see Figure 5.1). Averaging over these two subsets of anomalous trials may result in the biphasic condition contrast in which the average N400 is more negative and the average P600 is more positive in the incongruent condition than in the baseline.

Crucially, and in contrast to multi-stream accounts, RI theory predicts both an N400 and a P600 increase on the same trials: On RI theory, both the mapping of word forms to word meanings (retrieval) and the mapping of word meanings into an updated utterance meaning representation (integration) are constrained by utterance context – the utterance meaning constructed so far (see Figure 5.2). Hence, for the processing of a single word, the single-stream architecture makes a fundamental

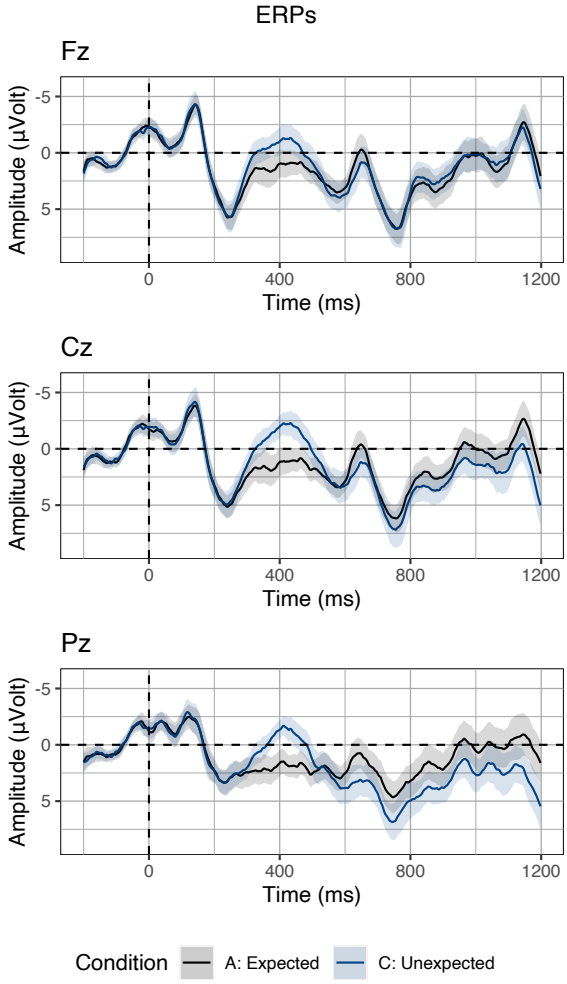


FIGURE 5.3: Grand-average ERPs on three midline electrodes (Fz, Cz, Pz) for two conditions of Chapter 3 that manipulated target word expectancy. Waveforms were averaged per condition from the per-subject, per-condition averages. Error ribbons indicate confidence intervals based on standard errors computed across subjects.

prediction: Due to the shared dependency of retrieval and integration on the utterance meaning constructed so far, words that are more effortful to retrieve should also be more effortful to integrate. Consequently, N400 amplitude and P600 amplitude should be negatively correlated. This prediction is supported by quantitative model estimates generated by the computational instantiation of RI theory (Brouwer, Delogu, Venhuizen, & Crocker, 2021) for a recent ERP study by Delogu et al. (2019). Comparing the N400 and P600 estimates generated by this model for all words in the stimuli (i.e., not just the target words), we indeed find a negative correlation (r = -0.62). The model estimates thus confirm the prediction of RI theory that words with a more negative N400 amplitude should, generally, also induce a more positive P600 amplitude.

Thus, multi-stream models and RI theory account for biphasic effect patterns, by assuming very different processing architectures that, critically, make opposing

predictions for the modulation of the N400 and the P600 within-trial: On multi-stream accounts, the presence of an N400 increase for an anomalous trial predicts the absence of a P600 increase and, vice versa, the absence of an N400 increase for an anomalous trial predicts the presence of a P600 increase. In contrast, RI theory predicts that N400 amplitude and P600 amplitude should be negatively correlated in that more negative N400 amplitudes should co-occur with more positive P600 amplitudes. To investigate the single-trial dynamics of the N400 and the P600, we re-analyse the ERP data presented in Chapter 3. While the full experiment crossed expectancy with lexical association, we here focus only on the expectancy manipulation. Indeed, both multi-stream models and RI theory offer a possible explanation for this expectancy manipulation at the effect level. We will examine here, however, whether they can also account for the data at the single-trial level.

Quantitatively, the prediction of multi-stream models can be expressed, slightly unintuitively, by a *positive* correlation between N400 amplitude and P600 amplitude relative to the grand average of two conditions with a biphasic effect (cf. Figure 5.3). That is, on multi-stream accounts, a trial that results in an increase in N400 amplitude, should not result in an increase in P600 amplitude. As a consequence, P600 amplitude should be more negative than the grand average in this case. Conversely, a trial that results in an increase in P600 amplitude, should not result in an increase in N400 amplitude. Hence, in this case, N400 amplitude should be more positive than the grand average. Taken together, this predicted pattern thus results in a *positive* correlation between N400 amplitude and P600 amplitude at the single-trial level: If P600 amplitude becomes more positive, N400 amplitude should not diverge from baseline, and hence be more positive than the grand average, and vice versa.

The prediction of RI theory, on the other hand, can be expressed by a *negative* correlation between N400 and P600 amplitudes at the single-trial level. That is, RI theory assumes that both retrieval and integration are expectation-based processes: The expectations about upcoming word meaning (retrieval) and utterance meaning (integration) both derive from the utterance meaning representation constructed so far. Hence, it is due to this shared dependency on the unfolding utterance meaning representation that RI theory predicts unexpected words to generally – on a by-trial basis – be more difficult to retrieve and more difficult to integrate. This results in the prediction that there is a negative correlation between N400 amplitude and P600 amplitude, because more negative N400 amplitudes should co-occur with more positive P600 amplitudes, and conversely, more positive N400 amplitudes with more negative P600 amplitudes. After the analysis of ERP data that elicited a biphasic effect (from Chapter 3), we also test the generalisability of the proposed single-trial neurodynamics to ERP data that elicited only monophasic effects between conditions (Delogu et al., 2019).

5.2 Method

Both multi-stream models and RI theory can explain condition contrasts resulting in a biphasic N400-P600 effect. In order to investigate whether, and if so, how the N400 amplitudes of single trials correlate with the P600 amplitudes of the same trials, we re-analyse the data in the Expected and Unexpected conditions of Chapter 3 (Table 5.1; Figure, 5.3). Our analyses focus on three midline electrodes, as we did not observe hemispheric differences in the topography of the N400 effect and the P600 effect. In the electroencephalography (EEG) experiment, 120 items were presented to 40 participants. After artefact rejection, 2027 trials remained in the subset of the Expected and Unexpected conditions. Sentences were presented using rapid serial visual presentation, whereby individual words were presented centrally on the screen for 350 ms with a 150 ms inter-stimulus interval. After presentation of each sentence, participants were instructed to provide a binary plausibility judgement. The EEG was re-referenced offline to the average of the left and right mastoid electrodes and band-pass filtered between 0.01 and 30 Hz. Data were baseline corrected using a 200 ms pre-stimulus interval. For full detail on experimental design, electrophysiological recording and processing, refer to Chapter 3.

5.2.1 Towards Single-Trial Dynamics: Naive Binning-Based Approach

An initial approach to investigate the interrelation of N400 and P600 amplitudes would be to compare their raw amplitudes. Computing their correlation, we find that, in fact, single-trial N400 amplitudes (300 ms - 500 ms) and P600 amplitudes (600 ms - 1000 ms) are positively correlated ($r = 0.67$; correlation computed for electrode Pz, where both the N400 effect and P600 effect were maximal; see Chapter 3, for topographic maps). That is, trials with more negative N400 amplitudes also appear to exhibit more negative P600 amplitudes and vice versa. At face value, this supports the multi-stream explanation rather than RI theory. To validate whether this positive correlation between the amplitudes in the two time windows is indeed specific to the ERP components of interest, we compute per-trial averages in the N400 time window in order to split the data into three equal-sized bins. This binning is then applied to visualise the entire waveforms. With regard to the predictions, we then examine whether the bins derived from the N400 time window also induce an ordering in the P600 time window. The resulting bins for electrode Pz, on which both the N400 effect and the P600 effect were maximal in the original experiment, are displayed in Figure 5.4. While some of the typical peaks and troughs of visually elicited language ERPs are visible in the bins, it is striking that the bin-averaged waveforms diverge immediately after stimulus onset, i.e., the point from which baseline correction takes effect. This immediate divergence of the bins before the N400 time window casts doubt on the idea that the correlation between the single-trial averages in the N400 and the P600 time window, as visualised by the three bins, captures

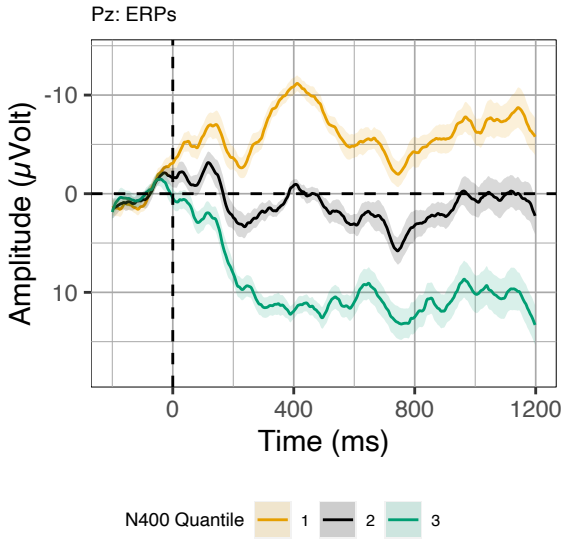


FIGURE 5.4: EEG signals binned by N400 averages (300 ms - 500 ms) in the Expected and Unexpected conditions of Chapter 3 on electrode Pz. Error ribbons indicate confidence intervals based on standard errors computed across quantiles.

(only) systematic N400 variability.

To understand how these bins arise, it is useful to consider the kinds of noise and variability present in single-trial EEG data (Figure 5.5). Overall, single-trial EEG signals are characterised by a low signal-to-noise ratio. Unsystematic variation, i.e., variability not elicited by the stimulus, comes in the form of random noise, periodic signals, such as alpha waves, or as monotonous voltage drifts that are becoming more negative or positive over time (highlighted by regression lines in Figure 5.5). Single-trial N400 time window averages will thus be driven by unsystematic variability (such as voltage drifts) to a much larger extent than by the underlying N400 amplitude within this trial. The averaging of ERP signals per condition and/or per subject removes drifts from average ERPs if they are occurring randomly, that is, if they do not systematically co-occur with specific conditions and/or subjects. When computing three N400 bins based on the "raw" N400 time window average (Figure 5.4), we are however grouping the data based on a property of the signal itself and, hence, the resulting bins are not independent of the noise. Thus, the bins may be more strongly driven by the overall amplitude magnitude of the signals than by true N400 amplitude. This explanation is supported not only by the immediate divergence of the bin-averaged waveforms after stimulus onset but also by the overall magnitude of the highest and lowest bin (compare the magnitude to the condition averages of Figure 5.3). Importantly, some kinds of noise, such as voltage drifts, are correlated in two consecutive time windows, which could thus alternatively explain a positive correlation of N400 amplitudes and P600 amplitudes. Hence, in order to group the EEG data based on some characteristic of the signal itself, such as the size of the N400 in a single trial, or to compute correlations between consecutive time

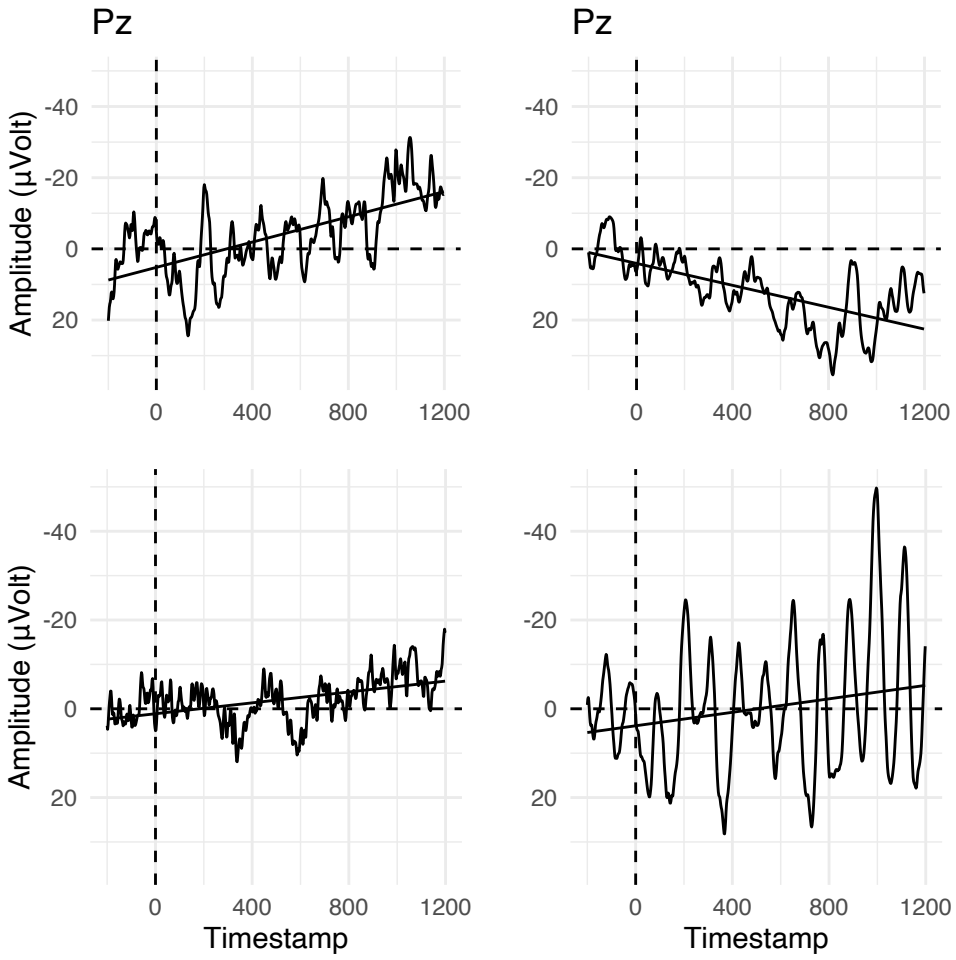


FIGURE 5.5: Four randomly selected single-trial waveforms from the Expected and Unexpected conditions in Chapter 3. Regression lines indicate voltage trends over time.

windows, it is necessary to separate voltage drifts from the systematic N400 modulations.

A naive approach to remove the voltage drift from the binning would be to compute the average of the N400 time window and subtract from it the average voltage on that trial computed from 0 to 1200 ms post-stimulus onset (cf. the traditional procedure for applying baseline correction). Crucially though, these "average N400 minus average Segment" voltages are *only* used to arrive at the bins, which are then used to visualise the *unaltered* data as bin-averaged waveforms (making this approach different from baseline correction). That is, we do not alter the displayed data in any way: the subtraction procedure only affects the assignment of trials to bins. Interestingly, if we apply subtraction-based binning, the resulting average waveforms better resemble typical condition-average ERP waveforms (Figure 5.6). Most strikingly, the waveforms do not diverge immediately after stimulus onset. Rather, it is only around 300 ms (the beginning of the N400 time window) that the waveforms start to diverge, suggesting that the subtraction procedure may indeed have recovered aspects of systematic N400 variability in the single-trial N400

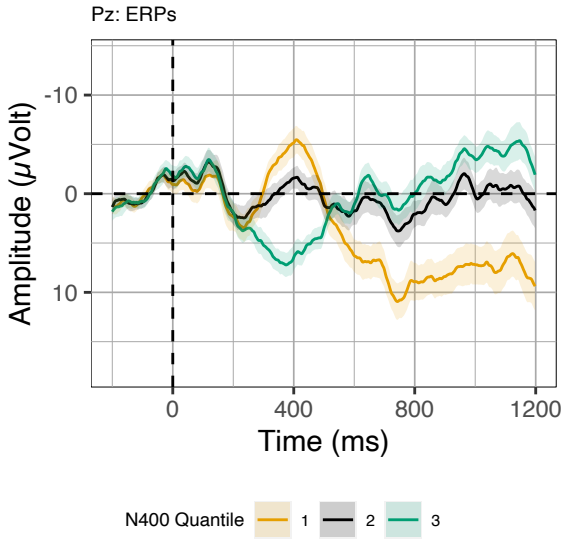


FIGURE 5.6: EEG signals grouped by bins obtained by subtracting average Segment voltage (0 ms - 1200 ms) from average N400 voltage (300 ms - 500 ms) in the Expected and Unexpected conditions of Chapter 3. Error ribbons indicate confidence intervals based on standard errors computed across quantiles.

voltages while removing random voltage drifts. Crucially, moving to the P600 time window (from around 600 ms post-stimulus onset), the ordering of the N400 bins flips. That is, according to the subtraction-based bins, the more negative the N400 amplitude, the more positive is P600 amplitude. The validity of the obtained binning is strengthened by the morphology of the resulting waveforms, which, compared to Figure 5.4, suggest clearer N400 components and P600 components with more typical peaks and latencies, as well as no large differences before them (pre 300 ms). This subtraction-based binning approach can also be related back to quantitative correlations, by computing the partial correlation between N400 amplitudes and P600 amplitudes that accounts for their correlation to Segment voltage: While the raw correlation between N400 amplitude and P600 amplitude was positive, the partial correlation that factors out Segment voltage is negative ($r = -0.49$, correlation computed for electrode Pz).[2] In sum, the naive subtraction-based binning approach suggests that the bin with the largest N400 amplitudes also includes the largest P600s, indicating a negative correlation on a by-trial basis between the two components.

Hence, if voltage drifts are accounted for, the correlation as well as the binning, which are both based on the single-trial EEG data, are incompatible with the explanation of biphasic data that we articulated for multi-stream accounts: The architecture of most multi-stream models suggests that trials which induce more negative N400 amplitudes do not trigger a P600 increase and vice versa. The language processing architecture proposed by Retrieval-Integration theory directly predicts the obtained pattern, because retrieval effort (N400) and integration effort (P600) should

[2]Computed using the `ppcor` package for `R` (S. Kim, 2015).

be correlated negatively at the target word. While these results indeed form initial support for a coupling of the N400 and the P600 at the single-trial level, this naive approach still suffers from shortcomings.

5.2.2 Towards Single-Trial Dynamics: Regression-Based Approach

The subtraction used in the binning process is rather crude and applies the same amount of subtraction (N400 $-$ 1 $*$ Segment) to all time samples. This is inadequate because voltage drifts tend to be directed (see Figure 5.5), i.e., they become more negative or positive over time (see also Hennighausen et al., 1993). Due to this directedness, it would be desirable to apply a variable amount of drift correction across time. Ideally, the optimal amount of voltage correction should be derived from the data itself at each time sample. This can be achieved straightforwardly by casting the research question into the perspective of rERPs (Smith & Kutas, 2015a), a regression-based ERP analysis technique. At the core of the rERP technique lies the observation that fitting a series of intercept-only regression models – one for each subject at each time sample – is mathematically equivalent to computing a grand-average ERP waveform from per-subject average waveforms, meaning that "all ERPs *are* rERPs" (Smith & Kutas, 2015a, p. 158). Building on this, more predictors can be added to the regression equations to model the variability around the mean, and across time samples in the EEG signal. For instance, the original data from both conditions could be modelled using a continuous predictor, such as cloze probability (see the analyses in Chapter 3). Here, however, we are interested in explaining the EEG signals recorded from each subject and at each time sample as a function of that signal itself in order to determine a possible coupling of N400 and P600 amplitude.

Hence, our rERP models[3] include the average N400 voltage (300 - 500 ms) and the average Segment voltage (0 - 1200 ms) as trial-level predictors (see Alday, 2019, for a similar approach to applying baseline correction). We apply the analysis method to three midline electrodes and compute separate N400 and Segment predictors for each electrode. Predictors are z-standardised and inverted. While the inverting results in positive correlations to be expressed by negative coefficients on the N400 predictor (and vice versa), it will aid intuitive ERP-like visualisation of the resulting model coefficients. We arrive at a set of models of the following form:

$$y_{ets} = \beta_{0ets} + \beta_{1ets} * N400_{ets} + \beta_{2ets} * Segment_{ets} + \epsilon_{ets} \tag{5.1}$$

These regression equations – one each electrode e, time sample t, and subject s – model the observed data y by computing estimated data $\hat{y}_{ets}$ (equal to $y_{ets} - \epsilon_{ets}$). The intercept term β_0 will equal the average of the data for the current selection of subject and time sample. As both other predictor terms, N400 and Segment, are

[3]Statistical analyses were implemented in `Julia` (Bezanson et al., 2017).

computed per trial, they allow us to capture any auto-correlations present in the signal. Specifically, the Segment voltage predictor fitted by coefficient β_2 will capture the extent to which the EEG signal, across time samples, is explainable by overall segment magnitude. Seeing that voltage drifts tend to be directed, becoming more positive or negative over time (Figure 5.5), we expect that the Segment voltage coefficient should increase over time. The N400 predictor, on the other hand, captures the extent to which N400 amplitude explains variability in the EEG signal, over and above what is explained by the Segment predictor. The combined presence of both predictors in the models effectively leads to a variable weighting of both predictors over time, which is expressed in the magnitude of the coefficients. Hence, our rERP analysis should be superior to the invariable subtraction-based approach. We expect the N400-average predictor, fitted by coefficient β_1, to be a very good predictor for the N400 time window itself. Our prediction about the N400-P600 single-trial dynamics is addressed by inspecting the coefficients of the N400 predictors in the P600 time window (600 ms - 1000 ms). If a positive correlation exists between N400 and P600 amplitude, as predicted by multi-stream models, the N400 coefficient should extend its trend from the N400 time window into the P600 time window. If, on the other hand, N400 and P600 amplitude are inversely correlated, the N400 coefficient should flip sign when moving from the N400 to the P600 time window and predict more positive amplitudes from around 500 milliseconds post-stimulus onset.

Single-Trial Dynamics Across Conditions

The resulting coefficient graph (Figure 5.7; coefficients are added to the intercept) demonstrates that the Segment predictor becomes active – relative to the intercept – immediately after stimulus onset and, indeed, the coefficient increases over time, suggesting that it captures monotonically increasing and decreasing voltage drifts. Critically, the Segment coefficient drops back to the intercept during the N400 time window, indicating no contribution to explaining the signal. This is simply because the N400 predictor captures both systematic and random variability in its own time window very well. After the N400 time window, the previous trend of the Segment predictor continues. In sum, the coefficients of the Segment predictor indicate that trials that are more negative overall tend to become more negative over the course of the segment and those that are more positive overall become more positive over the course of the segment. The coefficient for the N400 predictor, on the other hand, indicates only a small contribution prior to the N400 time window and the directionality of the coefficient indicates that more negative voltages in the N400 time window predict more negative voltages prior to the N400 time window. In the N400 time window itself, the coefficient of the N400 predictor increases in magnitude. Here, the N400 predictor presumably models both the systematic N400 variability and the Segment drifts present in this time window (cf. the "raw" N400 bins above, Figure 5.4).

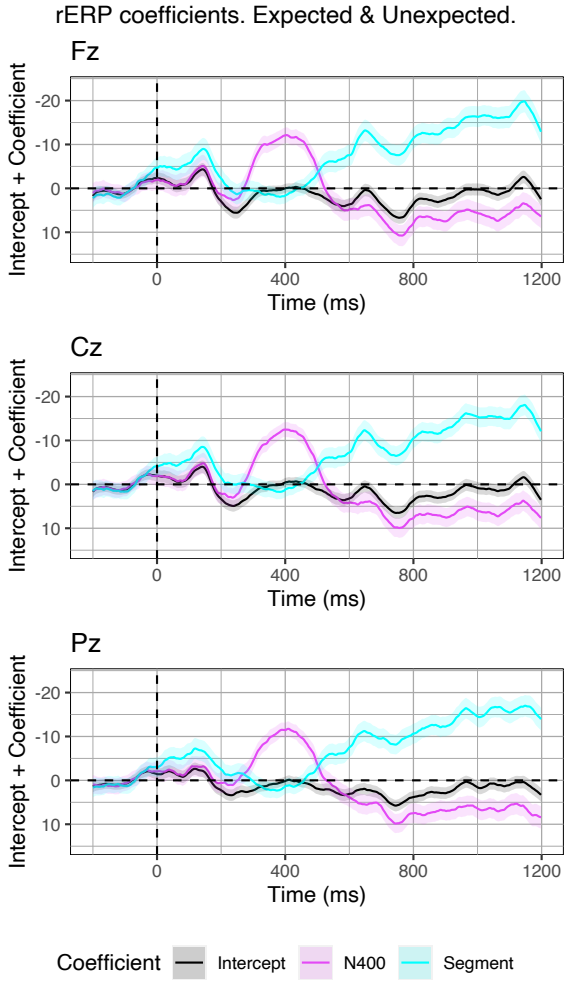


FIGURE 5.7: Model coefficients (added to their intercept) across time on three midline electrodes (Fz, Cz, Pz) for regression models fitted on two conditions of Chapter 3. Coefficients express the extent to which single-trial N400 amplitude (averaged from 300 ms - 500 ms) and Segment amplitude (averaged from 0 ms - 1200 ms) explain the EEG signal across time. Error ribbons indicate standard errors on the coefficients.

The critical aspect of the rERP analysis is the behaviour of the N400 predictor in the P600 time window (600 ms - 1000 ms). Indeed, in the P600 time window, the coefficient of the N400 predictor changes sign, indicating that trials that were more negative in the N400 time window are predicted to become more positive in the P600 time window. Importantly, due to the presence of the Segment predictor, drift-related variability in the N400 predictor is factored out when determining the latter's best-fit coefficient – at least to the extent to which the Segment predictor accounts for the drift-related variability. In sum, the rERP analyses, in which Segment correction is optimised for each subject and time sample, support the initial results derived from the naive subtraction-based binning approach: Trials with more negative N400 amplitudes also induce more positive P600 amplitudes.

Crucially, the rERP approach still suffers from one shortcoming: It is currently

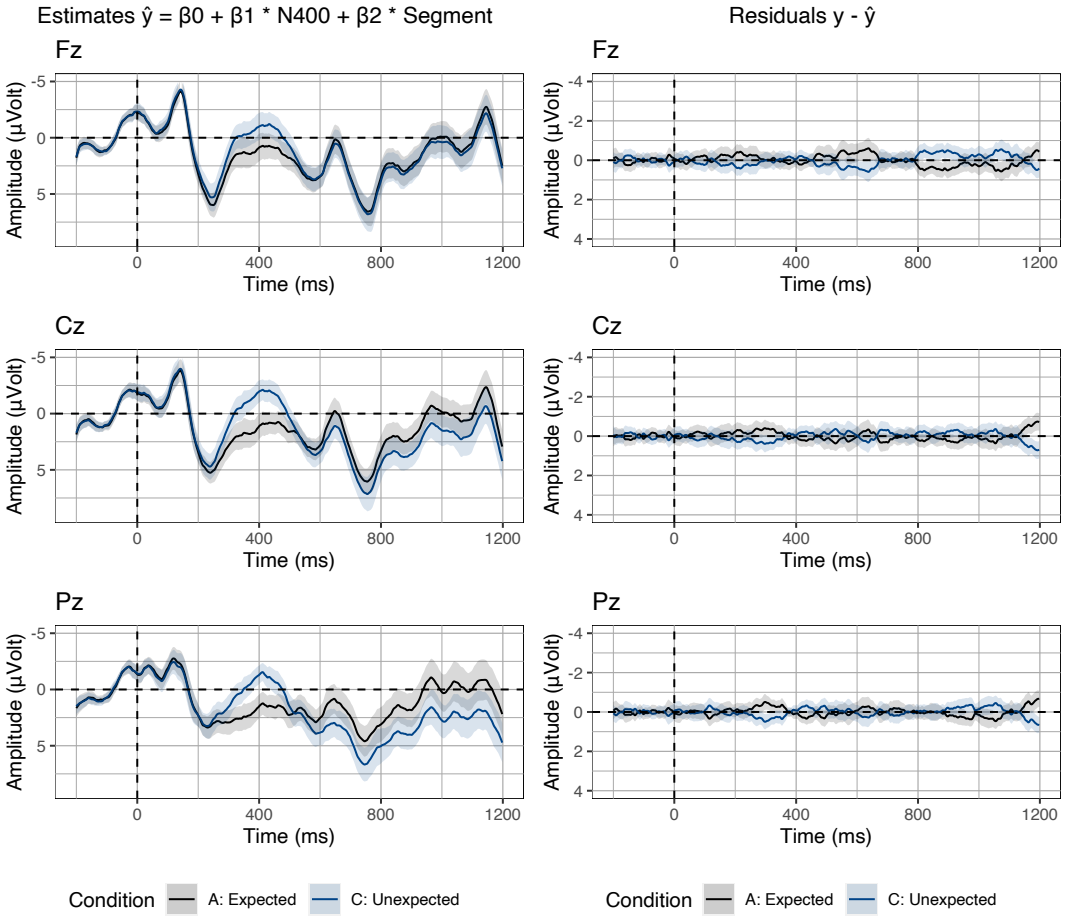


FIGURE 5.8: Forward estimates (left) and residual error (right) on three midline electrodes (Fz, Cz, Pz) from a set of regression models fitted using Equation 5.2.2. Estimates and residuals were split per condition. Error ribbons indicate confidence intervals based on standard errors computed across subjects.

not possible to quantify the extent to which the predictors – which are derived from the signal itself – pick up on N400-P600 dynamics or on noise that is correlated across time windows. In order to clarify this issue, we return to the traditional approach of removing randomness in EEG signals, which is the averaging of noisy single-trial EEG recordings to average ERPs. The intuition is that noise which randomly occurs with the grouping factor used for averaging (e.g., conditions) will be removed in the average ERPs. Inspired by this traditional approach, we evaluate our rERP analysis by measuring the extent to which the N400 and the Segment predictor are able to recover the two conditions underlying the current data (see Figure 5.3, Table 5.1). In order to evaluate the rERP models against the two conditions, we use the regression coefficients to compute the estimates ($\hat{y}$ in Equation 5.2.2) for the entire data set. We then group the estimates by the original two conditions (Figure 5.8, left), in order to determine the extent to which the estimates reproduce the biphasic N400-P600 effect pattern. Indeed, compared to the original two conditions (Figure 5.3), the estimated

data appear to capture both the N400 effect and, more importantly, the P600 effect of the Unexpected relative to the Expected condition.

To quantify the difference between the observed data and the estimated data, we compute the residual error: $y - \hat{y}$ (Figure 5.8, right). We find that the residual error – averaged per time sample and participant and then split up by condition – is close to zero, indicating that the rERP models recover the effect structure of the observed data. Strikingly, while the effect structure clearly differs across the three midline electrodes (e.g., compare the absence of a P600 effect at Fz to the presence of such an effect at Pz), the coefficient graphs look very similar. That is, at all electrodes, the coefficient for the N400 predictor suggests that more negative N400 amplitudes also induce more positive P600 amplitudes. While this may initially appear to be inconsistent, we will later address in detail how a monophasic effect structure can indeed yield the pattern of coefficients such as the one observed at Fz (see Section 5.2.2).

To further decompose the extent to which specifically the N400 predictor – and not the Segment predictor – captures the condition contrast, we compute their isolated estimates and residuals. To do so, we use the models fitted with both predictors present and neutralise the influence of one of the predictors on the forward estimates by setting the predictor values to their average, which is zero for z-standardised predictors. Crucially, this does not involve refitting the models, and thus the coefficients remain unchanged: That is, we only re-estimate data, using the same set of fitted coefficients, while neutralising different predictors.

As the coefficients for the z-standardised predictors adjust the by-trial estimates in terms of their deviation from the grand average, as given by the intercept, a first step is to isolate the contribution of the intercept to the estimates by neutralising *both* the N400 and Segment predictor (see Figure 5.9, row 1):

$$\hat{y}_{ets} = \beta_{0ets} + \beta_{1ets} * 0 + \beta_{2ets} * 0 \tag{5.2}$$

Next, to compute the isolated estimates of the N400 predictor while neutralising the influence of the Segment predictor, we re-estimate the data using the following equation (see Figure 9, row 2):

$$\hat{y}_{ets} = \beta_{0ets} + \beta_{1ets} * N400_{ets} + \beta_{2ets} * 0 \tag{5.3}$$

Conversely, to isolate the contribution of the Segment predictor, we neutralise the influence of the N400 predictor, as shown in the following equation (see Figure 5.9, row 3):

$$\hat{y}_{ets} = \beta_{0ets} + \beta_{1ets} * 0 + \beta_{2ets} * Segment_{ets} \tag{5.4}$$

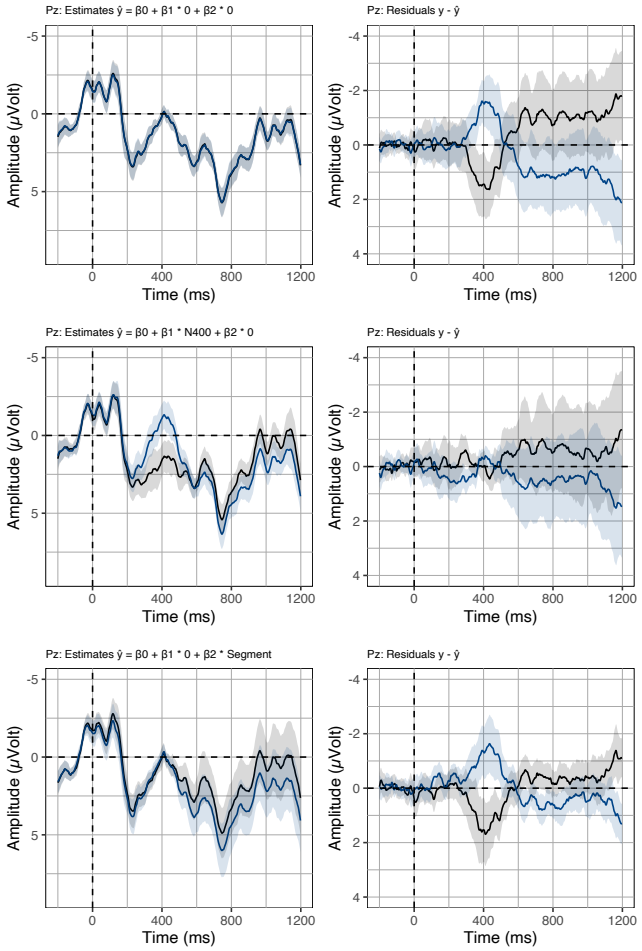


FIGURE 5.9: Isolated forward estimates (left) and residual error (right) computed from rERP models fitted with all predictors present (Equation 5.2.2). Estimates and residuals were split per condition. Rows contain the isolated estimates and residuals of the intercept (row 1), the intercept plus the N400 predictor (row 2), and the intercept plus the Segment predictor (row 3). Error ribbons indicate confidence intervals based on standard errors computed across subjects.

Plotting the isolated estimates and their residual error (Figure 5.9), we first find that, trivially, the intercept, which is equal to the average of the data in our models, is a good model of the conditions pre-N400, but does not accurately capture the difference between conditions in the N400 and the P600 time window (top row). Adding the N400 predictor to the computation of the forward estimates reveals that, indeed, N400 amplitudes allow us to not only model the N400 effect, but also reduce the residuals in the P600 time window, indicating that, indeed, N400 amplitudes are predictive of P600 amplitudes (middle row). Lastly, turning to the Segment predictor, we find that, in fact, Segment voltage also models part of the P600 effect in the data (bottom row). This is most likely the case because the P600 is a long, sustained

component and hence the Segment predictor also contains systematic P600 variability. Indeed, we find that single-trial P600 voltages (600 - 1000 ms) and single-trial Segment voltages (0 ms - 1200 ms) are strongly correlated (r = 0.94). Despite the fact that the P600 effect is in part modelled by Segment voltage, the isolated estimates of the N400 predictor (Figure 5.9, middle row) reveal a unique contribution in explaining P600 variability, over and above what is accounted for by the Segment predictor.

Single-Trial Dynamics Within-Condition

While the estimates reveal that our rERP models successfully account for the ERPs at the condition level, it is an open question to what extent the observed N400-P600 interrelation is driven by the Unexpected and the Expected condition, respectively. Indeed, in Chapter 3, we also conducted a post-hoc analysis that explored whether the graded expectancy of the target word in the Expected condition (Cloze probability: mean = 0.67, SD = 0.23, range = 0.17 - 1) also induced graded retrieval effort (N400) and integration effort (P600). An rERP analysis in which the EEG was modelled as a function of log-Cloze probability suggested that, indeed, not only N400 amplitude but also P600 amplitude was continuously related to target word expectancy. Hence, neither the N400 nor the P600 responses appear to be specifically elicited by the violation of the main verb's selectional restrictions that was employed in the Unexpected condition, which is in line with the assumption of RI theory that both components are continuous indices of processing effort. In the current analyses, we would thus expect that, similarly, the correlation of N400 amplitude and P600 amplitude should also be observable in the Expected condition alone and not be driven by the Unexpected items alone. Thus, in order to validate that the N400-P600 interrelations we found are not qualitatively different in the Expected and Unexpected condition, we also fit the rERP models for the two conditions separately, using the same regression equations as above. While it is now not possible to validate model fit against the effects observed between conditions (cf. Figure 5.9), the model coefficients for the regressions that were fitted on the two conditions separately do not suggest any qualitative differences between the conditions on midline electrodes (see Figure 5.10, for the coefficients at electrode Pz).

Hence, our novel approach reinforces that both the N400 (Kutas & Hillyard, 1984) and the P600 (Chapter 4) are continuous indices of processing effort but importantly go beyond these earlier findings by also suggesting that the two ERP components are negatively correlated at the single-trial level both for well-formed sentence completions (Expected condition) and violating target words (Unexpected condition). However, while we found evidence for negatively correlated N400 and P600 amplitudes in a design that resulted in a biphasic effect between conditions, an open question is how these proposed within-trial dynamics can be reconciled with ERPs that exhibit only monophasic effects.

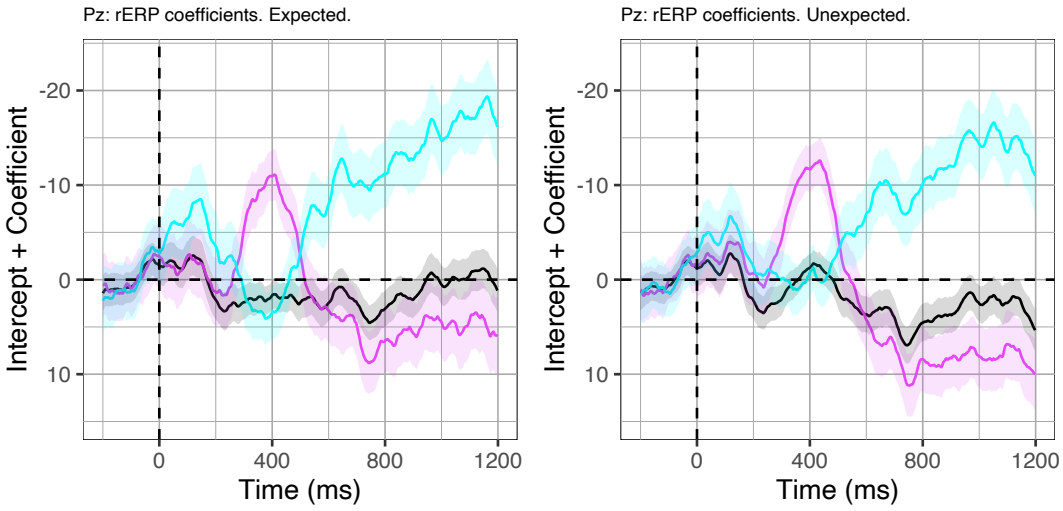


FIGURE 5.10: Model coefficients (added to their intercept) across time on electrode Pz for regression models fitted separately on the Expected and Unexpected condition of Chapter 3. Coefficients express the extent to which single-trial N400 amplitude (averaged from 300 ms - 500 ms) and Segment amplitude (averaged from 0 ms - 1200 ms) explain the EEG signal across time.

Single-Trial Dynamics in Monophasic Effect Structures

While linguistic manipulations often elicit biphasic effects between conditions, there are many crucial ERP studies in which only an N400 effect or only a P600 effect was reported relative to baseline (due in part to spatiotemporal component overlap). Intuitively, it may seem that these monophasic *effects* would speak against or at least limit the N400-P600 within-trial dynamics predicted by RI theory. However, while experimental manipulations can be constructed such that retrieval effort is equal across conditions, leading to the absence of an N400 effect, or such that integration effort is equal across conditions, leading to the absence of a P600 effect, this is not necessarily at odds with the general proposal of RI theory that expectations derived from the utterance meaning representation constructed so far will modulate both retrieval and integration. That is, even in data in which no N400 or P600 effect is observed between conditions, a *within-condition* correlation between N400 amplitude and P600 amplitude may be present at the single-trial level.

To illustrate this point, we turn to the data presented by Delogu et al. (2019) which, relative to baseline, revealed only a sustained N400 effect in one condition and only a P600 effect in another condition (but see Brouwer, Delogu, & Crocker, 2021; Delogu et al., 2021). During the experiment, a context sentence was presented which introduced a scenario, which was then followed by a critical sentence, presented word-by-word, containing the target word (see Table 5.2).[4] The experiment

[4]EEG data were processed the same way as described previously for Chapter 3 in Section 5.2. For full detail on experimental design, electrophysiological recording and processing, see Delogu et al. (2019).

Baseline	John entered the restaurant. Before long, he opened the menu and ...
Event-related violation	John left the restaurant. Before long, he opened the menu and ...
Event-unrelated violation	John entered the apartment. Before long, he opened the menu and ...

TABLE 5.2: Example stimuli from Delogu et al. (2019). Stimuli were transliterated from German. Context sentences were presented as a whole, critical sentences were presented using rapid serial visual presentation. Target words were underlined for this table.

consisted of three conditions in which the context sentence was either associated ("John entered/left the restaurant") or unassociated ("John entered the apartment") to the target word in the second sentence ("Before long he opened the menu and ..."). Both manipulated conditions created a violation of world knowledge (opening the menu after leaving the restaurant or after entering the apartment). However, target word meaning in the *event-related* violation is associated to the context whereas it is unassociated in the *event-unrelated* violation. Delogu et al. (2019) found a P600 effect but no N400 effect for the event-related condition, relative to the baseline condition (Figure 5.11). For the event-unrelated condition, an N400 effect but no P600 effect was found relative to the baseline.

The finding that the event-related violation elicits no N400 effect and only a P600 effect relative to baseline is in line with RI theory: The context associatively facilitates target word retrieval similarly in both the baseline and the event-related violation condition, explaining the absence of an N400 effect. Integration, however, is more effortful in the event-related violation condition than in the baseline condition, leading to an increase in P600 amplitude. Importantly, for the event-unrelated condition, RI theory predicts a biphasic N400-P600 effect relative to control, as both retrieval and integration should be more effortful than in baseline condition. However, only an N400 effect and no P600 effect were observed. The absence of the predicted P600 *effect* in the event-unrelated condition relative to baseline is explainable in terms of spatiotemporal component overlap (Luck, 2005) between the N400 and the P600 component (see Brouwer, Delogu, and Crocker, 2021, for evidence and Brouwer and Crocker, 2017, for a general discussion). The N400 and the P600, being opposite in polarity, partly cancel each other out in the scalp recorded signal, which may result in the attenuation – or even absence – of a P600 *effect* between conditions in the observed data. Indeed, in a follow-up study, the N400 effect disappears and the P600 effect re-emerges if the event-unrelated condition is compared to a similarly unassociated baseline condition (Delogu et al., 2021). Critically, while component overlap can lead to puzzling effect structures, this is not necessarily a problem for analyses of single-trial data: For instance, in the contrast of the event-unrelated condition to the baseline condition, average P600 amplitude may be equal in both conditions, which may seem difficult to reconcile with the large difference in average N400 amplitude when assuming correlated N400 and P600 amplitudes. However, while average N400 amplitudes may be offset in the two conditions, there may still be a correlation between ERP components within-condition.

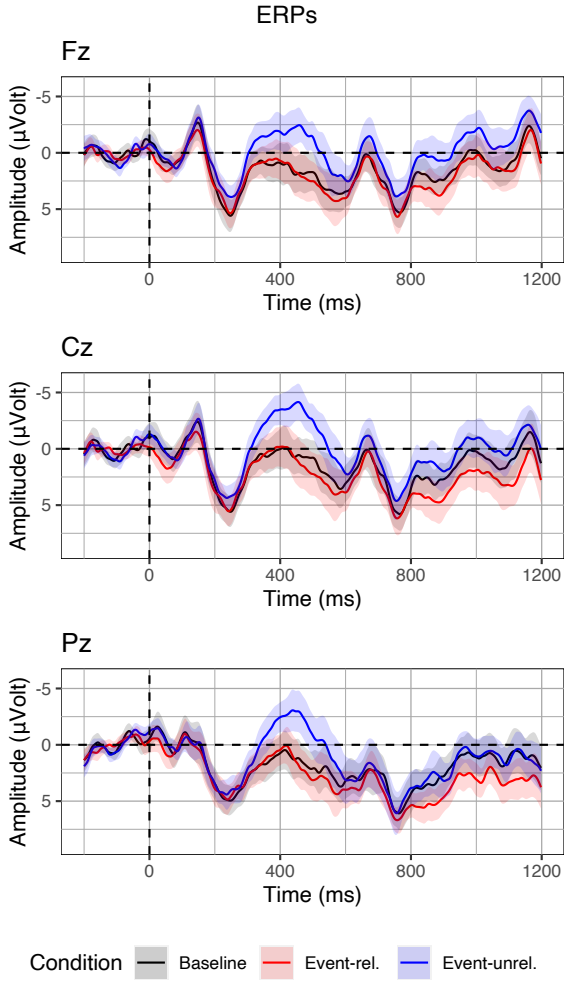


FIGURE 5.11: Average ERPs on three midline electrodes (Fz, Cz, Pz) in the baseline, event-related violation, and event-unrelated violation condition of Delogu et al. (2019). Error ribbons indicate confidence intervals based on standard errors computed across subjects.

In order to determine the single-trial dynamics in the Delogu et al. (2019) data, we conduct our analyses separately in the three conditions (analogous to Section 5.2.2). Again, this means that we cannot rely on evaluating the regression models against the effect structure observed between conditions. While we previously quantified the extent to which the regression models capture the P600 effect between conditions, there is no P600 effect for the contrast of the event-unrelated condition relative to baseline (due in part to spatiotemporal component overlap, Brouwer, Delogu, & Crocker, 2021; Delogu et al., 2021). Similarly, while there is a P600 effect for the event-related condition relative to baseline, here the average N400 amplitudes – and hence the N400 predictor values in the regression models – do not differ between conditions. Hence, assessing the extent to which the rERP models capture the effect structure across conditions is not informative. As before, however, the regression coefficients are still informative. Fitting rERP models separately for each condition, we find similar patterns as before (Figure 5.12 shows the coefficients at

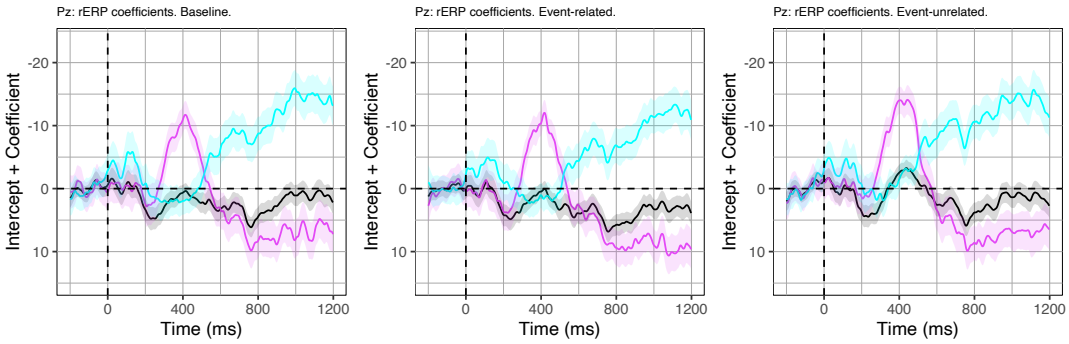


FIGURE 5.12: Model coefficients (added to their intercept) across time on electrode Pz for regression models fitted separately on the experimental conditions of Delogu et al. (2019). Coefficients express the extent to which single-trial N400 amplitude (averaged from 300 ms - 500 ms) and Segment amplitude (averaged from 0 ms - 1200 ms) explain the EEG signal across time. Error ribbons indicate standard errors on the coefficients.

Pz). Indeed, in each of the three analyses, the intercept term is equal to the average of the condition. The coefficient of the N400 predictor suggests, as before, that the variability around the mean is correlated in the N400 and the P600 time window: Within each of the three conditions of Delogu et al. (2019), more negative N400 amplitudes co-occur with more positive P600 amplitudes within-trial. In sum, our analyses do suggest that while the average N400s and average P600s may or may not differ between conditions, there is a within-trial correlation between the two ERP components within-condition. Indeed, these findings are also consistent with the computational model instantiation of RI theory that generated N400 and P600 estimates for the items in the Delogu et al. (2019) study. Within-condition, we find that the N400 and P600 amplitudes predicted by the computational model instantiation for the target words are negatively correlated (Baseline : $r = -0.62$; Event-related violation: $r = -0.52$; Event-unrelated violation: $r = -0.50$).

5.3 Discussion

Ultimately, any viable model of the neurocognition of language comprehension should explain how the N400 component and P600 component of the ERP signal are modulated at the single-trial level. While most computational instantiations of neurocognitive models do indeed make such trial-level predictions, the statistical analysis and interpretation of N400 and P600 modulations in ERP data are often focused on the effect level, comparing condition averages in predefined time windows. We have argued that this focus on effects limits our ability to decide between models and that we may improve upon this situation by moving from the effect level to the level of single trials.

We demonstrate this approach by teasing apart two explanations of biphasic N400-P600 effect patterns: On Multi-stream accounts, the N400 increases and the P600 increases are thought to stem from different pools of trials. That is, certain trials elicited by semantic anomalies that induce an N400 increase should not induce a P600 increase, whereas other trials that do not induce an N400 increase should induce a P600 increase. On RI theory, on the other hand, trials with more negative N400 amplitudes are predicted to also exhibit more positive P600 amplitudes. Critically, a set of regression models with single-trial N400 averages as predictor is able to explain systematic variability in the P600 time window and recovers the effect structure observed for the expectancy manipulation in Chapter 3. Further, an analysis of the Expected condition in isolation suggests that this relation is not specific to target words that violate selectional restrictions of the verb (Unexpected condition: "Then ate the lumberjack the axe"). This forms strong support for the explanation of RI theory and demonstrates that N400 amplitudes and P600 amplitudes are correlated at the single-trial level. Importantly, we also demonstrate that the predicted correlation between N400 and P600 amplitude is not generally at odds with monophasic effect patterns, as we found similar N400-P600 couplings within the individual conditions of Delogu et al. (2019). The key to this explanation are differences in per-condition average N400 or P600 amplitude, which may, for instance, be induced by strong priming or spatiotemporal component overlap. In sum, our results are in line with RI theory and not only eschew the need for multi-stream accounts but present explicit counter-evidence for the single-trial dynamics that follow from multi-stream architectures.

In contrast to these multi-stream accounts, RI theory posits a single-stream architecture in which expectation-based language comprehension is driven by an utterance meaning representation that is updated with every incoming word. During processing of a word, the utterance meaning representation constructed so far influences both the mapping of word forms to word meaning representations (retrieval/N400) and the updating of the utterance meaning representation with the retrieved word meaning (integration/P600). Due to the strong influence exerted by the utterance meaning representation on both retrieval and integration, N400 amplitude and P600 amplitude are predicted to be inversely correlated: Words that require more effort to retrieve, will generally be more effortful to integrate, and, consequently, more negative N400 amplitudes should, generally, co-occur with more positive P600 amplitudes. Note that the relationship between N400 amplitude and P600 amplitude is strictly *correlational*. That is, beyond the effects of spatiotemporal component overlap (see Brouwer and Crocker, 2017, for discussion; also see Brouwer, Delogu, and Crocker, 2021), there is no direct *causal* relationship between *latent* N400 amplitude and P600 amplitude in the signal itself. Rather, on RI theory, there is a causal relation between both the retrieval process underlying the N400 and the integration processes underlying the P600 to the utterance meaning representation constructed so far. It is *this* mechanistic dependence of both retrieval and integration

on the unfolding utterance representation that underlies the observed correlation in the signal itself.

As a consequence of this architecture, RI theory assumes that both the N400 and the P600 components are elicited by every word during language comprehension. Hence, biphasic N400-P600 patterns should be considered to be part of the default ERP signature of language processing. Crucially, this proposal is not at odds with the absence of N400 or P600 *effects* in certain condition contrasts. Rather, monophasic *effects* would be explained through conditions consisting of stimuli that are matched in the degree to which they make retrieval (N400) or integration (P600) effortful. Further, spatiotemporal component overlap between the N400 and the P600 can result in the partial cancellation of ERP components, which can render the observed condition-averaged waveforms unrepresentative of the underlying latent components (Brouwer & Crocker, 2017; Brouwer, Delogu, & Crocker, 2021; Delogu et al., 2021).

Additionally, the neurocomputational RI model directly predicts continuous N400 and P600 amplitude modulations, rather than binary increase patterns. This is critical since the N400 component has been shown to be a graded processing index (Kutas et al., 1984) and a similar gradedness has recently been demonstrated for the P600 (Chapter 4; see also the post-hoc analyses in Chapter 3). Hence, models of the electrophysiology of language comprehension should aim to generate continuous estimates of processing cost that reflect the graded nature of ERPs.

Lastly, it is worth noting that our single-trial analysis also goes beyond the item level: Both model-derived and human-derived processing estimates (such as Cloze probability) are computed for stimuli, abstracting over the notion of individual participants, who may experience variable processing effort. Our analyses are, however, informed by single-trial N400 amplitudes and suggest that even at this level of granularity, N400 and P600 amplitude are correlated. We interpret this as converging evidence for previous studies demonstrating that individual participants' understanding and knowledge drive expectation-based language comprehension (Troyer & Kutas, 2020; Troyer et al., 2020).

5.4 Conclusion

Most theories of the electrophysiology of language comprehension are informed by, and make predictions about ERP effects between conditions. There are multiple shortcomings with this approach: Focusing on effects bears the risk of artificially dichotomising the demonstrably continuous sensitivities of the N400 and the P600 and hence may obscure crucial aspects of EEG data that could inform theories. Further, spatiotemporal component overlap between the N400 and the P600 may result in a divergence between the observed ERP effects and the underlying, latent component structure. Finally, competing accounts for a range of ERP data at the effect

level assume fundamentally different language processing architectures. Here, we addressed these shortcomings by examining ERP data at the single-trial level. To do so, we articulated trial-level predictions of competing theories – multi-stream models and RI theory – for biphasic N400-P600 patterns observed between conditions. We then investigated the within-trial dynamics of the N400 and the P600 component. Using a regression-based approach, we quantified the extent to which single-trial N400 amplitudes are predictive of their consecutive P600 amplitudes. We provide first evidence that their amplitudes are continuously and inversely correlated: Trials with larger N400 amplitudes also exhibit larger P600 amplitudes. Further, we have shown that this finding is not limited to biphasic effect patterns, but also extends to monophasic effect patterns.

The finding that increases in N400 and P600 amplitude are coupled within-trial supports the single-stream view proposed by Retrieval-Integration theory and appears inconsistent with the processing architecture proposed by many multi-stream models, which predicts that any given trial should elicit either an N400 or a P600 increase. Our results illustrate that in order to further dissociate competing theories of the electrophysiology of language comprehension, models should make quantitative single-trial level predictions and, crucially, ERP analyses must evaluate these predictions at the trial level, rather than at the effect level.

Chapter 6

General Discussion and Conclusions

6.1 Summary of Results

This thesis started from the proposal that language comprehension is fundamentally grounded in two expectation-based mechanisms: retrieval of word meaning from long-term memory and integration of retrieved word meaning into an unfolding utterance meaning representation. These processes form the core of Retrieval-Integration (RI) theory, which offers a neurocognitive account – with an explicit neurocomputational instantiation – of the interaction of retrieval and integration and relates the two processes to empirical measures of word processing effort in the neural and the behavioural domain. Specifically, RI theory posits that the N400 component of the Event-Related Potential (ERP) signal indexes lexical retrieval and that the P600 component indexes integration. Further, reading times are posited to strongly correlate to comprehension-centric surprisal and directly relate to the integrative effort indexed by the P600. The positions of RI theory are, however, not universally accepted: Several models hold that the integration and surprisal are indexed by the N400 component instead. Contrasting these competing stances, this thesis investigated several key predictions of RI theory.

On RI theory, both retrieval and integration are taken to be expectation-based processes, as each word contributes meaning to the utterance representation and thereby constrains which continuations are more likely than others. Thus, expectancy is predicted to be correlated to both retrieval and integration. Critically, this means that dissociating the neural indices of the two processes based on expectancy effects alone is impossible. This issue can, however, be overcome by manipulating other stimulus properties that should uniquely modulate one process but not the other, thus allowing us to identify which ERP component corresponds to retrieval and integration, respectively. Specifically, while retrieval (N400) is predicted to be sensitive to both lexical association and expectancy, no such relation of association to integration (P600) is predicted. This hypothesis was investigated empirically using a context manipulation design (Design 1; Figure 6.1) that fully crossed lexical association and expectancy. Indeed, the ERP experiment elicited N400 modulations from

Design 1

Cond.		Expectancy		Association	Target
A: A+E+	Yesterday	sharpened	the lumberjack,	before he the wood stacked,	the axe ...
B: A-E+	Yesterday	sharpened	the lumberjack,	before he the movie watched,	the axe ...
C: A+E-	Yesterday	ate	the lumberjack,	before he the wood stacked,	the axe ...
D: A-E-	Yesterday	ate	the lumberjack,	before he the movie watched,	the axe ...

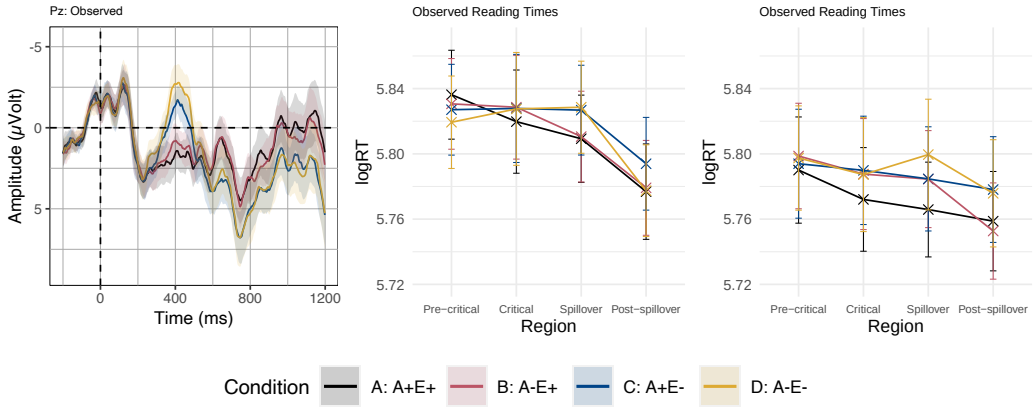


FIGURE 6.1: Top: Design 1 employed for Chapter 3, crossing expectancy of the target word with an independent manipulation of lexical association. Bottom: Results of the corresponding event-related potential and self-paced reading studies.

both lexical association and expectancy, but P600 modulations from expectancy only (Chapter 3). Further, a post-hoc analysis suggested a continuous relation between word expectancy and N400 amplitude (replicating earlier findings; Frank et al., 2015; Kutas & Hillyard, 1984), and, as a novel finding, a continuous relation also between word expectancy and P600 amplitude. Reading times recorded in two self-paced reading experiments using the same stimuli were increased for unexpected compared to expected stimuli on spillover regions. Association effects on reading times were less reliable, as a slowdown induced by low lexical association was observed when participants responded to comprehension questions, but not when they were rating the stimuli for plausibility in a binary judgement task.

The exploratory finding that P600 amplitude may be continuously related to expectancy is, in fact, directly predicted by RI theory. On RI theory, the P600 is taken to be elicited not just by impossible or violating continuations but by every word in an utterance, *continuously* reflecting integration effort. This possible continuous relation was further investigated in a dedicated study (Design 2; Figure 6.1). In this experimental design, we first presented a context paragraph in which a story was introduced. Importantly, the target word was mentioned several times already in the context paragraph in order to maximally facilitate lexical retrieval of target word meaning in the manipulated final sentence. As the main manipulation, final sentences were constructed such that the target word rendered the utterance meaning plausible, less plausible, or implausible. An initial self-paced reading experiment revealed that reading times gradually increased, the less plausible the target word,

Design 2

Context
A tourist wanted to take his huge suitcase onto the airplane ...

Condition	Continuation
A: Plausible, no attraction	Then dismissed the lady the tourist ...
B: Less plausible, attraction	Then weighed the lady the tourist ...
C: Implausible, no attraction	Then signed the lady the tourist ...

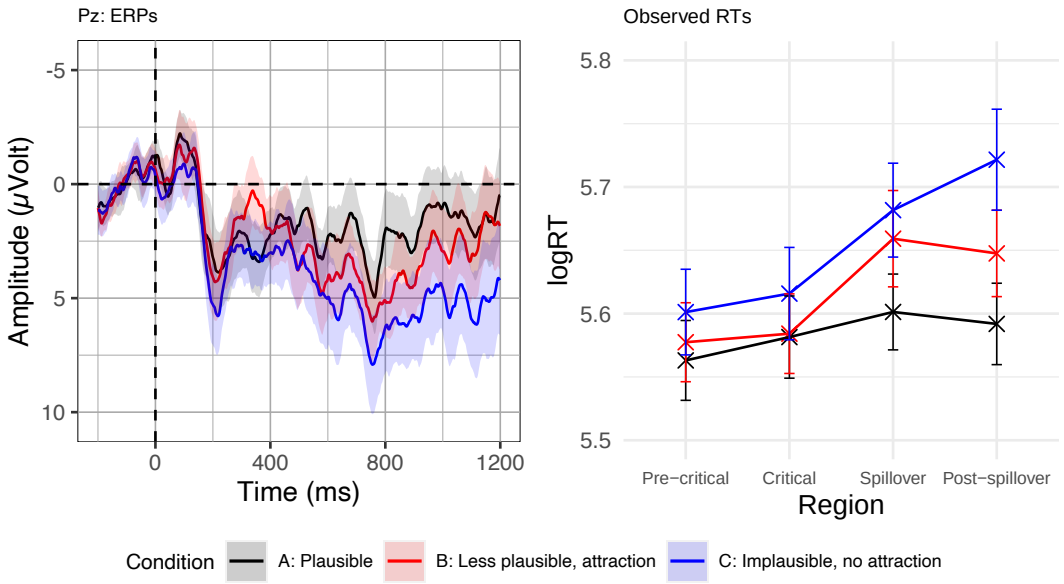


FIGURE 6.2: Top: Design 2 employed in Chapter 4, establishing the target word in a preceding context and manipulating plausibility across three levels. Bottom: Results of the corresponding event-related potential and self-paced reading studies.

suggesting differential integration effort. In the ERP study, we found that plausibility is indeed inversely and continuously related to P600 amplitude, with the least plausible target words eliciting the most positive P600. The design furthermore tested the predictions of RI theory directly against those of a group of multi-stream models, which predict either an N400 effect or a P600 effect relative to baseline, depending on the availability of a semantically attractive alternative interpretation. However, our stimuli elicited no N400 effect relative to baseline, even when no semantically attractive interpretation is available.

While theories of language comprehension are often informed by binary ERP effects observed between conditions, any viable model should ultimately explain how the N400 component and the P600 component are modulated at the single-trial level. RI theory specifically predicts that words that are unexpected given the utterance meaning representation constructed so far should, generally, be both more effortful to retrieve and more effortful to integrate. Because of this, the amplitude of the

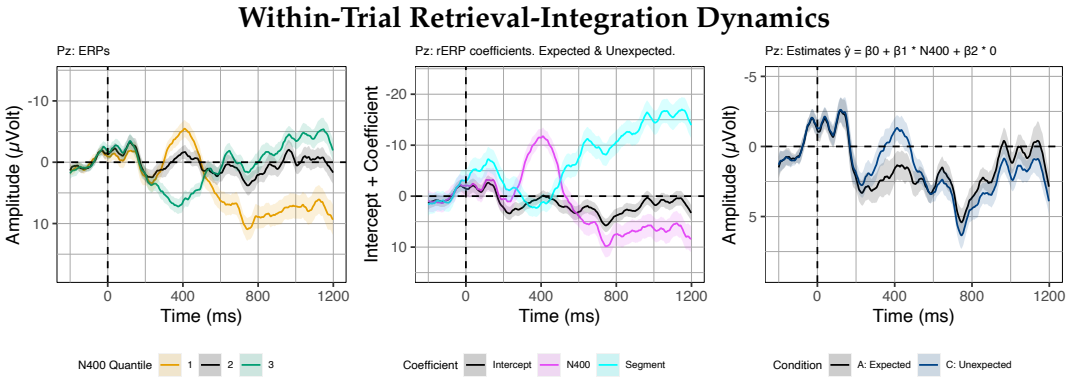


FIGURE 6.3: ERP data of Design 1, binned by subtracting Segment voltage from N400 voltage (left), regression coefficients from modelling the ERP signal as a function of N400 voltage and Segment voltage (middle), and isolated estimates computed from only the N400 predictor in the same regression models (right).

N400 and the P600 should be negatively correlated at the single-trial level, in that trials inducing more negative N400 amplitude should also induce more positive P600 amplitude. By re-analysing ERPs from earlier studies (Chapter 3 and Delogu et al., 2019) we found that this correlation is indeed present at the single-trial level (Figure 6.3). This result provides further evidence against multi-stream models, whose architecture suggests that any individual unexpected word should induce an increase in either N400 amplitude *or* P600 amplitude, but not both.

As we discuss below, our results are incompatible with both traditional interpretations of the N400 and the P600 and with the processing architecture proposed by multi-stream models in particular. Rather, our findings are parsimoniously explained by the single-stream RI model, which posits the N400 to continuously indexes retrieval effort and the P600 to continuously index integration effort.

6.2 Implications for the Neurocognition of Language

6.2.1 The N400 Indexes Retrieval

Traditionally, the N400 has been interpreted as an index of semantic integration (Brown & Hagoort, 1993, 2000; Hagoort et al., 2004). Interestingly, a large body of N400 results indicates that this component at least *also* indexes processing effort related to isolated word meaning which would not be expected to induce meaning integration (see Kutas & Federmeier, 2011, for an overview). This raises the question of to what extent N400 modulations observed while participants comprehend whole utterances are necessarily reflecting integration effort.

Indeed, Design 1 demonstrated that lexical association and expectancy have separable influences on the N400 during sentence processing (Figure 6.1). Critically, the manipulation of association was explicitly designed to not alter the expectancy of the target word, hence excluding an explanation of the association-induced N400

modulations in terms of integration effort. Based on this result, we exclude the pure integration view of the N400 (see also van Berkum, 2009, for discussion).

The findings can, however, still be explained by the hybrid view of the N400, which posits that the N400 indexes both semantic integration and lexical retrieval (Baggio & Hagoort, 2011; Lau et al., 2016; Nieuwland et al., 2020). However, in Design 2, we found that target words that vary gradually in utterance meaning plausibility elicit no N400 modulations if word meaning of the target word is primed equally across conditions. That is, even though the target word differed in how difficult its meaning was to integrate with the prior context, we observed no N400 modulations. Thus, we also exclude the hybrid view of the N400, as this account would predict the N400 to be modulated by both association – which it was – and by plausibility – which it was not.

Another variant of the integration view was developed within multi-stream models which were put forward in response to semantic P600 studies that often found no N400 effect for implausible relative to plausible target words (see Bornkessel-Schlesewsky & Schlesewsky, 2008; Brouwer et al., 2012; Kuperberg, 2007, for a review). On these multi-stream accounts, the N400 was maintained as an index of integration effort by stipulating that no increase in N400 is produced when a semantically attractive alternative interpretation is made available by the input (e.g., by going with the plausible reading that *the athletes threw the javelin* when reading "the javelin has the athletes thrown", Hoeks et al., 2004). To achieve this, multi-stream accounts posit that integration – as indexed by the N400 – proceeds in a structure-insensitive, quasi-compositional (Rabovsky & McClelland, 2020) manner (see also Li & Ettinger, 2023; Ryskin et al., 2021).

Indeed, multi-stream models can employ the notion of semantic attraction to explain some association-induced N400 modulations, typically by postulating that an implausible target word may be successfully integrated with the associated context within the semantic stream. However, the stimuli of Design 1 were deliberately created such that they do not support any dependency between the target word and the associated adverbial clauses, thereby ruling out any structural or semantic attraction, which would be prerequisite for an explanation of the N400 effect between unexpected-associated and unexpected-unassociated target words in terms of the "good-enough" processing effects (Ferreira & Patson, 2007; Rabovsky & McClelland, 2020) proposed also by multi-stream models. Hence, this quasi-compositional integration view of the N400 would have to be extended into a hybrid account on which the N400 indexes both (structure-insensitive) integration and retrieval, in order to fully explain the data of Design 1 within such a processing architecture.

Crucially, however, in Design 2, we specifically developed stimuli that *do* make a semantically attractive alternative interpretation available in one condition (Condition B) but *not* in the other (Condition C; see Figure 6.2). While we did observe a P600 increase for the condition that makes an alternative interpretation available (in

line with multi-stream accounts), the N400 increase did *not* resurface for the implausible condition without semantic attraction, thereby falsifying the key prediction of the quasi-compositional integration view of the N400.

The interpretation of the N400 that is directly supported by the results of this dissertation is the lexical retrieval account of the N400 (Brouwer et al., 2012; Kutas & Federmeier, 2000, 2011; Lau et al., 2009; Lau et al., 2008; van Berkum, 2009, 2010), i.e., the view that the N400 continuously indexes the effort involved in the retrieval of word meaning from long-term memory. Critically, by assuming an *expectation-based* notion of retrieval, this explanation also accounts for expectancy effects on the N400, as found in Design 1 and many previous experiments manipulating Cloze probability (Kutas and Hillyard, 1980; Kutas and Hillyard, 1984, for an overview, see Van Petten and Luka, 2012), or as found in studies examining the link between N400 amplitude and corpus-based surprisal (Aurnhammer & Frank, 2019a; Ettinger, 2020; Frank et al., 2015; Merkx & Frank, 2021; Michaelov et al., 2023; Michaelov & Bergen, 2020; Parviz et al., 2011; Szewczyk & Federmeier, 2022). Further, the retrieval view of the N400 is also in line with the complete absence of N400 effects in Design 2, as target words were repeated several times throughout a preceding context paragraph, thereby facilitating lexical retrieval equally across conditions.

6.2.2 The P600 Indexes Integration

The P600 was originally described and investigated in studies that manipulated syntax, leading to an initial interpretation as an index of syntactic processing effort or syntactic integration (Friederici, 1995; Hagoort et al., 1999; Kaan et al., 2000; Kaan & Swaab, 2003; Osterhout & Holcomb, 1992). However, semantic illusion studies, which often did not elicit an N400 effect, frequently reported P600 effects for manipulations of sentences that were semantically implausible but syntactically well-formed, relative to plausible control sentences. This was also the case for both Design 1 and Design 2 in which syntactically well-formed and unambiguous sentences that differed in target word expectancy or plausibility induced P600 effects, thus adding to the literature of semantic P600s. In line with previous results (see Chapter 2 for an overview), the interpretation of the P600 as indexing *only* syntactic processing is therefore excluded.

In response to semantic illusion studies, the P600 was re-interpreted within multi-stream architectures that posited the P600 to reflect a conflict between competing semantic analyses. However, P600 effects have also been observed for implausible sentences that do not induce competing analyses relative to plausible controls (e.g., Chow & Phillips, 2013; Delogu et al., 2019, 2021). Similarly, in Design 1, we observed P600 effects for unexpected continuations that do not make semantically attractive alternative interpretations available. The same was the case in the results obtained for Design 2: The P600 effect remained even in absence of a semantically attractive alternative interpretation, which would be required to explain the absence

of the N400 effect and the presence of the P600 effect. Thus, we rule out this conflict detection/resolution interpretation of the P600 entirely.

An alternative interpretation of the P600 as an index of semantic integration was proposed by Brouwer et al. (2012). In line with this interpretation, we found the P600 to be modulated by expectancy but, importantly, not by lexical association (Design 1), thus adding to the large body of expectancy-related P600 effects, which typically result in biphasic N400-P600 patterns (see Van Petten & Luka, 2012, for an overview). As an important novel observation, we also established that the P600 response is continuous, with graded differences in integration effort manifesting as graded differences in P600 amplitude. These continuous modulations thus suggest that the P600 is elicited by every word in an utterance and are at odds with proposals that hold the P600 to be a binary response to strongly implausible utterance meanings (e.g., Bornkessel-Schlesewsky et al., 2011; Kuperberg et al., 2020). Our P600 findings are thus in line with previous studies demonstrating that any difficulty in establishing a coherent meaning representation based on syntactic, semantic, and pragmatic information, appears to trigger an increase in P600 amplitude (Burkhardt, 2006, 2007; Cohn & Kutas, 2015; Delogu et al., 2019; Dimitrova et al., 2012; Hoeks et al., 2013; Regel et al., 2010; Schumacher, 2011; Spotorno et al., 2013; Xu & Zhou, 2016).

6.2.3 A Single-Stream Account of the N400 and the P600

In this dissertation, we have emphasised theoretical accounts that make predictions for both the N400 and the P600 within architectural proposals that specify how their underlying processes interact. Because multi-stream accounts posit that N400 and P600 increases depend on the presence/absence of semantic attraction, they predict that any single anomalous sentence continuation processed by a single participant (i.e., any trial) should induce either an increase in N400 amplitude or in P600 amplitude (but see Li and Ettinger, 2023, for a model capable of producing biphasic modulations). In our study of within-trial N400-P600 dynamics, we found that rather than being unrelated,[1] the amplitude of the N400 and the P600 are negatively correlated at the single-trial level: Trials inducing more negative N400 amplitude generally also induce more positive P600 amplitude. Hence, this study provides counter-evidence for the single-trial N400-P600 dynamics proposed by multi-stream models. Instead, the observed N400 and P600 coupling supports a single-stream account, within which the processes underlying the N400 and the P600 are often jointly modulated, e.g., because both underlying processes are expectation-based.

We have argued that neither an integration view nor a hybrid view of the N400 can be maintained based on the findings of this dissertation. Similarly, an interpretation of the P600 in exclusive terms of syntactic integration cannot be upheld, whereas an interpretation of the P600 as a reflection of conflicting analyses is directly

[1] In our approach, the prediction of multi-stream models, in fact, corresponds to a positive correlation, since N400 and P600 deviations are specified relative to the grand average of ERPs exhibiting a biphasic effect (see Chapter 5).

opposed. As such, multi-stream accounts of the N400 and the P600 become increasingly difficult to maintain since our findings contradict two of their key predictions: The resurfacing of the N400 effect in the absence of a plausible alternative interpretation and the isolated modulation of the N400 and the P600 at the single-trial level. Retrieval-Integration theory, on the other hand, combines the retrieval view of the N400 and the integration view of the P600, both of which remained after our studies excluded competing interpretations. Hence, we argue that the sum of our findings is parsimoniously explained by this single-stream, Retrieval-Integration account in which expectation-based retrieval (N400) and integration (P600) are two key processes active during incremental language comprehension.

6.3 Expectation-Based Retrieval and Integration in Language Comprehension

Rather than assuming a multi-stream architecture within which different streams process the input in parallel, Retrieval-Integration theory proposes a single-stream architecture consisting of two main operations: Retrieval, linked to the N400, during which word forms are mapped to word meanings, and integration, linked to the P600, during which retrieved word meanings are mapped into an updated utterance meaning representation. The process thought to underlie the N400 (retrieval) feeds its output to the process underlying the P600 (integration). Further, both processes also receive input from the output of previous integration and, hence, both retrieval and integration are constrained by the utterance meaning representation constructed thus far. Thus, RI theory proposes a formal description of expectation-based retrieval and integration, and how these processes interact during language comprehension.

Because of the dependency of both retrieval and integration on what has been understood so far, both the N400 and the P600 are predicted to be modulated by expectancy. However, while the utterance meaning representation constructed so far may lead to a pre-activation of potentially upcoming word meaning, thereby facilitating retrieval of word meaning from long-term memory, the memory state can also be altered associatively. That is, every processed word may lead to the activation of memory representations of frequently co-occurring concepts, regardless of whether these concepts would be expected based on comprehension of the utterance. The influences of expectancy and association on retrieval effort were investigated in Design 1, which revealed separable and additive effects of both factors: Unexpected words induced more negative N400 amplitude relative to expected words while presenting associated lexical material before the target word induced a reduction in N400 amplitude for both expected and unexpected words. Indeed, N400 attenuations by association can be so strong that they may override any difference in N400 amplitude that would be predicted by expectancy (see Brouwer et al., 2012, for discussion).

Investigating this proposal in Design 2, we repeatedly mentioned the target word throughout a context paragraph and indeed observed no N400 differences between conditions that differed in plausibility (in line with previous empirical evidence; see Chapter 5).

While integration effort has previously been found to be reflected in P600 amplitude (see Chapter 2 for an overview), these studies typically employed binary designs contrasting plausible with implausible continuations. Similarly, in Design 1, we found increased P600 amplitudes for unexpected relative to expected target words, while lexical association did not modulate the P600. However, on Retrieval-Integration theory, the amplitude of the P600 is explicitly predicted to continuously index integration effort. We gathered first evidence in support of this prediction in a post-hoc analysis of the baseline condition of Design 1, where we found continuous N400 *and* P600 modulations as a function of cloze probability, thus suggesting that the P600 component is not only elicited by impossible or violating continuations. Design 2 was developed to further investigate this question in a dedicated experiment. Consistent with the prediction of RI theory, we found that conditions varying in plausibility across three levels (plausible, less plausible, implausible) elicited increasingly positive P600 amplitude – a relationship that was modelled statistically by offline plausibility ratings on a Likert scale.

Lastly, RI theory also makes predictions about joint N400 and P600 modulations at the single-trial level: Specifically, due to the aforementioned dependence of both retrieval and integration on the utterance meaning constructed so far, words that are more effortful to retrieve should, generally, be more effortful to integrate. In the electrophysiological domain, this proposal predicts that the amplitude of the N400 and the P600 should be negatively correlated, in that more negative N400 amplitudes should co-occur with more positive P600 amplitudes. In a study of within-trial retrieval-integration dynamics, we indeed found evidence for a negative within-trial correlation between N400 amplitude and P600 amplitude in ERPs that exhibited biphasic effects between conditions (Design 1). Importantly, we demonstrated that the same N400-P600 dynamics also underlie ERPs that exhibited only monophasic effects between conditions (Delogu et al., 2019).

In sum, our investigations provide strong support for a single-stream architecture in which expectation-based retrieval (N400) and integration (P600) are interlocking, core processes underlying language comprehension. Hence, *both* the N400 and the P600 should be elicited by every word in an utterance, and be considered part of the default ERP signature of language processing.

6.3.1 The Behavioural Correlates of Retrieval and Integration

Our understanding of expectation-based language comprehension is supported not only by neural but also by behavioural evidence. While the focus of the empirical

work conducted for this thesis lies on event-related brain potentials, we also investigated the behavioural domain by collecting reading time data for both experimental designs in self-paced reading studies. These experiments were designed to illuminate how processing effort during retrieval and integration manifests behaviourally. That is, we were interested in how modulations of the N400 and the P600 relate to increases and decreases in reading times. Indeed, in the computational instantiation of RI theory proposed by Brouwer, Delogu, Venhuizen, and Crocker (2021), estimates of comprehension-centric surprisal (Venhuizen et al., 2019) are taken to be proportional to reading times. The model thus proposes a strong relation of surprisal to integrative processing and, hence, predicts a close link of reading times to P600 amplitude (see also Brouwer, Delogu, Venhuizen, and Crocker, 2021).

In the first two self-paced reading studies, participants were presented with the materials of Design 1. In the first version, the task for the participants was to respond with a binary plausibility judgement to every sentence they read. We found that unexpected words were read significantly slower on the first of two spillover regions (Figure 6.1, middle). The same design was also validated with a different task.[2] Using comprehension questions, we found that unexpected words were read slower than expected words across both spillover regions (Figure 6.1, right). Further, a relation of reading times to word expectancy was found also within the baseline condition, replicating earlier findings of a continuous relation of behavioural measures to corpus-based surprisal (e.g., Fernandez Monsalve et al., 2012; Frank & Thompson, 2012; Smith & Levy, 2013). These behavioural results from Design 1 indeed suggest that reading times at least *also* correlate reliably with the P600.

Furthermore, in Design 1, we found evidence for behavioural effects of lexical association. This modulation was, however, only observed when participants answered comprehension questions and not when rating plausibility. Speculatively, the absence of association effects while rating plausibility could be explained by the fact that the participants may have realised that the associated/unassociated adverbial clause does not inform plausibility. That is, only the expectancy manipulation provided relevant information for the binary plausibility judgement task (focusing on the relation between the main verb, "sharpened/ate", and the target word, "axe"), whereas comprehension questions were also asked about the associated/unassociated adverbial clauses ("before he the [wood stacked/movie watched]"). Hence, the influence of lexical association on reading times could be more pronounced in the experiment with comprehension questions due to task relevance.

The presence of the association-induced reading time modulations also suggests an interesting pattern between the ERP data and the reading time data of Design 1: The reading times on the Spillover region appear to mirror the pattern of the N400

[2]The self-paced reading studies were conducted as web-based studies since a virus pandemic made testing in-lab unfavourable. Due to this circumstance, there was limited control over the experimental environment, which lent the choice of task particular importance.

in that both lexical association and expectancy modulated the dependent variable, whereas on the Post-spillover region, only an effect of expectancy was observed – mirroring the P600 modulations. This pattern raises the question of to what extent retrieval effort may also be reflected in reading time measures, and, more generally, to what extent and how reading time signatures extending over several regions can or cannot be attributed to different underlying cognitive processes.

The link of reading times to the P600 was explored further in Design 2. We found graded reading time increases that corresponded to graded differences in plausibility. Interestingly, the corresponding ERP study elicited no N400 modulations (due to strong but equal priming across conditions) but graded P600 modulations. That is, even though previous research found that both the N400 and reading times scale with word predictability (Fernandez Monsalve et al., 2012; Frank et al., 2015; Smith & Levy, 2013), this link does not uphold when the target word is primed equally across conditions while differing in plausibility. This pattern between P600 amplitude and reading times in the absence of any N400 modulations thus provides strong evidence for a direct link between comprehension-centric surprisal, the P600, and reading times. A post-hoc analysis further corroborated this case: In an rERP analysis, we found that average per-item reading times from the self-paced reading study model the P600s in the corresponding ERP experiment.

In sum, the behavioural results indicate a reliable connection between expectation-based influences on P600 amplitude and reading times. By comparison, reading time modulations by association (and hence the N400) were less robust and more short-lived (Design 1). Most importantly, the relation of the P600 to reading times was even upheld when no N400 modulations were observed at all (Design 2). Thus, the behavioural results of this thesis support the direct link of reading times to comprehension-centric surprisal and hence to the integrative processing effort indexed by the P600 (as argued for by Brouwer, Delogu, Venhuizen, & Crocker, 2021).

6.3.2 The P600 as a Continuous Index of Comprehension-Centric Surprisal

We presented evidence that the amplitude of the P600 is sensitive to expectancy as well as to plausibility and that it is directly related to reading time modulations. The N400 was found to be sensitive to expectancy as well as to lexical association and, importantly, no N400 modulations were observed when target word meaning was primed equally strongly across conditions. These results suggest that the P600, rather than the N400, is a continuous index of integrative processing during expectation-based language comprehension. This conclusion has important consequences for the notion of surprisal in language comprehension (Hale, 2001, 2003; Levy, 2008).

Surprisal has oftentimes been operationalised using language modelling approaches, which estimate the probability of a word given the words preceding it.

However, while language model surprisal has been shown to be a good predictor of neural (Frank et al., 2015, who found a significant relation to the N400) as well as behavioural measures (e.g., Fernandez Monsalve et al., 2012; Frank & Thompson, 2012), comprehension is influenced not only by linguistic experience but also by world knowledge (Venhuizen et al., 2019). Further, the goal of the human language processing system is to comprehend. Because of this, Venhuizen et al. (2019) and Brouwer, Delogu, Venhuizen, and Crocker (2021) argue that a notion of surprisal centred around comprehension rather than language statistics alone should better account for how linguistic experience and world knowledge combine in informing a rich, probabilistic utterance meaning representation. In the model by Brouwer, Delogu, Venhuizen, and Crocker (2021), it is this utterance meaning representation, which shapes expectations about upcoming word meaning – during expectation-based retrieval – and utterance meaning – during expectation-based integration.

Against this backdrop, current research examines which neuro-behavioural correlates of expectation-based language processing may directly index such a comprehension-centric notion of surprisal. Based on ERP results by Delogu et al. (2019) and a self-paced reading replication using the same stimuli, Brouwer, Delogu, Venhuizen, and Crocker (2021) argued that while retrieval effort – as indexed by the N400 – often *correlates* with utterance-level surprisal, the P600 reflects the latter more directly. This is strengthened by the observation that reading times, which were the first established indices of surprisal (Hale, 2001; Levy, 2008; Smith & Levy, 2013), closely mirrored the modulation pattern of the P600 when component overlap is taken into account (Brouwer, Delogu, & Crocker, 2021).

Adopting the same comprehension-centric view on expectation-based language comprehension, we found corroborating evidence for this argument: The modulation patterns of the amplitude of the P600 reliably resembled those of reading times across spillover regions. We found that P600 amplitude varied continuously as a function of utterance meaning expectancy and plausibility and, crucially, that it did so even when no N400 modulations were elicited. The P600 component was elicited by every word, ranging from expected, plausible target words to unexpected, implausible, and violating target words. Based on these results, we argue that future electrophysiological investigations of surprisal should adopt a comprehension-centric view on expectation-based language processing. Such an investigation should focus on the integrative effort indexed by the P600, rather than on the retrieval effort indexed by the N400.

6.4 Conclusions

The goal of the language comprehension system is to understand the message being communicated. We have argued that two important mechanisms that must be involved during utterance comprehension are retrieval, the process by which word

meaning is accessed in long-term memory, and integration, the process by which an incrementally constructed utterance meaning representation is updated with retrieved word meaning. Crucially, each incoming word contributes meaning to the utterance representation and thereby makes some continuations more likely than others. Thus, we argued that both retrieval and integration can be facilitated based on what has been understood so far and, hence, are expectation-based processes.

To investigate the role of retrieval and integration in expectation-based language comprehension, we employed ERPs – a neurophysiological method that provides a multi-dimensional window into the nature and time course of language comprehension. Reviewing key ERP data and language processing theories, we have argued that establishing electrophysiological indices of processes such as retrieval and integration has direct consequences for our understanding of the organisation of the language comprehension system as a whole. Even though empirical results are abundant, the field has converged neither on a generally accepted formalisation of the processes necessary for comprehension, nor on their mapping to ERP components. The resulting uncertainty hinders progress in the description of the neurocognitive architecture of the language comprehension system, which thereby impedes the effective design and unambiguous interpretation of experimental studies. This situation can, however, be remedied by formalising verbal theories of language comprehension into computational models, from which predictions about ERP components can be derived. One theory that offers such a formal model of expectation-based retrieval and integration, while specifying their relationship to ERP components, is Retrieval-Integration theory (Brouwer et al., 2017; Brouwer et al., 2012). This single-stream model posits that the N400 component indexes retrieval effort and that the P600 component indexes integration effort. To examine these expectation-based processes, we tested several key predictions of Retrieval-Integration theory that contrast competing interpretations and models of the N400 and the P600.

Based on the presented data, we have argued that the traditional interpretation of the N400 as an index of integration cannot be upheld, regardless of whether a compositional or a quasi-compositional notion of integration is adopted. Similarly, a hybrid view of the N400, on which this component indexes both integration and lexical retrieval is not supported by the evidence. Instead, our findings are in line with a retrieval-only view of the N400.

Turning to the P600, our results do not support its exclusive interpretation as an index of syntactic processing effort. Furthermore, interpreting the P600 as a reflection of conflicting analyses (e.g., generated within a multi-stream architecture) is incompatible with our data. Instead, we found that within utterances, the P600 is elicited by plausible and implausible/violating target words alike and that its amplitude continuously indexes the effort involved in updating an utterance meaning representation with new incoming word meaning.

As we consider the processes underlying the N400 and the P600 to be intertwined, we have also examined how the N400 and the P600 are jointly modulated

and presented evidence that their amplitudes are negatively correlated. We take this finding as an indication that due to the top-down influence of expectancy on both retrieval and integration, words that are more effortful to retrieve are, generally, more effortful to integrate. This finding directly opposes theories that predict isolated increases in N400 *or* P600 amplitude depending on the absence/presence of a semantically attractive alternative utterance interpretation (e.g., multi-stream models).

Investigating the notion of comprehension-centric surprisal in the behavioural domain, we found that reading times are directly related to the integrative processing indexed by the P600 and argued that their correlation to the N400 is only indirect.

We conclude that the sum of our findings is parsimoniously explained by the single-stream Retrieval-Integration model, an explanation which eschews the need for multi-stream architectures and the notion of quasi-compositional integration. Expectation-based retrieval and integration should be considered central mechanisms of language comprehension, and, hence, the N400 and the P600 are proposed to form part of the default ERP signature of incremental utterance comprehension. Future research into the neurocognition of expectation-based language understanding should adopt a comprehension-centric view on expectancy/surprisal and thus focus on the integrative effort indexed by the P600.

Appendix A

Regression-Based ERP Estimation

Throughout this dissertation, we apply rERPs (Smith & Kutas, 2015a, 2015b), a regression-based (re)-estimation technique. Although put forward as a tool to analyse event-related potentials (ERPs), the technique can be applied similarly to other dependent measures. We showcase its use for both ERPs and reading times. Further, the technique is not limited to least-squares regression, but it can easily be extended to linear-mixed effects regression. In this appendix, we re-trace how the rERP method derives directly from the traditional ERP averaging procedure, explain its general principles and argue for its advantages and unique potential as an analysis approach that can help to develop a deeper understanding of the data.

A.1 rERPs as ERP Averaging

To display a single ERP waveform at a single electrode, the traditional procedure is to first compute the average waveform for each participant that took part in the experiment and then compute the mean of the resulting per-subject averages to obtain the grand-average ERP. This two-step procedure has two effects: First, the ERP of each participant is weighed equally, even though there is typically not the same amount of data for each participant, due to artefact rejection. Second, the variability across subjects can be visualised around the average waveform as an error ribbon, for instance, using the standard error multiplied by 1.96 to yield a confidence interval.[1] As a result of this procedure, we obtain a waveform displaying the average potential over time and across subjects.

This operation can be formalised equivalently as a set of linear models containing only an intercept (Equation A.1).

$$\hat{y}_{ts} = \beta_{0ts} * 1 \tag{A.1}$$

The set consists of one model fitted for each subject and at each time sample within that subject. Thus, the outcome variable y_{ts} corresponds to the average voltage at time sample t of subject s. At this granularity of subject and time sample,

[1] The variability around the mean computed across subjects also informs traditional statistical analyses.

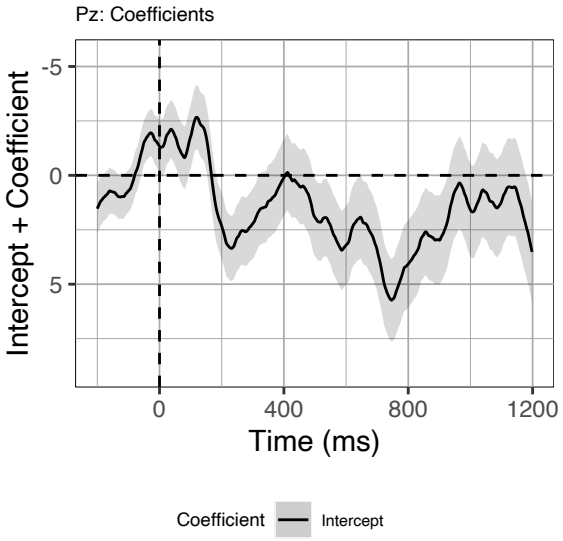


FIGURE A.1: Coefficients from a set of intercept-only models, mathematically equivalent to the average obtained from the traditional grand-average waveform computing procedure. Error ribbons indicate confidence intervals, computed as 1.96 times the standard error around the mean, across subjects.

we are left with a vector of scalar voltages which correspond to all the experimental items recorded for this subject (that have not been rejected). To model the data, we have only provided an intercept term to each model, a coefficient which is fitted on a predictor consisting only of ones with its length equal to the vector of the outcome variable. Given just this intercept term, the intercept β_0 that reduces the sum of the squared error most is equal to the average of the data (see Smith & Kutas, 2015a, for the full algebraic equivalence). Thus, after fitting, the coefficients in the set of models contain the average voltages of each subject at each time sample. In line with the traditional averaging procedure, we can now compute the grand average of the β_0 terms across time – averaging across subjects – and indicate the variability across subjects as an error ribbon. To reiterate, the coefficients of the set of intercept-only regression models are mathematically equivalent to the traditional way of computing the grand-average ERP (Smith & Kutas, 2015a).

In many ERP experiments, scientists make use of experimental conditions to investigate specific effects, which become manifest as a difference in the dependent measure between one condition and another. For instance, half of the experimental items in a psycholinguistic experiment (Chapter 3) presented well-formed sentences ("Yesterday, sharpened the lumberjack the axe") in which the target word axe is expected, whereas the other half of the items changed the context ("Yesterday, ate the lumberjack the axe"), rendering the target word unexpected (stimuli transliterated from German, original word order preserved).

To investigate the difference between conditions using the traditional procedure, we compute the average potential across time for each participant *and* condition. The

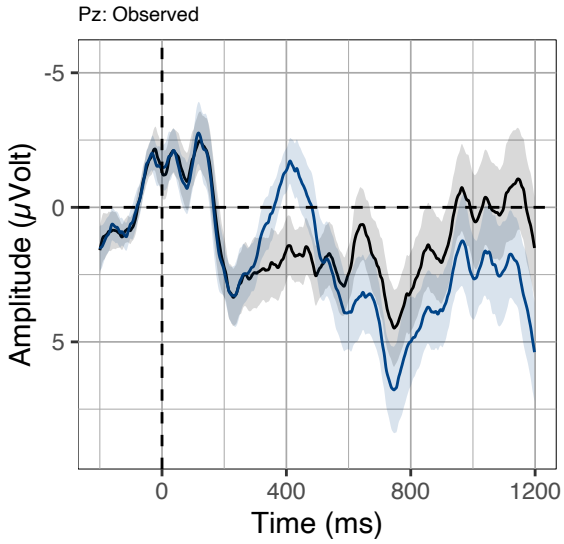


FIGURE A.2: Grand-average ERPs computed per condition for a contrast manipulating the expectancy of the target word (Design 1, Condition A vs. C).

resulting per-participant, per-condition averages are then averaged per condition and we can compare the average waveforms of the two conditions as well as the variability around them across subjects. In this study, we found that the ERP of the unexpected compared to the expected target words was more negative around 300 to 500 milliseconds and more positive from 600 milliseconds onwards. This too can be computed equivalently by a set of linear models (Equation A.2).

$$\hat{y}_{ts} = \beta_{0ts} + \beta_{1ts} x_1 \tag{A.2}$$

Again, we obtain the intercept coefficients, which represent the average of the data at time sample t and for each subject s. Going beyond the average, the difference between the conditions is captured by introducing the coefficient $\beta 1$ which is multiplied by a corresponding predictor x_1. This predictor codes for the two conditions by representing the two conditions numerically, for instance as 0.5 and -0.5.[2] Hence, we start with the average of the data for each subject s at each time sample t, given by β_{0ts}, and then use the coefficient β_{1ts} to offset the waveforms at any time samples in which the averages of the two conditions are not identical. We can then use the coefficients (Figure A.3, left) to re-compute the per-condition voltages of each subject at each time sample (Figure A.3, right). Again, the estimated outcome computed by this set of models is mathematically equivalent to the traditional way of computing the per-condition average ERP.

[2]Note that the choice of coding will have consequences for the interpretation of the intercepts. Other coding methods, such as treatment-coded conditions (0 and 1) are equally valid for certain hypotheses. Throughout this appendix we will only use predictors which average to zero, as this ensures the equivalence of the intercept to the arithmetic mean of the dependent variable.

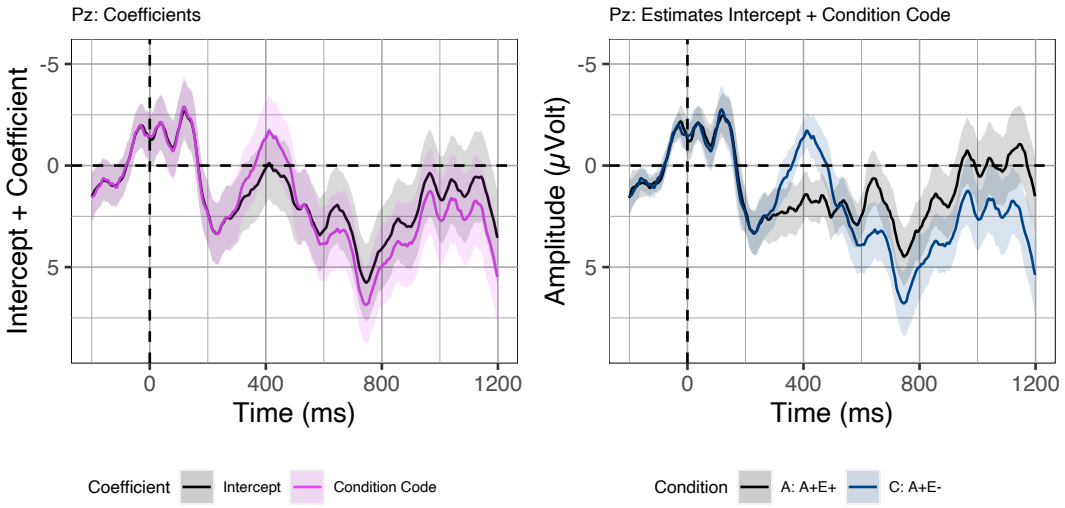


FIGURE A.3: Condition-contrast coded models recovering the original condition difference in the data. Left: Coefficients for the intercept and the contrast (added to the intercept). Right: Estimated ERPs, which are mathematically equivalent to traditional per-condition average ERPs, as shown in Figure A.2.

A.2 Beyond Conditions: Continuous Predictors

In linear models, information about what data belongs to which condition can be represented numerically. As outlined above, this is achieved by assigning the same value with opposing polarity to observations belonging to two conditions, respectively. However, we can also take a different perspective on the modelling task and not provide the linear models with this information. Rather, we quantify the relevant stimulus properties that create the difference between the conditions directly. In the above example, the difference in expectancy of the target word ("axe") in the two conditions can be measured by its cloze probability - a continuous measure of word expectancy on a probability scale of 0 to 1. Note that while most of the cloze probabilities in the unexpected condition are zero, there is indeed variability in expectancy within the expected condition. Thus, we not only model the difference between two conditions but can make quantitative predictions on a continuous scale of expectancy. To preserve the desirable property of the intercept representing the average of the data, it is necessary to transform the continuous predictor to a scale on which its mean equals zero. Further, if we work with several numerical predictors and want to assess their relative influence, predictors should be comparable. To achieve this, the predictor values are centred, whereby the mean of the predictor is subtracted from each value, and z-standardised, whereby each value is divided by the predictor's standard deviation, expressing it as a z-value (Equation A.3).

$$z_i = \frac{x_i - \bar{x}}{\sigma} \tag{A.3}$$

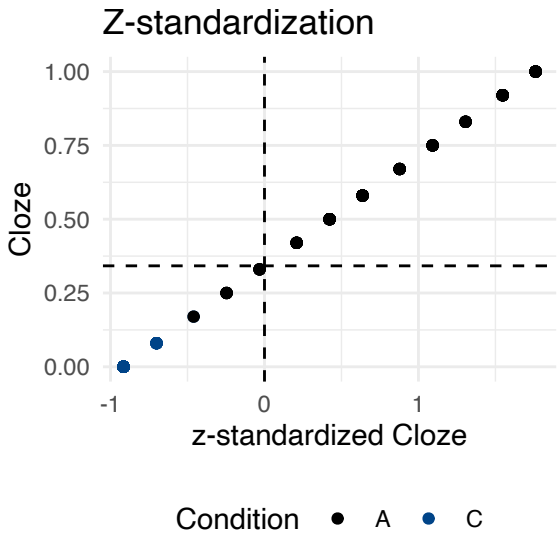


FIGURE A.4: Effect of z-standardization on Cloze probability in the Aurnhammer et al. (2021) expectancy manipulation. Unique Cloze probabilities are shown. Averages on the original and the z-standardised scale are shown by horizontal and vertical lines.

As a result of the z-standardisation, the values are now expressed as standard deviations with an average of zero. This is a linear transformation, meaning that the relative differences between the predictor values remain the same. The values are merely projected linearly to another scale (compare the scales in Figure A.4). Instead of providing the models with a condition-coded predictor, we can now model the data using the cloze probabilities of the stimuli directly. To do so, we use models of the form displayed in Equation A.2, where the predictor vector x_1 contains the standardised cloze probability values. We now obtain a set of fitted models that predict the average of the data (using the intercept coefficient β_{0ts}) and any variability in the data that can be explained by the cloze probability values (using the coefficient $\beta_1 x_{1ts}$) at each time sample and subject. This set of equations can then be used to compute the forward estimates $\hat{y}_{ts}$, i.e. the amplitudes that are predicted by the set of models fitted at each time sample and for each subject, given some cloze probability value. Figure A.5 shows the corresponding model coefficients (left) and the estimated data (centre). Recall again that the model was not provided with explicit information about the conditions. We are merely visualising the estimated data using the conditions as a grouping that is applied to the estimated data afterwards. The graph suggests that cloze probability captures the condition averages rather well. In essence, we are now visually comparing the observed data y (Figure A.2) to the estimated data $\hat{y}$ (Figure A.5, centre). A key way to quantify this difference is to compute the residual error between y and $\hat{y}$. To understand the residual error, consider that y (rather than $\hat{y}$) can also be written out as the linear model equation A.2 to which the residual error ϵ was simply added (Equation A.4).

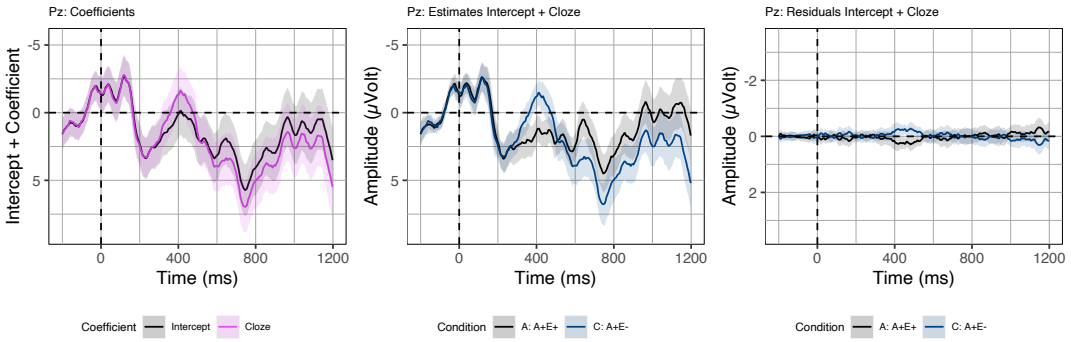


FIGURE A.5: Coefficients (left), estimated ERPs, and residual error for a model using cloze probability to model an expectancy-induced condition difference.

$$y_{ts} = \beta_{0ts} + \beta_{1ts}x_1 + \epsilon_{ts} \tag{A.4}$$

The error term is used to capture any difference between the observed data y and the estimates $\hat{y}$. Hence, the residual error can be computed simply as y - $\hat{y}$. Note that when we determine the set of coefficients that best predict the observed data, we are doing nothing else than minimising the sum of the squared error. In the case of rERPs, we have a model for each time sample and subject and thus we obtain a vector of residual error terms for each model. Just like in the case of the estimates above, we can visualise the average error for the two conditions over time samples and use compute confidence intervals across subjects to visualise the variability in average error across subjects as a ribbon.

This provides us with a useful tool to examine and potentially improve the fit of the models to the data. For instance, we can look at the residual plot for different predictor transformations. In this particular example, we find that computing the logarithm of cloze probability (before z-standardising) provides a better fit to the data in the N400 time window (Figure A.6). Clearly, however, we want to ground this decision not only in eye-balling a residual graph. To sum up the residual error in a single number, we can also inspect measures such as the sum of the squared error or compute the averages of model quality criteria like Akaike's information criterion (AIC) or the Bayesian information criterion (BIC). These numbers can then be averaged across all models or the models within a time window of interest. At 400 ms, the sum of the squared error, averaged across the per-subject models, is indeed lower for the log-Cloze model than for the untransformed cloze model ($15613.94 < 15626.37$).

When modelling EEG signals as a function of stimulus properties we are of course not limited to simple regression, i.e. linear models with an intercept and a single additional predictor. To include a second predictor, the regression equation is simply extended with more predictor terms through addition (Equation A.5).

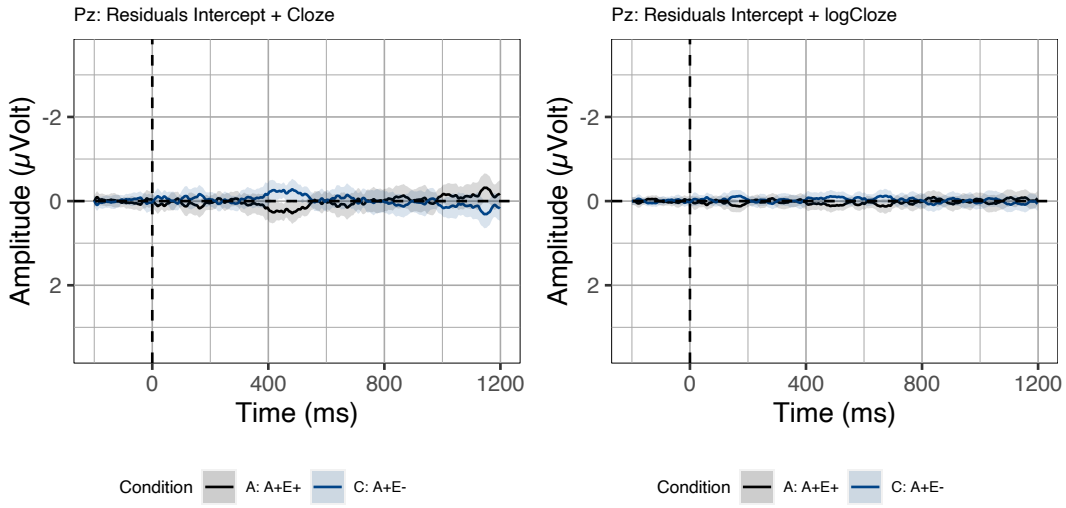


FIGURE A.6: Residual error for rERP analyses with untransformed (left) and log-transformed (right) cloze probability as predictor.

$$\hat{y}_{ts} = \beta_{0ts} + \beta_{1ts}x_1 + \beta_{2ts}x_2 \tag{A.5}$$

For instance, the example experiment from Chapter 3 actually used a two-by-two design crossing expectancy, measured by cloze probability, and lexical association, measured by ratings on a seven-point scale. Figure A.7 (left) displays the per-condition averages of the four resulting conditions, in which expectancy is low in conditions C & D and association is low in conditions B & D. Modelling the influence of Cloze and association (Equation A.5), we obtain a set of best-fit coefficients which we can visualise across time (Figure A.7, right). This allows us to trace the strength of any given predictor across time. As we saw before, the intercept term is equal to the average of the data, yielding the average ERP waveform. Importantly, in the visualisations shown here, the coefficients are always added to their intercept term, which expresses them relative to the average of the data rather than relative to zero.[3] As a result of this addition and the z-standardisation of the predictor values, the coefficient waveform shows the predicted waveform for predictor values 1 standard deviation above the mean, as this equals $\beta_j * 1$. The coefficients show that lower association ratings predict a small negative wave in the N400 time window (300 ms - 500 ms) and lower cloze probability predicts both a more negative N400 and a more positive P600 (from around 600 ms on), relative to the intercept. To achieve a more legible coefficient graph, both predictors have been inverted, such that the highest cloze probabilities are now the lowest values. As a result of this, the coefficients for

[3]Note here, that the error ribbons around the coefficients are computed differently than the EEG data error ribbons. Here, we rather compute the standard error of each individual model by following the standard equation used in regression and then use the average standard error across subjects to visualise the variability in coefficients.

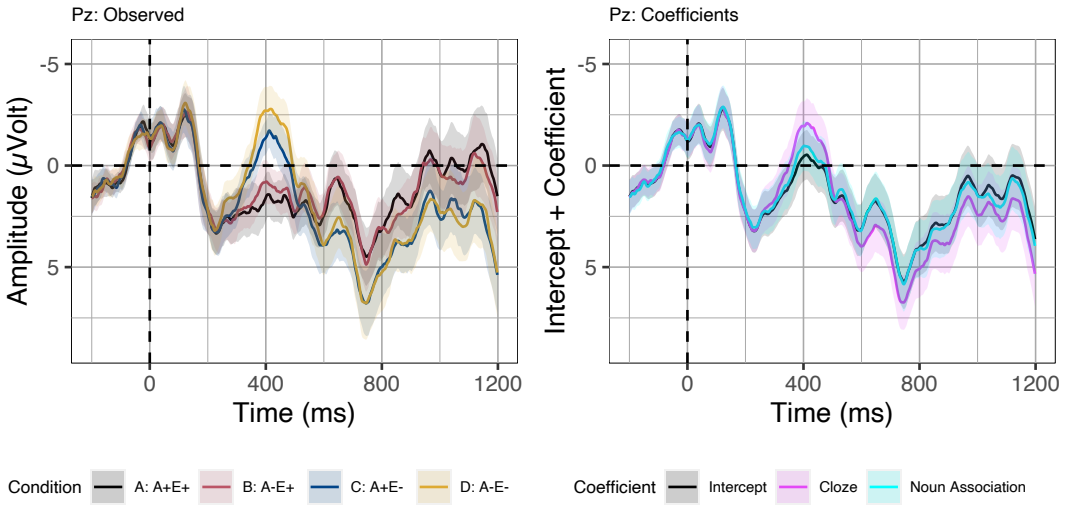


FIGURE A.7: Observed data of Design 1 showing N400 differences induced by expectancy (E- vs. E+) and lexical association (A- vs. A+) and a P600 effect of expectancy (E- vs. E+). Coefficients for an rERP analysis with cloze probability and lexical association ratings as predictors.

the N400 time window are more negative for low cloze and low association items.

A.3 Computing Isolated Forward Estimates

We can now use the fitted two-predictor rERP models to rebuild the four-condition ERP complex step by step by computing different forward estimates. By setting both cloze probability and association to their average values - zero in the case of z-standardised predictors - our estimates for all four conditions are exactly the same (Figure A.8, first row, left columns). Next, the association coefficient is multiplied by the actual association ratings (x_2). The estimates now exhibit a negativity in the N400 time window and, relative to the intercept-only model, some of the error in the N400 time window has been reduced (second vs. first row of the second column). Next, the association values (x_2) are fixed to their average again and the cloze coefficient is multiplied by the actual cloze probability values (x_1). As a result of this, the models now separate conditions A and B from conditions C and D: The Unexpected conditions exhibit an N400 (300 - 400 ms) and a P600 (from 600 ms) effect relative to the Expected conditions (third row). As the corresponding residual graph shows, the error has already been reduced a lot relative to the intercept-only model (third vs. first row, second column). Lastly, if we provide both coefficients, we effectively capture all four conditions of the observed data very well (fourth row). We turn to an inferential evaluation of the predictors in the following section.

Using this approach we can effectively isolate the influence of different stimulus

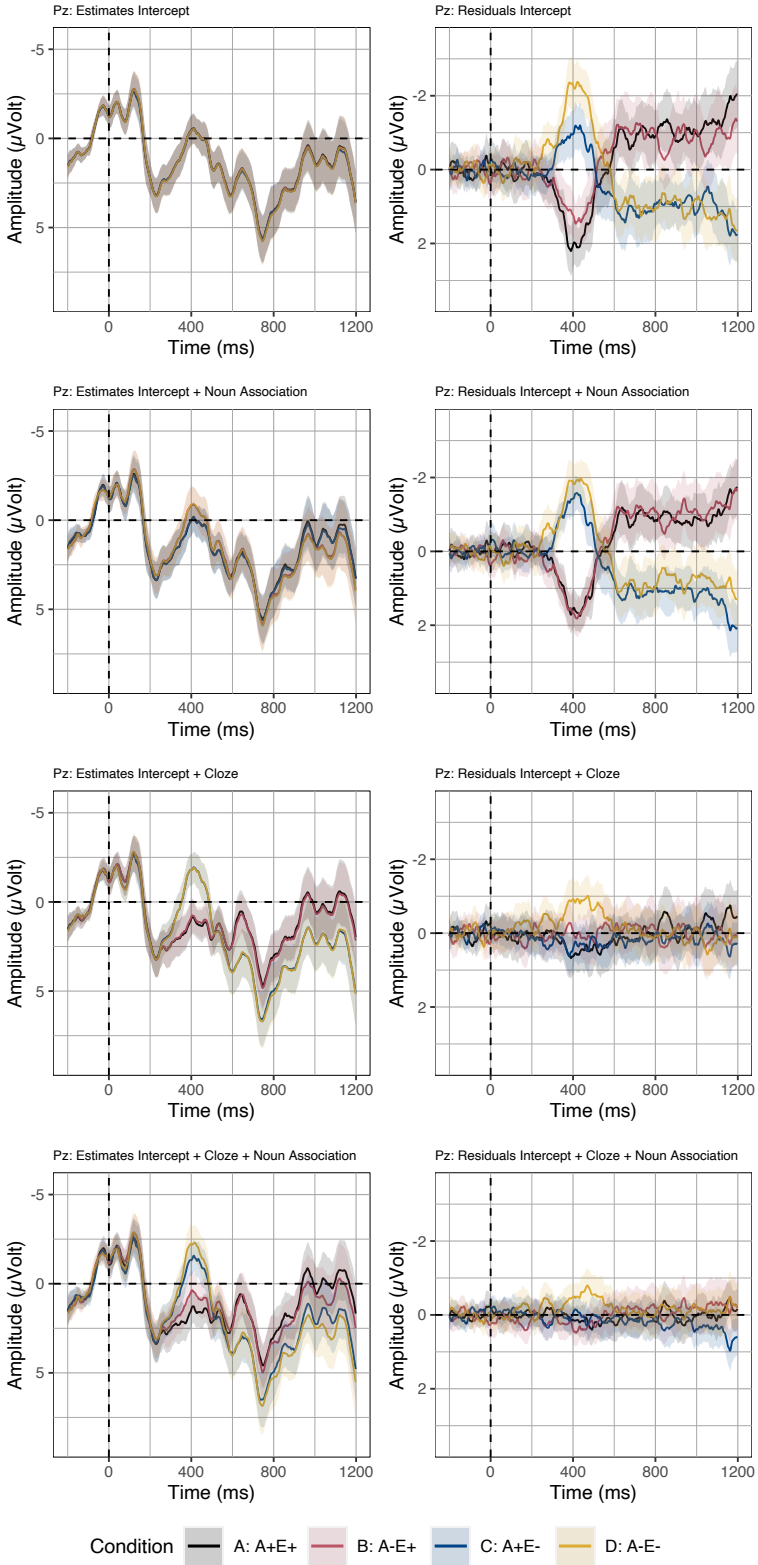


FIGURE A.8: Forward estimates (left) and residual errors (right) for models fitted using Cloze probability and noun association as predictors. Rows contain the isolated contribution of the intercept, the intercept and noun association, the intercept and cloze probability, and all three predictors together.

properties. This approach is extremely powerful in situations in which a single condition average is influenced by several stimulus properties at the same time. Further, Brouwer, Delogu, and Crocker (2021) showed that in a case where two ERP components with opposite polarity - one negative (N400) and one positive (P600) - spatiotemporally overlap and cancel each other out, rERPs can be used to estimate the underlying waveforms in isolation, given an appropriate experimental design. Whereas in the above case, we computed the isolated waveforms using the original predictor values, we can also generate the estimated waveforms of arbitrary predictor values. For instance, we may be interested in what the estimated waveform would look like for an item with either average, 1 standard deviation above average, maximal, or minimal cloze probability. Figure A.9 displays these estimates based on models that were fitted on the baseline condition of Aurnhammer et al. (2021) in which the target word varied in expectancy. This analysis demonstrates that rERPs provide a tool to go beyond experimental conditions and explore the variability within a single condition. Hence, this technique is also suitable for data that does not have experimental conditions, such as naturalistic language comprehension data (cf. Frank & Willems, 2017).

A.4 Inferential Statistics

Researchers developing ERP studies are often interested in assessing whether the predictors used in the model are statistically significant or not. These inferential statistics are derived from the coefficients. The t-values are computed by dividing the coefficients by their standard error, resulting in a measure of effect size. We can then compare t-values to the t-distribution, obtain a p-value, and – applying an arbitrary threshold – make a statement about the statistical significance of the predictor. In the case of rERPs, we obtain a t-value for each coefficient in each model. If we want to visualise the t-values and p-values across time, we realise that we now have a vector of inferential statistics for each time sample, as models were fitted separately for each subject. One solution to this problem is to fit an across-subjects regression model at each time sample, i.e. not fitting a separate model for each subject. Note that as a result of this, the varying amount of data for each subject leads to coefficients that are not identical to those obtained by computing the average coefficients from the within-subjects approach. Another solution is to use linear-mixed effects regression (LMER) and model the variability across subjects as random effects - an extension that is expanded upon below. Using either solution, we obtain just a single t-value (or z-value in the case of LMERs) and p-value for each time sample and we can now visualise these values across time (Figure A.10).

A different problem is that there are still many null hypothesis tests and many p-values. This poses a multiple comparisons problem, which needs to be addressed in order to control for the inflated false discovery rate. A simple way to correct this

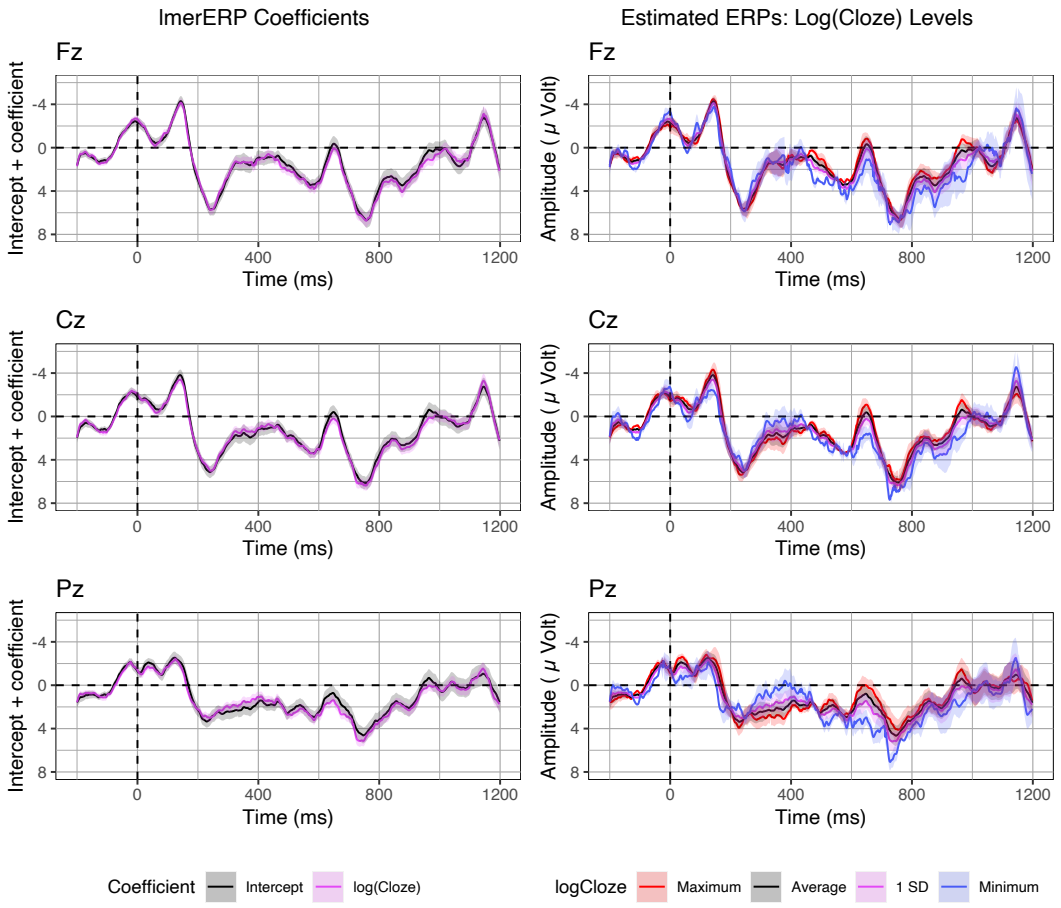


FIGURE A.9: Modelling the baseline condition of Design 1 by Cloze probability. Coefficients (left) and the estimated waveforms for different Cloze probability values (analysis based on LMERs; see Section A.6).

is to apply multiple comparisons correction, for instance following the method proposed by Benjamini and Hochberg (1995). Note that this approach treats all p-values independently, even though adjacent time samples are correlated. Further, applying this method to all time samples may lead to a very strong correction. As most ERP studies that employ null hypothesis statistical testing make predictions about specific time windows of interest, a sensible middle ground is to treat the time samples within this time window of interest as pertaining to one family of hypotheses, to correct p-values within time window and simply disregard the p-values on all other time samples. It is worth noting that there are considerable researcher degrees of freedom in making decisions about which p-values to treat as one family, which p-values to treat separately, and which to disregard. Indeed, while these decisions could be exploited for p-hacking and other questionable research practices, this is an issue that affects multiple comparison problems in general and is not specific to rERPs in any way. The specific strength of rERPs lies not in assessing statistical significance but in the ability to build an understanding of how different stimulus

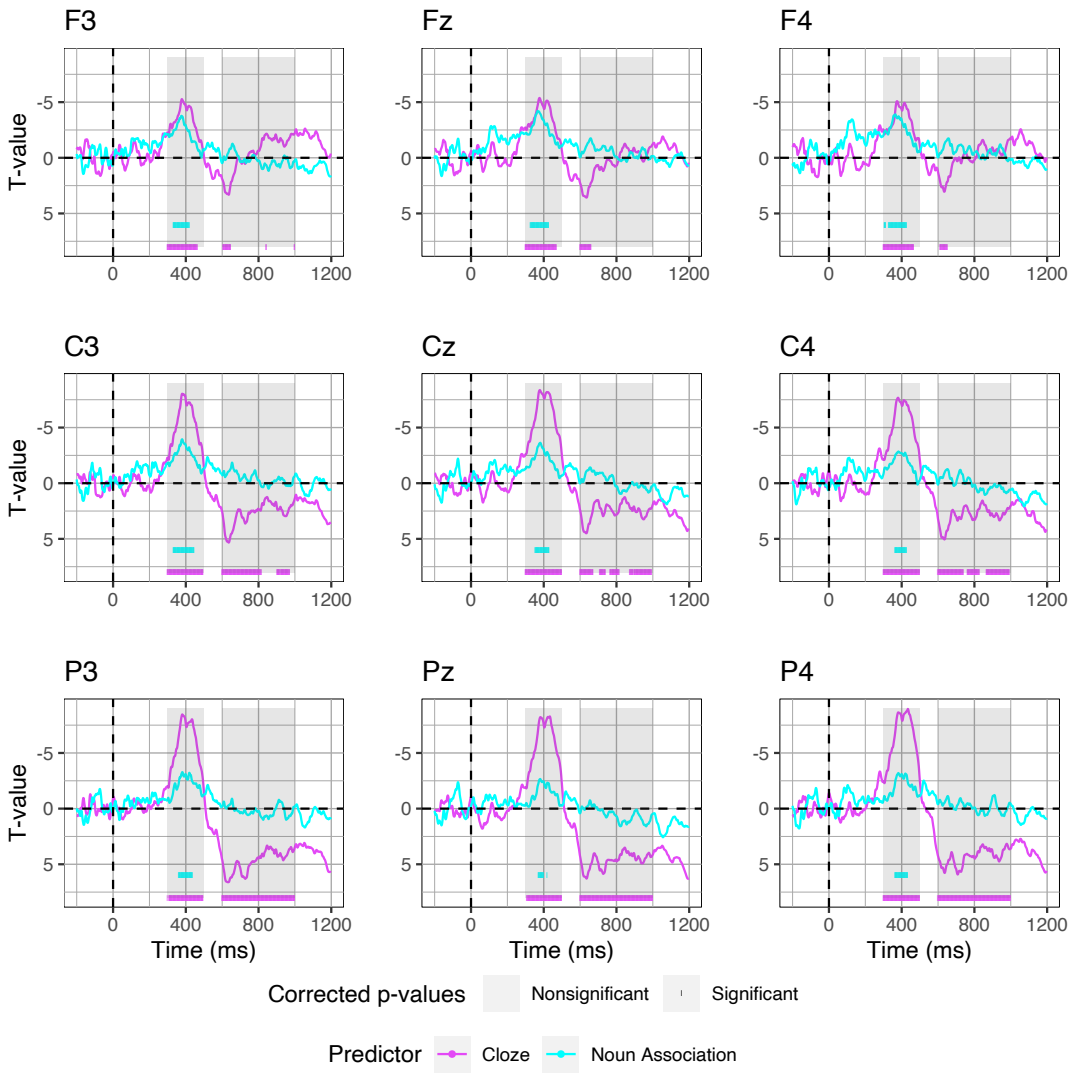


FIGURE A.10: Inferential statistics for the predictors cloze probability and noun association on nine electrodes, computed from regression across subjects. Bars indicate time samples with significant p-values after multiple comparisons correction.

properties shape the observed ERP in isolation and in combination at full temporal resolution. That said, for the data presented in Chapter 3, the inferential statistics indicate that after multiple comparisons correction, both Cloze probability and noun association significantly predict the ERP signal in the N400 time window, whereas in the P600 time window, only Cloze probability is significant when considering linear models fitted across subjects (Figure A.10).

A.5 Modelling Scalp Distributions

ERP components are often characterised by specific scalp distributions. That is, voltage deflections are stronger on some groups of close-by electrodes than on others. In traditional analyses, for instance, using analysis of variance (ANOVA), special

predictors are created that contrast some electrode positions against others (e.g., anterior-posterior or laterality contrasts). When using rERPs, we do not need to change the model specification itself. Rather, we treat the different electrodes just like we treated the time samples. That is, we fit one model for each subject, at each time sample and each electrode. Reflecting this, we add an e to our subscript notation (Equation A.6).

$$\hat{y}_{ets} = \beta_{0ets} + \beta_{1ets}x_1 + \beta_{2ets}x_2 \tag{A.6}$$

Investigating differences across the scalp can be as simple as inspecting the estimates, residuals, coefficients, and t-values at different electrodes or on the entire electrode grid (Figure A.11). With regard to assessing statistical significance, we must be aware that the number of comparisons has increased even more. Just as we did with the time windows, it is probably sensible to select electrodes of interest within which to apply multiple comparisons correction to avoid arriving at a too conservative correction scheme.

When we are interested in scalp distributions of ERPs we often turn to a different visualisation technique, the topographic map. To compute a topographic map, we first calculate the difference between the voltages of two data sets, most commonly by subtracting the baseline condition from a manipulated condition, similar to computing what is known as a difference wave. Second, the difference in voltages on each electrode is averaged across several time samples, which typically fall into a time window of interest, such as the canonical N400 or P600 time windows. In the last step, we take the per-electrode and within-time-window voltage differences and apply an interpolation algorithm to estimate the voltages for the spaces between the actual electrodes. Applied to our example data, we can visualise the scalp distribution of the N400 effect observed for the Unexpected relative to the Expected condition by computing their difference, averaging across the time samples between 300 and 500 ms, and interpolating (Figure A.12).

In order to leverage this visualisation technique within the rERP framework, all we need to do is repeat the model fitting process on all the electrodes that should be included in the topographic map (typically all non-reference, non-eye electrodes). We can then draw topographic maps of coefficients, estimates, residuals, and t-values. As a concrete use case, rERPs can be helpful to study how several ERP modulations combine within a single condition contrast. The estimates can be used to decompose the total waveform of a single condition into the contributions made by different predictors, as we saw in Figure A.8 where condition D was modulated by both cloze probability and lexical association. This translates directly into topographic maps.

For instance, in Chapter 4, we leveraged rERPs to demonstrate that the topographic map of a single condition contrast with two manipulations can be broken down into the underlying topographic maps elicited by the two manipulations. In

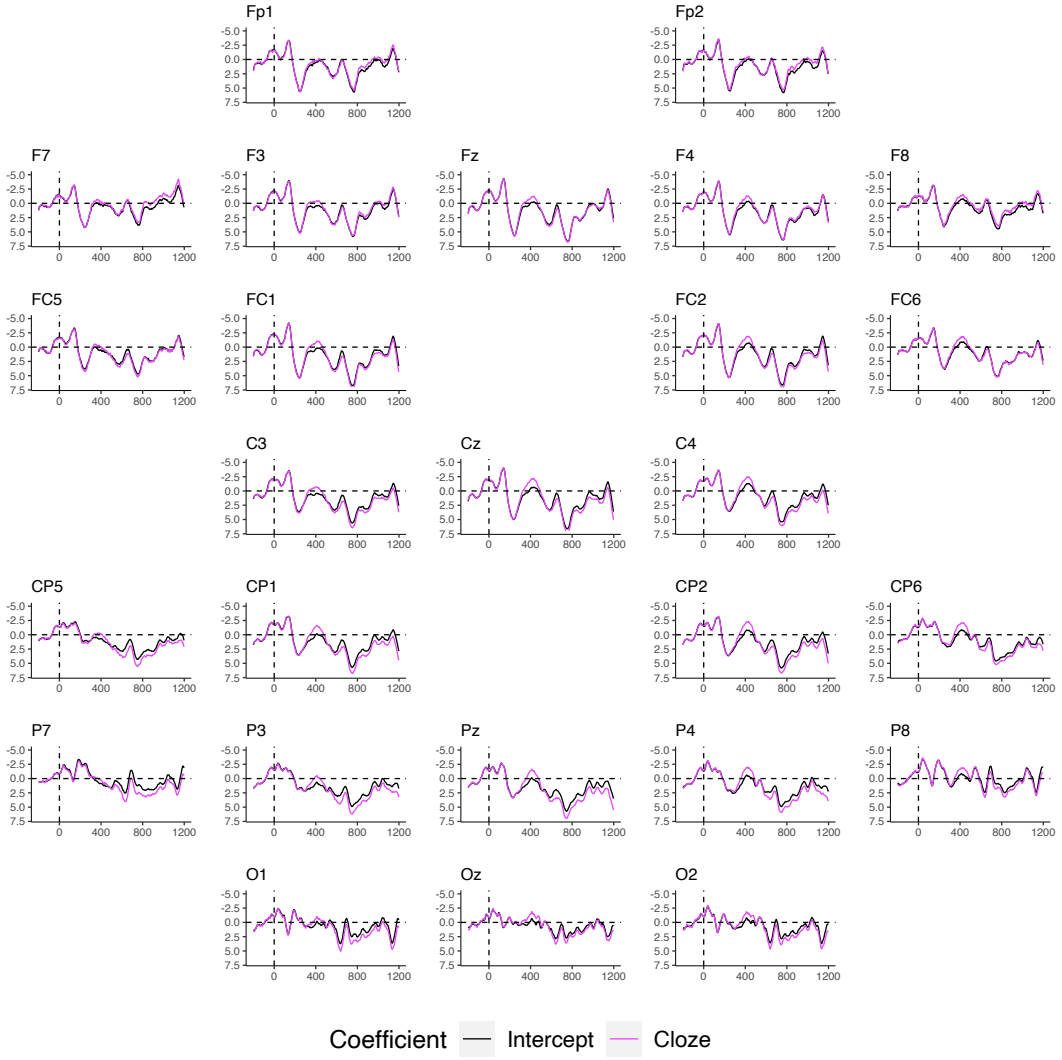


FIGURE A.11: Electrode grid with coefficients for the intercept and Cloze probability fitted on the data of two conditions manipulating expectancy (Design 1, Condition A vs. C). Standard errors around the coefficients are not shown for legibility.

this design, we first presented a context paragraph telling a story about a tourist at an airport who wants to take his huge suitcase onto the airplane. This context paragraph is then followed by one of two continuations. The control condition continued the story in a plausible manner: "Next, the lady dismissed the tourist". In the manipulated condition, the continuation was less plausible: "Next, the lady weighed the tourist". In the latter condition, the target word was not only less plausible than in the control condition, but additionally, the context raised expectations for the distractor word "suitcase", which was also introduced in the context paragraph. Hence, any obtained difference in the ERPs found for this contrast may be due to both target plausibility and distractor expectancy.

The topographic map of the observed data revealed a broadly distributed late positivity with several peaks (Figure A.13, right). Plausibility was modelled using

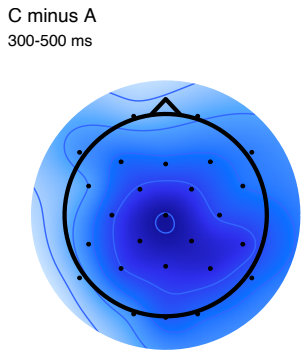


FIGURE A.12: Topographic map of the difference between low and high expectancy conditions in the N400 time window (300 - 500 ms).

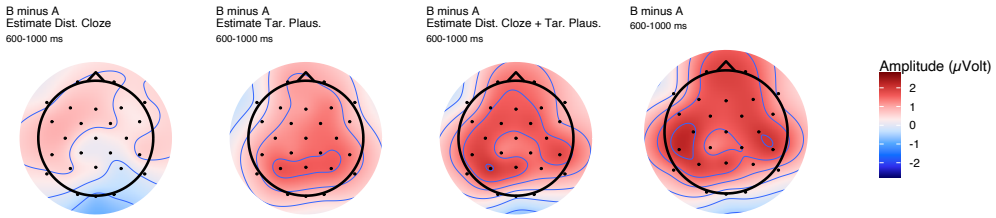


FIGURE A.13: Topographic maps of the estimated contributions of two predictors in isolation (left, middle left), their additive contribution (middle right), and the observed data to which the models were fitted (right).

plausibility ratings (1-7, where 7 is entirely plausible) and distractor expectancy was modelled using cloze probability values for the word "suitcase". Computing the forward estimates of the models with these two predictors and then visualising the estimates as topographic maps, we found that the observed estimated map for this condition contrast (Figure A.13, middle right) closely resembles the observed data (right). When isolating the forward estimates of one predictor by setting the other predictor to its average value, the topographic maps suggest that the estimated topography is the sum of a frontally and a parietally peaking sub-component which are predicted by the two independent variables - target word plausibility (middle left) and distractor cloze probability (left). Recall again that the predictors in the individual regression models are strictly additive, which means that the summed contribution of the two predictors is equal to the estimates generated when both predictors are active. Hence, the rERP method can be used to isolate the underlying topographic maps in situations where several components are active at the same time, given predictors that correlate with these sub-components.

A.6 Linear Mixed Effects Regression-Based ERPs

Statistical analyses typically have the goal to make statements about the population, rather than just the specific sample that was collected during data acquisition. Linear mixed effects regression (LMER) models allow to separate the variability that is specific to certain groups in the data from the variability that is present across those groups. In a typical psycholinguistic experiment, the subjects who took part in the study make up one grouping. In an ERP study, the N400 amplitudes of some subjects may be larger than those of others, in general. If such groupings are present in the data, the observations are said to be non-independent, potentially violating one of the assumptions of many statistical methods. These differences in average potential can be addressed in LMERs by additional intercepts fitted for each individual subject, yielding the difference of that subject's average to the mean across items,[4] which are called per-subject random intercepts in LMER terminology. Further, the N400 response to expectancy could result in larger N400 modulations in some subjects than in others. This can be captured as a coefficient for expectancy which is computed for each individual subject (across items). In LMER terminology, these coefficients are called per-subject random slopes. In fact, rERPs already achieve something very similar to per-subject intercepts and slopes, by fitting a separate model for each participant. Using LMER models, however, we can replace the separate models for each subject with a single model in which we capture subject-specific variability using random intercepts and slopes. Hence, only one model is computed for each electrode and time sample. In the regression equation, the random intercepts S_{0et} and random slopes S_{1et} are simply added to yield Equation A.7.

$$\hat{y}_{et} = \beta_{0et} + S_{0s} + (\beta_{1et} + S_{1s})x_1 + (\beta_{2et} + S_{2s})x_2 \tag{A.7}$$

In a typical psycholinguistic experiment, the data may not only be stratified by participants but also by the specific experimental items that were presented. Using LMERs, we can also separate the variability that is present across items from that which is specific to the individual items. Again, to achieve this, the per-item random intercepts (I_{0et}) and per-item random slopes (I_{1et}) are added to yield Equation A.8.

$$\hat{y}_{et} = \beta_{0et} + S_{0s} + I_{0i} + (\beta_{1et} + S_{1s} + I_{1i})x_1 + (\beta_{2et} + S_{2s} + I_{2i})x_2 \tag{A.8}$$

Using *lmerERPs* addresses one of the problems that we encountered while assessing statistical significance using rERPs, namely that for each time sample and electrode, there are vectors of t-values and p-values – one for each subject – which made visualisation difficult. In contrast, lmerERPs directly yield a single z-value and p-value for each predictor at each electrode and time sample, which can thus

[4] Note that due to a phenomenon in LMERs called "shrinkage" the per-grouping random intercepts are pulled to the fixed effect intercept. Thus, summing the fixed effect intercept and a per-grouping random intercept is not equal to the average of the dependent variable of that grouping level.

be visualised directly as shown in Figure A.10. Further, the coefficients on which the z-values and p-values are based have been adjusted for subject and item-specific variability.

However, lmerERPs are not without disadvantages. In least-squares regression, there is a single set of optimal coefficients[5] which are determined algebraically. In LMERs, on the other hand, model parameters are optimised using an iterative procedure, which tends to take more time. Further, this process is non-deterministic which means that there is no single set of correct model parameters and the final set may vary when applying different optimisation algorithms or even when applying the same algorithm twice. On a more pragmatic note, the faster computation speed of multiple regression may be considered preferable if we want to fit models on all electrodes, at high temporal resolution or simply to quickly test different predictor combinations.

A.7 Beyond ERPs: Regression-Based Reading Times

By repeatedly fitting linear models with the same predictor specification, the rERP method allows us to capture variability in predictor strength across relevant dimensions of the dependent variable, such as time samples and electrodes. Indeed, there are many experimental paradigms yielding dependent variables that extend over one or more dimensions, and indeed sometimes analysis techniques very similar to rERPs are employed to capture variability across these dimensions, such as the per-voxel regression analyses applied to functional magnetic resonance imaging (fMRI) data (see Smith & Kutas, 2015a, for discussion of similar analysis approaches). In this section, we will demonstrate how the generalised rERP approach can be applied to self-paced reading, a common behavioural paradigm in psycholinguistics research.

In self-paced reading studies, participants read a sentence word-by-word and press a button to proceed to the next word. Some words take longer to read than others and, typically, these variable reading times are the dependent variable of the subsequent analyses. For instance, in the above example, the reading time of the target word "axe" is slowed in the sentence "Yesterday ate the lumberjack the axe" compared to "Yesterday sharpened the lumberjack the axe" (Figure A.14, Critical Region). Critically, self-paced reading studies are characterised by a Spillover effect, meaning that the processing cost triggered by the target word often becomes manifest in increases in reading times on the following words. In the example, reading times are indeed increased on the two words following the target word "axe" (Figure A.14, Spillover and Post-Spillover region). Additionally, the Pre-Critical region is also shown to check whether there are any reading time differences prior to the presentation of the target word. In particular, context manipulations can lead to differential reading speed before target word presentation. In Chapter 3, we observed

[5] Assuming a strictly convex objective function.

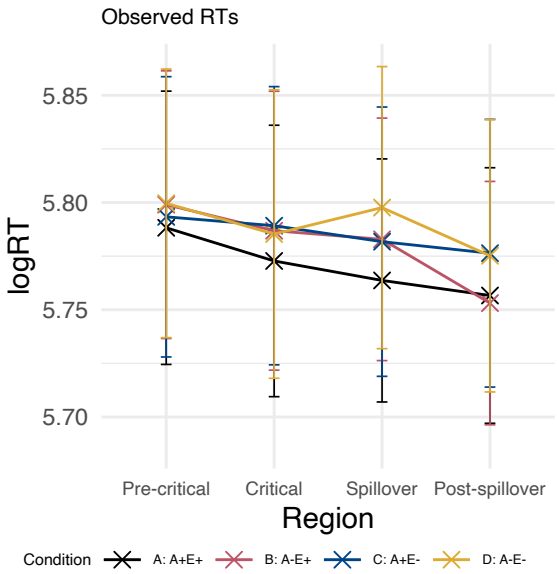


FIGURE A.14: Reading times observed in a self-paced reading experiment (Design 1) on four regions of interest.

differential reading time modulations across regions as a function of expectancy and lexical association. For Design 1, all manipulated conditions were slowed on the Critical region (B, C, D relative to A). The Spillover region showed additive effects of lexical association (B & D vs. A & C) and expectancy (C & D vs. A & B), whereas only the expectancy effect remained on the Post-Spillover region.

Because stimulus properties affect reading time differentially across regions, reading times can be considered as a temporally extended dependent measure, and an approach very similar to rERPs can be applied, which we name *rRTs*. To capture differences across reading time regions, a separate regression model is fitted at each region. Using the model specification with lexical association ratings and cloze probabilities as predictors (Equation A.5), we can inspect on which regions the predictors account for the data to what extent. As before, we can then plot each model's coefficients, which in this case are distributed across regions (Figure A.15, left).

Similarly, the coefficients resulting from the separate models can be used to compute the forward estimates and the residuals of the predictors in isolation and in combination (Figure A.16). The visual decomposition of the forward estimates can help to understand how and how well the statistical models explain the data. For instance, the residual error for the full model including both cloze probability and noun association (Figure A.16, last row) indicates that Condition B is not captured accurately on the critical region. One option to address this would be the inclusion of a multiplicative interactive term between cloze probability and noun association.

To assess the significance of the predictors, t-values and p-values can be computed from the models (Figure A.15, right). As before, this approach contains several independent null hypothesis tests, and the researcher must decide whether the separate reading time regions pertain to separate families of hypotheses or not. In

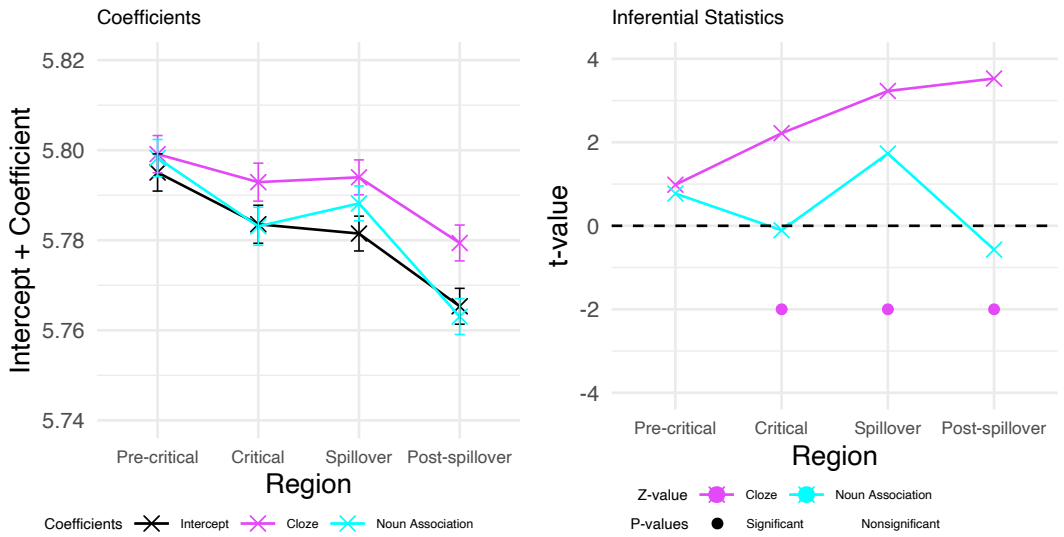


FIGURE A.15: Coefficients (left) and inferential statistics (right; t-values and p-values) for the predictors Cloze probability and noun association across regions.

the latter case, multiple comparisons correction should be applied.

Many other experimental paradigms record a dependent variable that is observed over several dimensions for each trial. It is worth thinking about how these dimensions can be modelled in the linear modelling framework instantiated by rERPs.

A.8 Summary

The foundation of the rERP method is the observation that ERP averaging is mathematically equivalent to computing a series of intercept-only regression models. In the words of Smith and Kutas (2015a, p. 158): "All ERPs *are* rERPs". Taking this perspective, we can extend our set of regression equations to include more predictors. These predictors can directly estimate the difference between experimental conditions, or they can be measures of the underlying stimulus properties themselves, capturing systematic variation around the average of the data, represented by the intercept. Visualising the resulting coefficients as waveforms allows us to trace when – across time samples – and where – across electrodes – the ERP signal varies systematically with the predictors. Further, we can use the coefficients to generate the forward estimates for predictors in isolation or in combination. This allows us to visualise what the condition-averaged waveforms would look like if only one of the stimulus properties is varied while the influence of the remaining properties is held constant. An important tool to understand the performance of the regression models is their residual error. Computed per condition and across time samples, the error can give insight into portions of the data in which the regression models

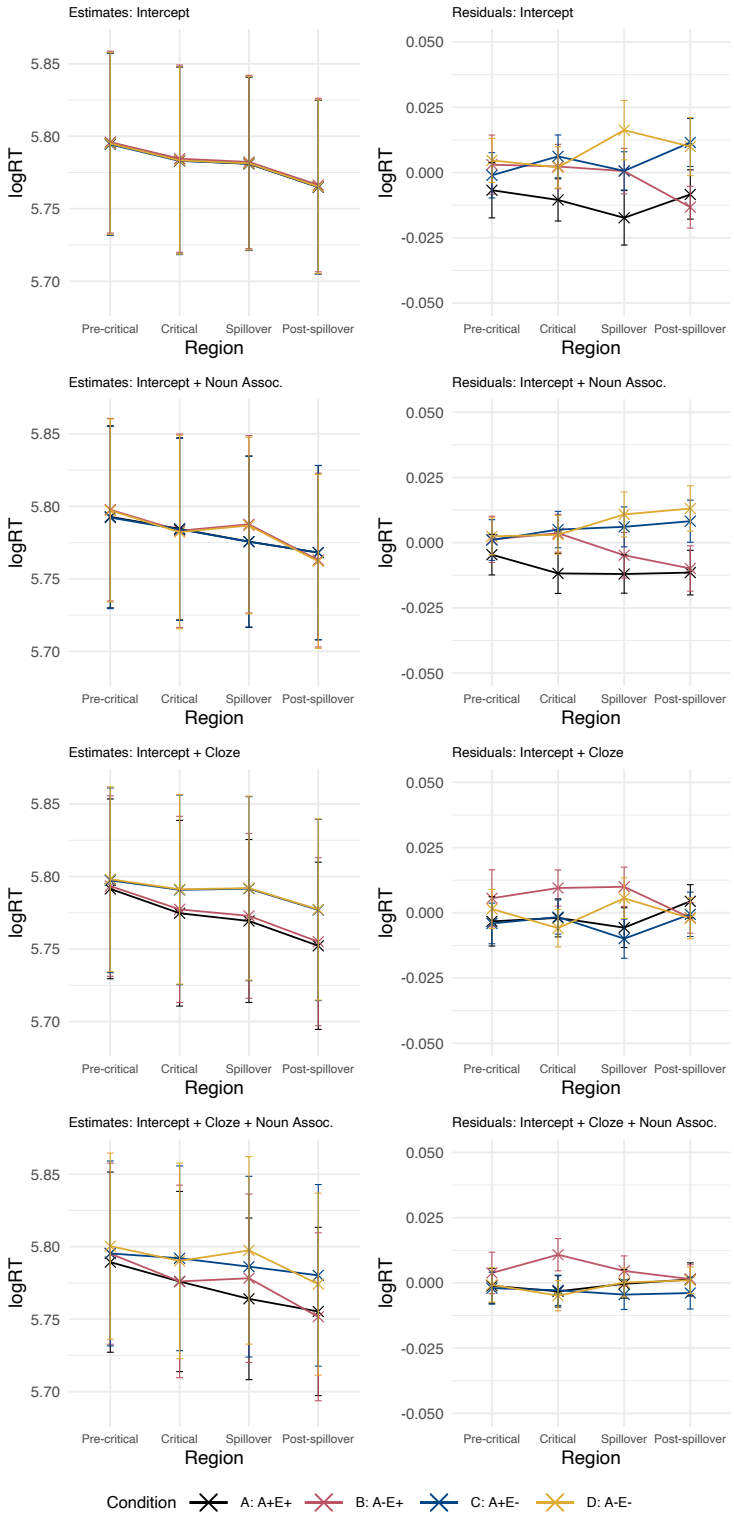


FIGURE A.16: Forward estimates (left) and residual error (right) for models fitted using Cloze probability and noun association as predictors. Rows contain the isolated contribution of the intercept, the intercept and noun association, the intercept and cloze probability, and all three predictors together.

over- or underestimate the observed data. The models also allow us to derive inferential statistics – t-values and p-values. This results in a multiple comparisons problem which must be addressed using adequate correction methods. All of the above can be visualised using the full array of EEG visualisation techniques, such as waveforms, differences waves, topographic maps, and so forth.

Further, the traditional regression approach can easily be extended to linear mixed effects models, allowing the inclusion of random effects for subjects, items, or any other grouping factor. Lastly, the approach underlying rERPs – repeating the same model fitting process across the dimensions over which the dependent variable is observed – can be applied to many other types of experimental paradigms (Smith & Kutas, 2015a).

A `julia` (Bezanson et al., 2017) implementation of the rERP technique is provided in the thesis repository:
https://github.com/caurnhammer/AurnhammerThesis.
The repository also contains visualisation functions written in `R`. Note that this code is not published as a package. Rather, the code is delivered as-is and works under the specific language and package versions detailed in the repository README. There will not be long-term support.

Appendix B

Stimuli

B.1 Materials of Design 1

1. Gestern [schärfte/aß] der Holzfäller, [bevor er das Holz stapelte / bevor er der Film schaute], die Axt und hackte die Holzscheite.

2. Nachdenklich [schärft / trinkt] der Barbier, [nachdem er den Rasierschaum aufgetragen hat / nachdem er die Topfpflanze gegossen hat], das Messer und rasiert den Kunden.

3. Sogleich [süßt / putzt] der Kellner, [nachdem er die Bestellung aufgenommen hat / nachdem er das Radio angemacht hat], den Kaffee und gießt die Milch dazu.

4. Vorsichtig [erhitzt / schlürft] der Bäckerlehrling, [nachdem er die Brötchen geformt hat / nachdem er die Fenster gekippt hat], den Ofen auf 180 Grad.

5. Zufrieden [entkorkt / glättet] der Winzer, [der die Reben geschnitten hat / der die Lohnabrechnung beendet hat], die Weinflasche für die Weinprobe.

6. Umgehend [löst / knotet] der Fahrgast, [der den Automaten bedient / der die Einkaufstasche abstellt], das Ticket für die Fahrt.

7. Am Abend [feiert / knabbert] der Gewinner, [der den Pokal erhält / der das Eis isst], den Sieg mit einer Party.

8. Endlich [fängt / baut] der Angler, [der den Köder auswarf / der die Zeitung las], den Fisch für das Abendessen.

9. Am Nachmittag [jätete / polierte] der Schrebergärtner, [der das Beet pflegte / der die Sonne genoss], das Unkraut und leerte den Kompost.

10. Eine Weile [lüftet / bastelt] der Lehrer, [bevor er die Tafel beschreibt / bevor er den Mantel aufhängt], das Klassenzimmer und nimmt ein Stück Kreide.

11. Schnell [reibt / flickt] der Chefkoch, [während er die Nudeln kocht / während er die Nachrichten hört], den Käse und gibt Öl in die Pfanne.

12. Vorgestern [probierte / bemalte] der Braumeister, [nachdem er den Kessel ausgeschaltet hatte / nachdem er den Boden gewischt hatte], das Bier das zuvor gebraut wurde.

13. Eine Zeit lang [mähte / räucherte] der Gartenhelfer, [nachdem er den Garten umgegraben hatte / nachdem er die Garage aufgeräumt hat], den Rasen hinter der Villa.

14. Achtsam [pflasterte / salzte] der Bauarbeiter, [der die Absperrungen umging / der die Steine schleppte], die Straße und den Bürgersteig.

15. Gestern [läutete / faltete] der Priester, [nachdem er den Kirchturm erklommen hatte / nachdem er die Armbanduhr ausgezogen hatte], die Glocken und freute sich.

16. Rasch [näht / versteckt] der Notarzt, [der den Verletzten betreut / der den Ohrring trägt], die Wunde mit fünf Stichen.

17. Eilig [bügelt / salzt] der Geschäftsmann, [nachdem er das Bügelbrett aufgestellt hat / nachdem er den Eistee ausgetrunken hat], das Hemd und die Hose.

18. Schnell [repariert / brät] der KFZ-Mechatroniker, [der den Motor ausgetauscht hat / der den Urlaub gebucht hat], das Auto und übergibt es an den Kunden.

19. Am Morgen [knetet / hört] die Bäckerin, [die den Kuchen backt / die den Stammgast begrüßt], den Teig für das Brot.

20. Am Nachmittag [pfeffert / druckt] der Grillmeister, [bevor er die Kohle anzündet / bevor er die Mücke verscheucht], das Steak und die Beilagen.

21. Schnell [entfacht / entstaubt] der Brandstifter, [der das Streichholz fallengelassen hat / der die Treppe hochgeklettert ist], das Feuer und rennt weg.

22. Nun [kämmt / speichert] der Friseur, [nachdem er die Kopfhaut massiert hat / nachdem er den Cappuccino ausgetrunken hat], die Haare und holt den Föhn.

23. Flink [spitzte / knotete] der Zeichner, [der die Karikatur entwarf / der den Atlas öffnete], den Bleistift und machte sich wieder an die Arbeit.

24. Heute [impft / feuert] der Arzt, [nachdem er die Spritze befüllt hat / nachdem er die Schublade geschlossen hat], den Patienten gegen Tetanus.

25. Unverzüglich [obduziert / pfeffert] der Pathologe, [nachdem er die Mordakte gelesen hat / nachdem er das Dokument unterschrieben hat], die Leiche in der Leichenhalle.

26. Am Nachmittag [signiert / schluckt] der Autor, [der die Geschichte erfunden hat / der den Vorhang zugezogen hat], das Buch für den Fan.

27. Sofort [begrüßte / feuerte] der Hotelier, [der die Koffer stapelte / der die Fische räucherte], die Gäste und überreichte ihnen die Schlüssel.

28. Vorsichtig [fällt / durchwühlt] der Holzarbeiter, [der die Säge gestartet hat / der den Transporter geparkt hat], den Baum neben der Schule.

29. Langsam [paffte / öffnete] der Kubaner, [der den Rauchring blies / der die Straßenbahn verpasste], die Zigarre und schlenderte über den Marktplatz.

30. Fröhlich [paniert / baut] der Koch, [nachdem er die Bratkartoffeln geschält hat / nachdem er die Ärmel hochgekrempelt hat], das Schnitzel für das Gericht.

31. Umsichtig [glasiert / durchsucht] der Konditor, [der die Backform eingefettet hat / der den Tisch abgewischt hat], den Kuchen für die Geburtstagsfeier.

32. Gestern [ölte / beheizte] der Fahrradfahrer, [bevor er die Bremse einstellte / bevor er die Limonade trank], die Kette und die Bremsen.

33. Hastig [frankiert / umgeht] der Postbeamte, [der das Postamt aufgeschlossen hat / der das Mittagessen bestellt hat], den Brief an die Agentur.

34. Fröhlich [erklimmt / kauft] der Wanderer, [nachdem er das Tal durchquert hat / nachdem er die Stoppuhr gestartet hat], den Berg und stellt sein Zelt auf.

35. Bedächtig [erteilt / biegt] der General, [der die Rekruten kommandiert / der die Getränke einschenkt], den Befehl zum Abmarsch.

36. Sofort [streicht / lenkt] der Maler, [nachdem er die Tapete angebracht hat / nachdem er den Hund gefüttert hat], die Wand des Büros.

37. Direkt [schwingt / speichert] der Schmied, [der den Amboss aufgestellt hat / der den Stuhl weggestellt hat], den Hammer auf das Metall.

38. Vor langer Zeit [regierte / briet] der König, [der die Reise unternahm / der den Gesang vernahm], das Land im Norden.

39. Zügig [hob / verdrückte] der Dirigent, [der das Tempo wechselte / der das Gähnen unterdrückte], den Stab und gab den Takt an.

40. Direkt [trimmt / hört] der Herrenfriseur, [nachdem er die Klinge ausgepackt hat / nachdem er das Wasserglas abgestellt hat], den Bart des Mannes.

41. Gestern [spülte / glättete] der Tellerwäscher, [bevor er den Tisch wischte / bevor er das Taxi rief], das Geschirr und die Brettchen.

42. Im Handumdrehen [knackt / kopiert] der Einbrecher, [der den Zahlencode erraten hat / der die Akte durchblättert hat], das Schloss und stiehlt die Wertsachen.

43. Routiniert [hackte / schraubte] der Küchengehilfe, [bevor er die Suppe umrührte / bevor er das Radio einschaltete], die Zwiebeln und den Knoblauch.

44. Umsichtig [schmolz / würzte] der Kerzenzieher, [nachdem er den Docht gekürzt hatte / nachdem er das Nähkästchen zugeklappt hatte], das Wachs und fuhr mit der Arbeit fort.

45. Gestern [mixte / gründete] der Barkeeper, [der den Wodka geöffnet hatte / der den Boden gefegt hatte], den Cocktail für den Touristen.

46. Am Abend [dirigiert / vernichtet] der Kapellmeister, [der das Konzert eröffnet hat / der die Unterschriften gesammelt hat], das Orchester und das Streichquartett.

47. Direkt [buchstabiert / dekoriert] der Sprachschüler, [nachdem er das Buch aufgeschlagen hat / nachdem er den Pausenhof verlassen hat], das Wort und notiert es.

48. Vorsichtig [umzäunt / fälscht] der Schäfer, [der die Schafe gehütet hat / der das Mittagessen vergessen hat], die Weide mit einem Draht.

49. Nachdenklich [hält / leert] der Pfarrer, [nachdem er die Bibel aufgeschlagen hat / nachdem er die Brille aufgesetzt hat], die Predigt für die Gemeinde.

50. Unverzüglich [möblierte / erntete] der Vermieter, [der den Mieter aufgenommen hatte / der die Überweisung erhalten hatte], die Wohnung mit neuen Möbeln.

51. Sofort [betätigt / knickt] der Killer, [der das Ziel anvisiert hat / der den Tee abgestellt hat], den Abzug und trifft das Opfer.

52. Heute [überbringt / erntet] der Bote, [als er den Empfänger angetroffen hat / als er das Foyer durchquert hat], das Paket mit der Arbeitskleidung.

53. Schnell [schoss / spürte] der Fußballspieler, [der das Fußballfeld überquert hatte / der die Jacke ausgezogen hatte], das Tor und jubelte über seinen Erfolg.

54. Konzentriert [steuerte / faltete] der Pilot, [der die Stewardess gerufen hatte / der den Krimi gelesen hatte], das Flugzeug durch die Wolken.

55. Rasch [entwarf / schüttelte] der Architekt, [bevor er den Bauplan erstellte / bevor er den Vortrag schrieb], die Skizze für den Bauherrn.

56. Geschwind [kehrt / wirft] der Schornsteinfeger, [der den Rauch einatmet / der die Katze verscheucht], den Kamin und verlässt das Grundstück.

57. Vorgestern [gewann / reparierte] der Politiker, [der die Stimmen zählte / der die Hände schüttelte], die Wahl zum Minister.

58. Sofort [schoss / verspeiste] der Reporter, [der die Kamera bediente / der die Zigarette rauchte], das Foto von der Unfallstelle.

59. Unverzüglich [entwickelte / trocknete] der Informatiker, [der den Computer hochfuhr / der den Energydrink trank], das Programm für seinen Boss.

60. Nun [verkündete / roch] der Richter, [nachdem er den Angeklagten hereingerufen hatte / nachdem er die Lesebrille aufgesetzt hatte], das Urteil und schloss die Verhandlung.

61. Zunächst [bestand / kochte] der Student, [der das Studium wiederaufnahm / der das Jackett auszog], die Prüfung in Mathematik.

62. Konzentriert [sang / reparierte] der Sänger, [der das Mikrofon umklammerte / der den Hut abnahm], das Lied über die Vergangenheit.

63. Zufrieden [mischte / würzte] der Anstreicher, [der die Pinsel gewaschen hatte / der die Plätzchen gegessen hatte], die Farbe für den nächsten Auftrag.

64. Einen Moment lang [gießt / ölt] der Gärtner, [der den Dünger verteilt hat / der die Opernmusik angestellt hat], die Blumen im Blumenkasten.

65. Gestern Nachmittag [stutzte / bezahlte] der Landschaftsgärtner, [der die Gartenschere geschärft hatte / der die Sonnenbrille aufgesetzt hat], die Hecke um das Grundstück.

66. Früher [beackerte / kopierte] der Landwirt, [der den Traktor gestartet hatte / der den Sonnenhut aufgesetzt hatte], das Feld mit einem Gaul.

67. Heute Morgen [flickte / schmeckte] der Schuster, [der den Absatz ausgetauscht hatte / der die Sirene gehört hatte], den Schuh und polierte ihn.

68. Schnell [packte / rührte] der Urlauber, [als er die Klamotten gefaltet hatte / als er die Gardinen aufgezogen hatte], den Koffer für die Kreuzfahrt.

69. Sorgfältig [erntet / näht] der Imker, [nachdem er den Bienenstock erreicht hat / nachdem er das Haus verlassen hat], den Honig aus den Waben.

70. Vorsichtig [lädt / schneidet] der Jäger, [der die Munition verschossen hat / der den Campingplatz verlassen hat], das Gewehr und zielt erneut.

71. Vorsichtig [mischte / schrieb] der Maurer, [der die Backsteine aufeinanderlegte / der die Nachbarn belauschte], den Mörtel für die Mauer.

72. Vorgestern [mahlte /verzierte] der Müller, [der die Mühle betrieb / der das Lied pfiff], das Mehl aus dem Korn.

73. Vorsichtig [melkte / kleidete] der Viehzüchter, [nachdem er den Stall ausgemistet hatte / nachdem er das Gemüse sortiert hatte], die Kuh und brachte ihr frisches Heu.

74. Zuerst [bezog / fälschte] der Hotelangestellte, [nachdem er den Bezug entfaltet hatte / nachdem er den Hamburger verspeist hatte], das Bett und das Kissen.

75. Augenblicklich [erreicht / reinigt] der Sprinter, [der das Rennen gewinnt der den Wind genießt], das Ziel und durchtrennt die Ziellinie.

76. Direkt [surft / reserviert] der Surfer, [der die Badehose angezogen hat / der den Tagesausflug unternommen hat], die Welle bis zum Strand.

77. Gut gelaunt [kontrollierte / zitierte] der Schaffner, [der das Zugabteil betreten hatte / der die Mittagspause beendet hatte], die Fahrkarten der Pendler.

78. Am Mittag [verlegte / verfasste] der Elektriker, [nachdem er den Strom abgeschaltet hatte / nachdem er den Anwohner informiert hatte], die Leitung im Gebäude.

79. Langsam [hisst / verzehrt] der Matrose, [nachdem er die Taue gekappt hat / nachdem er die Handschuhe angezogen hat], die Segel des großen Bootes.

80. Direkt [entspiegelt / verschließt] der Optiker, [nachdem er den Sehtest durchgeführt hat / nachdem er die Jahresabrechnung aufgestellt hat], die Gläser und spricht mit dem Kollegen.

81. Auf der Stelle [wirft / bucht] der Basketballer, [der den Ball gedribbelt hat / der die Sitzbank verlassen hat], den Korb und holt einen Punkt.

82. Ohne zu zögern [knackt / verknotet] der Perlentaucher, [der den Meeresgrund erreicht hat / der die Taschenlampe angemacht hat], die Muschel und holt die Perle.

83. Umsichtig [spannt / tippt] der Schütze, [der den Pfeil gespitzt hat / der das Laub zusammengekehrt hat], den Bogen und feuert den Pfeil ab.

84. Schnell [verlegte / naschte] der Dachdecker, [der das Dach abdeckte / der den Ehering auszog], die Ziegel und beendete den Arbeitstag.

85. Zuerst [stapelt / serviert] der Umzugshelfer, [der den Umzugswagen gefahren hat / der die Firma gegründet hat], die Kartons und die Kisten.

86. Heute früh [schwänzte / dekorierte] der Schüler, [der die Klasse wiederholt hatte / der das Handy verloren hatte], den Unterricht und wurde erwischt.

87. Am Abend [kontrollierte / probierte] der Türsteher, [der den Geldbeutel durchsuchte / der die Nachtluft genoss], den Ausweis und ließ die Frau in den Club.

88. Umsichtig [wechselte / bastelte] der Babysitter, [der das Baby geweckt hatte / der den Obstsalat zubereitet hatte], die Windel und warf sie in den Abfall.

89. Langsam [stemmte / verdaute] der Bodybuilder, [der das Fitnessstudio betreten hatte / der das Licht angemacht hatte], die Gewichte in die Höhe.

90. Umsichtig [repariert / nascht] der Klempner, [der das Wasser ausgestellt hat / der den Kirchenchor geleitet hat], das Rohr im Badezimmer.

91. Schnell [schreibt / lenkt] der Journalist, [der die Recherche durchgeführt hat / der das Portemonnaie gefunden hat], den Artikel für die nächste Ausgabe.

92. Heute Morgen [flickte / trank] der Radler, [der die Luftpumpe benutzt hatte / der den Schrank aufgeschlossen hatte], den Reifen an seinem Rad.

93. Umgehend [entleert / verdrückt] der Postbote, [der das Postauto geparkt hat / der den Kugelschreiber verloren hat], den Briefkasten und fährt weiter.

94. Hektisch [kauft / näht] der Junkie, [der den Dealer angerufen hat / der die Kapuze aufgezogen hat], die Drogen für die nächste Woche.

95. Zügig [leerte / bastelte] der Müllmann, [der den Müllwagen geparkt hatte / der den Hausbesitzer gegrüßt hatte], die Tonne und sprang auf den Müllwagen.

96. Hektisch [knallte / verfasste] der Reiter, [der das Pferd ritt / der die Landschaft durchquerte], die Peitsche und gab die Sporen.

97. Nach kurzem Überlegen [zückte / gründete] der Gangster, [der die Bank überfiel / der den Bus betrat], die Waffe und verlangte das Geld.

98. Zufrieden [kaperte / verrührte] der Pirat, [der die Flagge hisste / der die Fehde austrug], das Schiff und übernahm das Kommando.

99. Sofort [kauft / knabbert] der Börsenmakler, [der die Börse betreten hat / der den Whiskey eingeschenkt hat], die Aktie und berichtet seinem Auftraggeber.

100. Gestern [bohrte / warf] der Handwerker, [der die Bohrmaschine hielt / der das Kaugummi kaute], das Loch in die Decke.

101. Am Morgen [misst / kauft] der Arzthelfer, [der das Messgerät hält / der den Bildschirm anstellt], den Blutdruck und den Blutzucker.

102. Konzentriert [schwang / befüllte] der Torero, [der den Stier verwundet hatte / der das Gebet aufgesagt hatte], das Tuch und machte sich bereit.

103. Angespannt [zückte / kaute] der Ritter, [nachdem er den Kampf begonnen hatte / nachdem er die Brücke überquert hatte], das Schwert und griff den Gegner an.

104. Wachsam [steuert / kopiert] der Sanitäter, [der das Blaulicht eingeschaltet hat / der den Notizblock weggelegt hat], den Rettungswagen in Richtung des Krankenhauses.

105. Ohne zu zögern [verweigerte / bestellte] der Zeuge, [der den Angeklagten wiedererkannte / der den Gehrock umklammerte], die Aussage und wendete sich an seinen Anwalt.

106. Sogleich [schnürt / zerbricht] der Eiskunstläufer, [der die Eishalle erreicht hat / der das Bonbon gelutscht hat], die Schlittschuhe und begibt sich auf die Eisfläche.

107. Gut gelaunt [schwingt / zerhackt] der Cowboy, [nachdem er das Pferd gesattelt hatte / nachdem er den Kameraden gerufen hat], das Lasso und fängt das Tier ein.

108. Entspannt [schrieb / verzehrte] der Regisseur, [der den Plot konstruiert hatte / der den Kamin angemacht hatte], das Drehbuch für die Serie.

109. Sofort [entdeckt / serviert] der Astronom, [der das Teleskop aufgestellt hat / der die Uhrzeit notiert hat], den Stern am Firmament.

110. Gestern [lichtete / kochte] der Seemann, [der das Boot kommandierte / der den Atem anhielt], den Anker und verließ den Hafen.

111. Am Vormittag [knotete / roch] der Bergsteiger, [nachdem er den Haken festgebohrt hatte / nachdem er die Krähe verscheucht hatte], das Seil um den Aufstieg abzusichern.

112. Entspannt [dreht / schlürft] der Croupier, [der den Spieltisch vorbereitet hat / der die Frau beeindruckt hat], das Rouletterad und verrät die nächste Zahl.

113. Neulich [feilte / baute] die Kosmetikerin, [während sie die Maniküre durchführte / während sie das Gespräch belauschte], die Nägel und lackierte sie.

114. Sogleich [prophezeite / würzte] der Wahrsager, [der die Kristallkugel befragt hatte / der die Kerze angezündet hatte], die Zukunft der Familie.

115. Heute [beginnt / raspelt] der Archäologe, [der die Schaufel hervorgeholt hat / der das Taxi genommen hat], die Ausgrabung in der Ruine.

116. Zügig [bindet / ölt] der Florist, [der die Rosen gezüchtet hat / der die Fenster geschlossen hat], den Strauß für die Hochzeit.

117. Umgehend [spielt / schnibbelt] der Schauspieler, [der das Skript auswendig gelernt hat / der die Torte gebacken hat], die Rolle für das Theaterstück.

118. Am Abend [moderiert / wiegt] der Moderator, [der die Zuschauer unterhält / der die Weltmeere bereist], die Sendung für den Fernsehkanal.

119. Ohne zu zögern [komponierte / grillte] der Musiker, [der die Noten aufschrieb / beantwortete], das Stück für seine nächste Sonate.

120. Fröhlich [malte / bewohnte] der Künstler, [der die Leinwand aufgespannt hatte / der den Nachtisch zubereitet hatte], das Bild in seinem Atelier.

B.2 Materials of Design 2

1. Ein Tourist wollte seinen riesigen Koffer mit in das Flugzeug nehmen. Der Koffer war allerdings so schwer, dass die Dame am Check-in entschied, dem Touristen eine extra Gebühr zu berechnen. Daraufhin öffnete der Tourist seinen Koffer und warf einige Sachen hinaus. Somit wog der Koffer des einfallsreichen Touristen weniger als das Maximum von 30 Kilogramm.
   Dann [verabschiedete / wog / unterschrieb] die Dame den Touristen und danach ging er zum Gate.

2. Ein engagierter Lehrer sah eine alte Weltkarte in der Vitrine eines Antiquitätengeschäfts. Ein solch authentisches Artefakt schien dem Lehrer sehr geeignet für sein Klassenzimmer zu sein und er sprach die Verkäuferin an. Aufgeregt fragte der Lehrer die sympathische Verkäuferin, wie viel die Weltkarte kosten sollte. Obwohl er für eine zusätzliche Weltkarte selbst bezahlen musste, sagte der Lehrer der Verkäuferin, dass er dies gerne tun würde. Die Verkäuferin sagte daraufhin, wie beschämend es sei, dass die Schule nicht einmal für eine Weltkarte bezahlen würde.
   Dann [kaufte / küsste / füllte] der Lehrer die Weltkarte und danach verließ er das Geschäft.

3. Eine Redakteurin hatte von ihrer Firma eine Streifenkarte erhalten. Mit dieser Streifenkarte konnte die Redakteurin günstig mit dem Bus zur Arbeit fahren und musste nicht jedes Mal eine Karte bei dem Busfahrer kaufen. Leider hatte die Tochter der Redakteurin eines Tages eine Zeichnung auf die Streifenkarte gemalt. Deswegen hatte die Redakteurin etwas Angst, als sie bemerkte, dass der Busfahrer heute nicht gut gelaunt war, als sie ihm die Streifenkarte überreichte.
   Dann [stempelte / beschimpfte / aß] der Busfahrer die Streifenkarte und sofort fuhr er viel zu schnell weiter.

4. Während er einen Tisch baute, brach ein Schreiner seinen schönen Hammer in zwei Teile. Der Schreiner hatte den Hammer immer gemocht. Deswegen schien es ihm eine Schande, ihn einfach wegzuwerfen. Es erschien dem Schreiner eine viel bessere Idee, den Hammer von seinem Lehrling reparieren zu lassen.
   Dann [nahm / belächelte / aß] der Lehrling den Hammer und sofort machte er sich an die Arbeit.

5. Ein Opa wollte einen Apfelkuchen bei einem Konditor kaufen. Der Konditor versicherte dem Opa, dass der Apfelkuchen heute besonders gelungen sei. Der Opa schaute auf den Apfelkuchen in der Vitrine und sah glücklich den Konditor an.
   Daraufhin [verpackte / belächelte / spülte] der Konditor den Apfelkuchen und dann wandte er sich an den nächsten Kunden.

6. Eine Lieferbotin brachte einem nervigen Kunden eine Frühlingsrolle. Der Kunde forderte jedoch von der Lieferbotin eine neue Frühlingsrolle, da diese kalt war. Nach einer Stunde kehrte die Lieferbotin einfach mit derselben kalten Frühlingsrolle zum Kunden zurück.
   Nichtsahnend [nahm / begrüßte / reparierte] der Kunde die Frühlingsrolle und sogleich schloss er hinter sich die Tür.

7. In einem Restaurant unterhielt sich eine Vegetarierin mit einem befreundeten Metzger über eine Fleischwurst auf seinem Teller. Der Metzger sah die Vegetarierin an und erklärte, diese Fleischwurst zu essen, wäre ein reines Vergnügen. Er verglich es sogar damit, eine schöne Oper zu hören. Die Vegetarierin hielt dies jedoch für einen schlechten Vergleich und wies den Metzger darauf hin, dass ein Tier für diese Fleischwurst getötet worden war.
   Dann [durchschnitt / belächelte / mietete] der Metzger die Fleischwurst und sofort begann er zu essen.

8. Ein gemeiner Kutscher schlug seinen Gaul immer sehr heftig mit einer Peitsche. Eines Tages wurde der Kutscher dabei von einem Tierliebhaber beobachtet, der Mitleid mit dem Gaul hatte. Sofort lief der Tierliebhaber zum Kutscher und seinem Gaul und nahm ihm die Peitsche weg.
   Dann [bedrohte / streichelte / füllte] der Tierliebhaber den Kutscher und darüber hinaus forderte er ihn auf, den Gaul in Ruhe zu lassen.

9. Mitten im Meer sah ein Kapitän ein Pärchen auf einem kleinen Segelboot. Schon aus großer Entfernung konnte der Kapitän sehen, dass das Segelboot kaputt und das Pärchen in großer Not war. Schnell änderte der Kapitän seinen Kurs und steuerte zum Segelboot, um dem Pärchen zu helfen.
   Dann [bestieg / rettete / verschloss] der Kapitän das Segelboot und sofort half er dem Pärchen.

10. Da der Wasserhahn einer älteren Hausfrau nicht mehr aufhörte zu tropfen, rief die Hausfrau schließlich einen Handwerker. Zuerst betrachtete der Handwerker den Wasserhahn ausführlich und versuchte dann, ihn zu reparieren. Geduldig wartete die Hausfrau daneben. Nach einer Weile sagte der Handwerker, dass

der Wasserhahn schon zu kaputt sei und er einen neuen installieren müsse.
Daraufhin [lobte / ersetzte / knickte] die Hausfrau den Handwerker und noch lange ärgerte sie sich über die Mängel moderner Geräte.

11. In einer fremden Stadt buchte ein Urlauber eine Stadtführung. Der Guide freute sich über das Interesse des Urlaubers und schenkte ihm noch einen Flyer. Der Guide erklärte dem verwunderten Urlauber, dass der Flyer zusätzliche Informationen enthalte, auf die er selbst während der Führung nicht eingehen werde. Der Urlauber freute sich über den Flyer und dankte dem Guide.
Nach der Führung [faltete / lobte / kochte] der Urlauber den Flyer und dann machte er sich auf den Weg zu seinem Hotel.

12. Ein Paparazzi stellte seine große Kamera auf und wartete auf eine berühmte Schauspielerin. Es war eine sehr gute Kamera und er wollte unbedingt tolle Bilder schießen. Als die Schauspielerin den Paparazzi entdeckte, wurde sie sehr wütend, da sie nicht fotografiert werden wollte. Deshalb warf die Schauspielerin die Kamera um.
Daraufhin [bedrohte / schulterte / färbte] der Paparazzi die Schauspielerin und ferner sagte er, dass er sich so nicht behandeln lasse.

13. Ein Schneider und seine Assistentin suchten für eine neue Schaufensterpuppe, die der Schneider auf einer Messe ersteigert hatte, einen Platz in dem Laden. Zuerst stellte die Assistentin sie in den hinteren Teil des Ladens. Doch dann überzeugte sie den Schneider, die Schaufensterpuppe in die Nähe des Eingangs zu stellen, da das Licht dort besser war. Tatsächlich befand die Assistentin, dass die Schaufensterpuppe dort durch das viele Licht sehr gut zur Geltung komme.
Daraufhin [lobte / bewunderte / schnitt] der Schneider die Assistentin und dann sagte er, dass der Platz am Eingang eine gute Idee war.

14. Ein Schwimmer übte einen besonders schwierigen Sprung vom Sprungbrett, als er am Beckenrand ein Mädchen entdeckte. Seit einiger Zeit schon bewunderte er das Mädchen aus der Ferne, hatte sich aber nie getraut, es anzusprechen. Doch heute wollte der Schwimmer dies nachholen und ihm kam die Idee, dass er es mit dem anspruchsvollen Sprung vom Brett beeindrucken könnte. So wartete er einen Moment ab, in dem das Mädchen zum Brett blickte und sprang dann ins Wasser. Nach dem geglückten Sprung ging der Schwimmer sofort zu dem Mädchen und sprach es an.
Danach [musterte / bewertete / salzte] das Mädchen den Schwimmer und nach einer Weile verriet es ihm seine Handynummer.

15. Erfreut zeigte eine Sekretärin ihrem Chefarzt die neue Diktiermaschine. Damit konnte der Chefarzt seine Arztberichte nun selbst aufzeichnen und war nicht mehr auf die Hilfe seiner Sekretärin angewiesen. Bisher hatte sie nämlich seine Berichte selbst aufschreiben müssen. Deswegen freute sie sich besonders über die neue Diktiermaschine. Da der Chefarzt heute besonders viele Patienten gehabt hatte, schlug die Sekretärin ihm vor, die neue Diktiermaschine direkt auszuprobieren.
Dann [verabschiedete / enthüllte / leerte] der Chefarzt die Sekretärin und dann machte er Feierabend.

16. Eine Reporterin wollte einen Bericht über eine Farm schreiben. Dafür hatte sie sich ein paar Fragen überlegt, die sie dem Bauern stellen wollte. Am Hof angekommen begrüßte ein Mitarbeiter die Reporterin freundlich und brachte sie zum Bauern. Auf dem Weg erzählte der Mitarbeiter, dass er schon seit zwanzig Jahren auf der Farm arbeite. Beim Farmhaus angekommen, stellte der Mitarbeiter die Reporterin dem Bauern vor und wünschte ihnen ein erfolgreiches Interview.
Daraufhin [verabschiedete / befragte / ordnete] die Reporterin den Mitarbeiter und anschließend machte sie ein paar Fotos vom Bauernhof.

17. Ein Gärtner war sehr stolz auf seinen schönen neuen Rasenmäher, denn der Rasenmäher war so groß, dass man auf diesem sitzen und wie mit einem Auto herumfahren konnte. Das erzählte der Gärtner auch der kleinen Tochter seines Chefs. Begeistert fragte die Tochter des Chefs, ob sie auch mal fahren dürfe. Die Tochter kletterte neben den Gärtner auf den Sitz des Rasenmähers und sie drehten eine große Runde über die Wiese.
Danach [parkte / verabschiedete / halbierte] die Tochter den Rasenmäher und dann sagte sie begeistert, dass sie morgen wiederkommen würde.

18. Eine junge Dame wollte einen Edelstein von einem Juwelier beurteilen lassen. Stolz erzählte sie ihm, dass sie ihn von ihrer Großtante geerbt habe. Nun wollte die Dame von dem Juwelier wissen, um welche Art Edelstein es sich handelte. Der Juwelier betrachtete den Edelstein sehr lange und sagte dann zu der jungen

Dame, dass er sehr selten und wunderschön sei.
Entzückt [entlohnte / bestaunte / würzte] die Dame den Juwelier und danach bedankte sie sich für sein Fachwissen.

19. Ein Mechaniker machte einige Zaubertricks mit einem Schraubenzieher für seine kleine Nichte. Zu ihrer Überraschung war das Werkzeug plötzlich aus der Hand des Mechanikers verschwunden, doch kurz darauf zog er den Schraubenzieher hinter dem Ohr der Nichte hervor und lachte über ihren erstaunten Gesichtsausdruck. Geheimnisvoll erzählte der Mechaniker der Nichte, dass er gerade Magie benutzt habe, um den Schraubenzieher verschwinden zu lassen.
Verblüfft [nahm / bewunderte / kochte] die Nichte den Schraubenzieher und dann sagte sie, dass sie noch mehr Zaubertricks sehen wolle.

20. Ein Mopedfahrer war versehentlich gegen die Stoßstange eines Autos gefahren. Der Autofahrer verlangte nun, dass der Mopedfahrer für den Schaden aufkomme, doch dieser weigerte sich und sagte, dass die Stange ja überhaupt nicht beschädigt sei. Daraufhin rief der Autofahrer einen Polizisten zur Hilfe. Der Polizist eilte sofort herbei und begutachtete das Fahrzeug. Dann sagte der Polizist zum Mopedfahrer, dass dieser für die Reparaturkosten des Autofahrers aufkommen müsse.
Daraufhin [bestach / entschädigte / sortierte] der Mopedfahrer den Polizisten und außerdem entschuldigte er sich für den Unfall.

21. Ein Segler und seine Freundin hatten einen Bootsausflug gemacht. Nun wollten sie das Boot wieder am Steg festbinden. Die Freundin griff nach dem Strick und wollte dem Segler helfen, doch dieser sagte der Freundin, dass er keine Hilfe benötige. Daraufhin packte er den Strick und wollte einen Knoten binden. Plötzlich glitt dem Segler der Strick aus den Händen und fiel ins Wasser.
Daraufhin [ermahnte / schnappte / verschraubte] die Freundin den Segler und dann sagte sie, er solle etwas aufmerksamer sein.

22. Als Piraten von riesigen Goldschätzen auf einer kleinen Insel mitten im Meer gehört hatten, machten sie sich sofort auf den Weg, um sie zu suchen. Zu ihrer Überraschung entdeckten sie Einheimische auf der Insel, die die Goldschätze bewachten. Die Piraten versteckten sich vor den Einheimischen, um in Ruhe ihren Überfall vorbereiten zu können. Die Piraten warteten ab, bis die Einheimischen schliefen, um unbemerkt an die Goldschätze zu kommen.
Dann [versklavten / raubten / wechselten] die Piraten die Einheimischen und danach segelten sie Richtung Heimat.

23. Ein Junge verspürte Lust, einen Apfel zu essen. Erst gestern hatte er bei der Ernte geholfen und anschließend den vollen Korb nach Hause getragen. Bei dem Gedanken, wie schwer der Korb gewesen war und daran, wie frisch und saftig der Apfel sein musste, lief dem Jungen glatt das Wasser im Mund zusammen. Der Junge wusste, dass die Mutter den Korb mit seinem ersehnten Apfel im Keller versteckte.
Sofort [suchte / zerschnitt / schlug] der Junge den Korb und dann entschied er sich für einen großen roten Apfel.

24. Schon seit einiger Zeit bereitete sich ein Sportler auf einen großen Wettkampf im Ringen vor. Der Vater des Sportlers half ihm täglich beim Training, denn gemeinsam wollten sie den Juror mit einer guten Technik überzeugen. Der Vater kannte den Juror schon seit langer Zeit und wusste, dass der Juror sehr auf die richtige Technik achtete. Am Tag des Wettkampfes war der Vater sehr aufgeregt, doch der Sportler beeindruckte alle mit seiner hervorragenden Technik und gewann den Wettbewerb.
Danach [beglückwünschte / entdeckte / öffnete] der Juror den Sportler und außerdem lobte er dessen Sohn in höchsten Tönen.

25. Eine Geschäftsfrau hatte bei einer Auktion eine süße, alte Scheune ersteigert, die sie zu einer Bar herrichten ließ. Ihr Mann war nämlich Kellner und wollte sich schon lange selbstständig machen. Da der Mann in Bezug auf Ästhetik nicht sehr viel verstand, überließ er es ihr, die Renovierungsarbeiten anzuleiten. Diese hatten einige Zeit beansprucht, doch der Geschäftsfrau war das egal, denn sie war mit dem Resultat äußerst zufrieden. Die Bar war wunderschön geworden und hatte ihr altes Flair nicht verloren. Begeistert zeigte die Geschäftsfrau ihrem Mann die fertige Bar.
Daraufhin [umarmte / bestaunte / sortierte] der Mann die Geschäftsfrau und dann lobte er sie für ihren guten Geschmack.

26. Ein Rentner wollte auf einem Trödelmarkt sein altes Zelt verkaufen, mit dem er schon viele schöne Urlaube verbracht hatte. Deshalb wollte er nun einen neuen Besitzer finden, der genauso viel Freude daran haben

würde, wie er selbst sie gehabt hatte. Plötzlich tauchte ein kleines Kind neben ihm auf und starrte begeistert auf das Zelt. Das Kind stellte dem Rentner viele Fragen und erzählte ihm auch von seinen eigenen Campingausflügen mit der Familie. Schließlich fragte das Kind nach dem Preis für das Zelt.
Lachend [holte / tätschelte / aß] der Rentner das Zelt und dann schenkte er es dem Kind.

27. Als ein Lehrer seine Unterlagen holen wollte, bemerkte er, dass er seine Tasche nicht bei sich hatte. Erschrocken überlegte er, wo er die Tasche hatte stehen lassen. Ihm fiel ein, dass er sich eine Limonade hatte kaufen wollen, aber nicht genügend Kleingeld gehabt hatte. Deswegen war der Lehrer nochmal zurück ins Lehrerzimmer gegangen, um mehr Geld zu holen. Dort war er von einem Kollegen in ein wichtiges Gespräch verwickelt worden, sodass er die Limonade total vergessen hatte. In aller Aufregung über die Limonade hatte er bestimmt auch die Tasche in der Kantine stehen lassen.
Zurück in der Kantine [kaufte / fand / unterrichtete] der Lehrer die Limonade und dann suchte er seine Tasche.

28. Eine junge Bergsteigerin hatte eine neue Spitzhacke geschenkt bekommen und war nun erpicht darauf, diese sogleich an einer sehr steilen Bergwand auszuprobieren. Ihre Mutter hatte ihr die Spitzhacke erst am Tag zuvor gekauft, nachdem der Verkäufer der Mutter versichert hatte, dass es ein sehr gutes Modell sei. Am Morgen hatte die Mutter ihr viel Erfolg gewünscht und danach war die Bergsteigerin voller Tatendrang aufgebrochen, den Berg zu erklimmen. Doch leider brach die Spitzhacke durch, nachdem die Bergsteigerin schon eine Weile geklettert war und sie musste von der Bergwacht gerettet werden.
Im Krankenhaus [tröstete / verwünschte / stapelte] die Mutter die Bergsteigerin und hinterher betrachtete sie die kaputte Spitzhacke.

29. Ein Förster und eine Praktikantin gingen in den Wald, um Wild zu sehen. Der Förster schlug vor, auf einen Hochsitz zu klettern, da sie dort einen besseren Überblick haben würden. Nach einer Weile entdeckte die Praktikantin einen Hirsch. Der Hirsch war groß und hatte ein mächtiges Geweih. Doch er war sehr weit entfernt, weshalb die Praktikantin enttäuscht sagte, dass sie kaum etwas erkennen könne. Daraufhin holte der Förster ein Fernglas aus seiner Tasche und gab es ihr, damit sie den Hirsch sehen konnte.
Dann [umarmte / beobachtete / sammelte] die Praktikantin den Förster und anschließend bedankte sie sich für das Fernglas.

30. Ein Angeklagter wurde zum Gerichtssaal gebracht, wo der Richter und der Staatsanwalt schon auf ihn warteten. Der Mann wurde eines Raubüberfalls beschuldigt und heute war der erste Anhörungstag. Nachdem der Richter die Sitzung eröffnet hatte, trug der Staatsanwalt alle Punkte vor, die dem Angeklagten vorgeworfen wurden. Danach dankte der Richter dem Staatsanwalt und begann mit der Anhörung des Angeklagten.
Am Ende [konsultierte / befragte / kopierte] der Richter den Staatsanwalt und danach ließ er den ersten Zeugen herein.

31. Aufgeregt standen die Gäste in der Kirche und lauschten der rührenden Predigt des Pfarrers. Die Braut konnte den Moment kaum erwarten, in dem sie dem Bräutigam ihr Jawort geben und den Ring erhalten würde. Sie wusste, dass der Ring ein sehr besonderes Erbstück aus der Familie des Bräutigams war, das schon lange von Generation zu Generation weitergegeben worden war, und fühlte sich sehr geehrt, dieses zu erhalten. Als der Pfarrer die Predigt beendete und dem Brautpaar die Frage stellte, gaben sich der Bräutigam und die Braut das Jawort, während der Trauzeuge den schönen Ring hervorholte.
Glücklich [küsste / bewunderte / vereinfachte] die Braut den Bräutigam und dann übergab der Trauzeuge den Ring.

32. Während ein Ritter seinen Umhang anprobierte, besprach er das bevorstehende Turnier mit dem Burgfräulein. Das Burgfräulein fand den Umhang viel zu groß und schlug vor, ihn etwas zu kürzen. Aber der Ritter wollte nicht, dass das Burgfräulein irgendetwas veränderte. Er hatte den Umhang schon seit Jahren und dieser hatte dem Ritter bisher immer Glück gebracht.
Daraufhin [faltete / verspottete / entleerte] das Burgfräulein den Umhang und dann wünschte es dem Ritter viel Erfolg.

33. In einem Museum konnte eine Besucherin einen bestimmten Raum nicht finden. Verzweifelt versuchte sie, sich an der Wegbeschreibung auf ihrer Eintrittskarte zu orientieren, aber ohne Erfolg. Dann entdeckte die Besucherin eine Aufsichtsperson am anderen Ende des Raumes und fragte sie nach Hilfe. Die Aufsichtsperson erzählte, dass einige Leute Probleme mit der Wegbeschreibung auf der Eintrittskarte hätten. Die Aufsichtsperson nahm die Eintrittskarte der Besucherin und versprach, ihr den Weg zu zeigen.
Daraufhin [begleitete / studierte / erfand] die Aufsichtsperson die Besucherin und währenddessen erklärte sie ihr den Weg.

34. Ein Händler war auf dem Weg in den fernen Orient, um dort kostbare Gewürze einzukaufen. Dort angekommen begab er sich zum Marktplatz. Der Händler konnte schon von weitem die Rufe hören, mit denen die Sklaven zum Kauf angepriesen wurden. Der Händler fragte jemanden nach dem Stand mit den Gewürzen. Auf dem Weg zu den Gewürzen kam auch er an den Sklaven vorbei, welche seine fremdländischen Gewänder interessiert musterten.
Dann [grüßte / kaufte / versiegelte] der Händler die Sklaven und anschließend ging er weiter zu den Gewürzen

35. Ein Kind entdeckte in einem Schaufenster einen Teddybären, den es unbedingt haben wollte. Der Ladenbesitzer bemerkte die bewundernden Blicke des Kindes und nahm ihn vom Regal. Das Kind sagte dem Ladenbesitzer, dass es den Teddybären gerne kaufen würde, worauf der Ladenbesitzer ihm den Teddybären überreichte.
Dann [drückte / entlohnte / bastelte] das Kind den Teddybären und dann lachte es vor Freude.

36. Eine Hundeliebhaberin hatte ihren Nachbarn engagiert, um auf den Welpen aufzupassen, da sie über das Wochenende geschäftlich unterwegs war. Da die Hundeliebhaberin wusste, dass der Nachbar sich gut mit Tieren auskannte und den Welpen auch sehr gerne hatte, hatte sie keine Bedenken. Trotzdem war sie froh, als sie wieder zu Hause war. Als die Hundeliebhaberin die Haustüre aufschloss, rannte ihr der Welpe entgegen und der Nachbar begrüßte sie freundlich.
Daraufhin [entlohnte / drückte / sortierte] die Hundeliebhaberin den Nachbarn und außerdem bedankte sie sich für seine Zeit.

37. Ein Schuhverkäufer hatte gerade einem Kunden ein Paar Schuhe verkauft, als er beobachtete, wie draußen vor seinem Laden ein Dieb dem Kunden seine Geldbörse entwendete. Auch sah der Schuhverkäufer, dass dieser nichts davon mitbekommen hatte und der Dieb sich geschickt aus dem Staub machte. Der Schuhverkäufer blickte dem Kunden hinterher und rannte schnell nach draußen, um den Dieb aufzuhalten.
Dann [bemitleidete / verfolgte / hinterlegte] der Schuhverkäufer den Kunden und sofort erzählte er ihm von dem beobachteten Diebstahl.

38. Ein Eskimo wollte auf die Jagd gehen, um eine Robbe zu jagen. Er nahm seine Freundin als Begleitung mit. Auf dem Weg sagte der Eskimo zu der Freundin, dass sie sich ganz still verhalten müsse und sich nicht mehr bewegen dürfe, sobald sie die Robbe erblickten. Nach einer Weile entdeckte der Eskimo die Robbe in geeigneter Entfernung und zeigte sie der Freundin.
Dann [ermahnte / erschoss / verpackte] der Eskimo die Freundin und danach lud er sein Gewehr neu.

39. Nach einer Abendveranstaltung machte sich eine Tänzerin auf den Weg nach Hause. Sie beeilte sich, um schnell bei ihrer Tochter und der Babysitterin zu sein. Da die Babysitterin das erste Mal auf die Tochter aufgepasst hatte, wollte die Tänzerin schnell nach Hause, um nach dem Rechten zu schauen. Zuhause angekommen fand die Tänzerin eine glückliche Tochter und eine entspannte Babysitterin vor und war sehr erleichtert.
Dann [vergütete / umarmte / stapelte] die Tänzerin die Babysitterin und anschließend schickte sie diese nach Hause.

40. Ein Minister und sein Berater waren erzürnt über den Präsidenten aus dem Nachbarland, da dieser sich nicht an ein Handelsabkommen hielt. Daraufhin riet der Berater dem Minister, mit Sanktionen gegen den Präsidenten vorzugehen. Der Berater organisierte ein Treffen, bei dem der Minister dem Präsidenten seine Forderungen überbringen konnte.
Dann [verhandelte / feierte / schminkte sich] der Minister mit dem Präsidenten und dabei besprachen sie genauere

41. Seit Monaten hatte sich der Athlet mit der Trainerin darauf vorbereitet, bei dem wichtigsten Wettkampf des Jahres den Pokal zu holen. Die Trainerin trieb ihn hart an, da sie sicher war, dass er gute Chancen hatte. Und tatsächlich hatte sich die harte Arbeit gelohnt, denn der Athlet gewann den Pokal und überglücklich bedankte er sich bei der Trainerin. Stolz hielt der Athlet den Pokal in den Händen.
Im Hotel [polierte / bejubelte / verspeiste] die Trainerin den Pokal und anschließend stellte sie ihn auf den Tisch.

42. Ein Autor ging mit dem Hund spazieren, um an der frischen Luft neue Ideen für sein derzeitiges Buch zu bekommen. Der Autor hatte einen Ball dabei, da der Hund sehr verspielt war. Im Park angekommen, warf der Autor den Ball einige Meter weit. Sofort rannte der Hund dem Ball nach und brachte ihn brav zurück.
Daraufhin [nahm / tätschelte / zitierte] der Autor den Ball und wieder warf er ihn einige Meter weit.

43. Eine Oma und ein Kleinkind standen vor einem Hasenstall und streichelten das Kaninchen. Die Oma gab dem Kleinkind Löwenzahn, damit dieses das Kaninchen füttern konnte und dann ging sie noch mehr Löwenzahn holen. Doch plötzlich biss das Kaninchen das Kleinkind und dieses fing fürchterlich an zu weinen.
Daraufhin [fütterte / streichelte / strickte] die Oma das Kaninchen und nebenbei tröstete sie das Kleinkind.

44. Der Geschäftsführer und der Coach saßen nebeneinander und schauten einem bedeutenden Fußballspiel zu. Leider war die Mannschaft, die der Coach trainierte, deutlich unterlegen. Der Torwart hatte bisher fast keinen Ball gehalten. Der Geschäftsführer saß bekümmert auf der Bank und selbst die gute Leistung der anderen Spieler konnte die Ungeschicktheit des Torwarts nicht wieder gut machen. Auch der Coach wirkte verzweifelt, als der Torwart aus Versehen den Ball einem gegnerischen Spieler zuspielte, worauf dieser ein Tor schoss. Am Ende verlor die Mannschaft das Spiel. Entrüstet sagte der Geschäftsführer dem traurigen Coach, dass er mit dem Torwart reden wolle.
Letztendlich [suspendierte / umarmte / reparierte] der Geschäftsführer den Torwart und dann fuhr er immer noch wütend nach Hause.

45. Als ein Referendar den Weihnachtsmarkt seines Gymnasiums betrat, wurde er direkt von ein paar Schülern begrüßt. Die Schüler berichteten dem Referendar, dass sie eine Tombola organisiert hatten und nun versuchten, die Lose zu verkaufen. Die Schüler hatten schon sehr viele Lose verkauft und erzählten dem Referendar nun, was er alles Schönes mit den Losen gewinnen könne.
Amüsiert [kaufte / musterte / betrat] der Referendar die Lose und tatsächlich gewann er einen Preis.

46. Eine Mutter ging mit ihren eineiigen Zwillingen zum Doktor, da diese geimpft werden sollten. Im Behandlungszimmer des Doktors machte dieser Witze darüber, wie ähnlich sich die Zwillinge sahen und zeigte ihnen die Spritzen, die er schon vorbereitet hatte. Der Doktor versicherte ihnen, dass sie keine Angst vor den Spritzen haben müssten. Da die Spritzen mit ihren langen, dünnen Nadeln tatsächlich angsteinflößend aussahen, bekamen die Zwillinge trotzdem Angst.
Dann [nahm / verwechselte / bastelte] der Doktor die Spritzen und anschließend begann er mit der Impfung.

47. In einem Kriegsgebiet wollte ein Soldat eine Zivilistin unbemerkt an den gegnerischen Truppen vorbei schmuggeln, da es für sie sehr gefährlich war, allein unterwegs zu sein. Da sie sich in einem Kriegsgebiet befanden, hielt der Soldat die Waffe bereit. So schlichen die Zivilistin und der Soldat mit seiner Waffe still die Häuser entlang. Die Zivilistin war sehr erleichtert über die Hilfe und fühlte sich durch die Waffe auch sicher, doch plötzlich tauchte vor ihnen ein Panzer des gegnerischen Lagers auf.
Schnell [zückte / versteckte / durchkämmte] der Soldat die Waffe und sofort ging er in Deckung.

48. Ein Dirigent hatte ein neues Stück geschrieben und wollte es heute Abend zum ersten Mal dem Publikum zeigen. Lange hatte er mit dem Orchester geprobt und war gespannt auf die Reaktion des Publikums. Machte das Orchester heute Abend keinen Fehler, könnte das Stück die Karriere des Dirigenten voranbringen. Als der Abend gekommen war, betrat der Dirigent zusammen mit dem Orchester die Bühne, um dem Publikum das Stück zu präsentieren.
An diesem Abend [spielte / verzauberte / engagierte] das Orchester das Stück und das Publikum applaudierte.

49. Ein Doktorand hatte nach Jahren endlich seine Arbeit beendet und musste sie nun seiner Betreuerin und anderen Prüfern vorstellen. Obwohl der Doktorand eng mit der Betreuerin zusammengearbeitet hatte und wusste, dass die Arbeit sehr gut war, war er trotzdem sehr nervös. Vor der Prüfung ging der Doktorand noch einmal die wichtigsten Stichpunkte bezüglich der Arbeit durch, dann folgte er den anderen Prüfern ins Büro der Betreuerin.
Dort [begrüßte / verteidigte / reparierte] der Doktorand die Betreuerin und dann hielt er seinen Vortrag.

50. Eine Protestantin wollte nach Israel fliegen, um sich Jerusalem anzuschauen. Sie hatte die Reise geplant, seitdem der Pfarrer ihr Bilder von seinem Aufenthalt dort gezeigt hatte. Nun war die Reise fertig organisiert und der Abflug rückte immer näher. Doch die Protestantin machte sich Sorgen, da es in letzter Zeit vermehrt Unruhen gegeben hatte. So ging sie zu dem Pfarrer, um ihn um Rat zu fragen. Sie wollte, dass der Pfarrer ihr versicherte, dass sie sich keine Sorgen machen müsse. Dadurch würde sich die Protestantin bezüglich der Reise sicherer fühlen.
Daraufhin [segnete / bewilligte / las] der Pfarrer die Protestantin und dann wünschte er der Protestantin einen guten Flug.

51. Eine Erzieherin suchte einen Therapeuten auf. Dieser war ihr von einer Freundin empfohlen worden, nachdem sie über Symptome geklagt hatte. Die Symptome waren denen einer Depression ziemlich ähnlich und die Erzieherin hatte beschlossen, dass sie professionelle Hilfe von dem Therapeuten brauche. So war die Erzieherin sehr erleichtert gewesen, als sie endlich einen Termin bei dem Therapeuten bekommen hatte, da sich die Symptome in letzter Zeit noch verschlimmert hatten.
In der Praxis [erfragte / behandelte / tauschte] der Therapeut die Symptome und daraufhin verschrieb er ein Medikament.

52. Eine Designerin hatte den Auftrag bekommen, ein Buch grafisch zu gestalten. In dem Buch ging es um Geschichten über Eisbären. Die Geschichten waren für Kinder gedacht und der Verlag wollte, dass die Designerin die Eisbären bildlich darstellte. Nun hatte die Designerin die Geschichten über die Eisbären zu Ende gelesen und war bereit, mit der Arbeit zu beginnen.
Dann [malte / veranschaulichte / leerte] die Designerin die Eisbären und bis spät in die Nacht arbeitete sie an der Geschichte.

53. Eines Abends wurde ein Architekt von dem Bürgermeister angerufen. Dieser sagte, dass die Stadt eine neue Turnhalle zu bauen beabsichtigte. Er beauftragte den Architekten, einen Plan der Turnhalle zu erstellen und diesen in einer Rede vor dem Gemeinderat näher auszuführen. In der Rede solle er auf die besonderen Merkmale seines Entwurfes eingehen. Der Architekt versicherte, dass er sofort mit der Konzeption der Turnhalle beginnen werde und bedankte sich für die Tipps bezüglich der Rede.
Daraufhin [schrieb / entwarf / rief] der Architekt die Rede und dann goss er sich ein Glas Wein ein.

54. Ein Gitarrist wurde von einer Agentin engagiert, um zusammen mit einer Sängerin auf einer Party aufzutreten. Auf dem Weg zur Probe erzählte die Agentin dem Gitarristen, dass sie lange nach einem guten Musiker gesucht habe und glaube, dass seine Art zu spielen ausgezeichnet mit der Stimme der Sängerin harmonieren würde. Dann holte der Gitarrist sein Instrument und die Agentin sagte, dass die Sängerin schon bereit sei und sie direkt mit der Probe beginnen könnten.
Dann [verabschiedete / traf / kaufte] der Gitarrist die Agentin und dann ging er schnell zur Bühne.

55. In einem Museum war ein Kurator dabei, eine neue Ausstellung zu gestalten. Da es um plastische Kunst ging, hatte sich der Kurator von einer befreundeten Galeristin eine Skulptur geliehen. Gerade war die Galeristin eingetroffen und sie überlegten nun gemeinsam, wo die Skulptur am Besten zur Geltung kommen würde. Lange suchten sie nach einem geeigneten Platz und fanden schließlich einen. Mühevoll installierte der Kurator die Skulptur, während die Galeristin Anleitungen gab.
Danach [umarmte / betrachtete / sammelte] der Kurator die Galeristin und dabei dankte er ihr für ihre Hilfe.

56. Eine Studentin war mit einer Kommilitonin in einer Kneipe. Da sie danach noch in einem Club feiern gehen wollten, beschlossen sie, sich auf der Toilette frisch zu machen. Dann fragte die Kommilitonin die Studentin, ob sie ihre Wimperntusche ausleihen dürfe, da sie ihre eigene vergessen hatte. Sofort gab die Studentin ihr die Wimperntusche. Die Kommilitonin fragte eine Bedienung nach der Toilette und machte sich mit der Wimperntusche in der Hand auf den Weg zur Toilette.
Dann [betrat / benutzte / las] die Kommilitonin die Toilette und anschließend schminkte sie sich.

57. Ein Verbrecher war auf dem Weg zu einem Haus, wo ein Ermittler wohnte. Dieser untersuchte einen Fall, in den der Verbrecher verstrickt war. Deswegen wollte dieser den Ermittler aus dem Weg räumen. Am Haus angekommen verschaffte sich der Verbrecher Zutritt. Er wusste, dass es in dem Haus einen Schäferhund gab und bedacht achtete er darauf, dass der Schäferhund ihn nicht hörte. Bevor er den Ermittler suchte, gab er dem Schäferhund etwas zu Essen, um ihn abzulenken.
Dann [streichelte / erschoss / faltete] der Verbrecher den Schäferhund und danach machte er sich auf die Suche nach dem Ermittler.

58. Ein Beschuldigter und seine Anwältin betraten den Gerichtssaal, um bei der bevorstehenden Anhörung zu beweisen, dass der Beschuldigte die Tat nicht begangen hatte. Der Kläger saß schon an seinem Platz und warf den beiden böse Blicke zu. Die Anwältin ging noch ein paar ihrer Unterlagen durch, dann begann die Verhandlung. Der Kläger wurde nach vorne gebeten und von der Anwältin zur Tat befragt. Der Beschuldigte blickte nervös drein, als der Kläger ihn vor aller Augen der Tat bezichtigte.
Daraufhin [verteidigte / entließ / schwenkte] die Anwältin den Beschuldigten und dann wandte sie sich an den Richter.

59. Ein Junge ging mit seinem Kumpel zum See, da er schwimmen wollte. Der Kumpel hatte seine Angel dabei und erzählte dem Jungen, dass er heute einen Flussbarsch angeln wollte, von denen es viele im See gab.

Er hatte einen besonderen Köder dabei, mit dem er den Flussbarsch anlocken wollte. Der Junge wünschte dem Kumpel viel Glück mit dem Flussbarsch und machte einen Salto ins Wasser.
Daraufhin [angelte / beobachtete / trocknete] der Kumpel den Flussbarsch und danach ging er selbst ins Wasser.

60. Eine Schwangere betrat das Untersuchungszimmer einer Gynäkologin und wurde von der Gynäkologin freundlich begrüßt. Die Gynäkologin deutete auf die Liege im Zimmer und forderte die Schwangere auf, sich dort hinzulegen. Die Liege war etwas hoch eingestellt, doch die Schwangere schaffte es, hochzukommen und legte sich auf die Liege.
Daraufhin [verstellte / untersuchte / verordnete] die Gynäkologin die Liege und dann wandte sie sich der Schwangeren zu.

Bibliography

Alday, P. M. (2019). How much baseline correction do we need in ERP research? Extended GLM model can replace baseline correction while lifting its limits. *Psychophysiology*, *56*(12), e13451. https://doi.org/10.1111/psyp.13451

Aurnhammer, C., Crocker, M. W., & Brouwer, H. (2023). Single-trial neurodynamics reveal N400 and P600 coupling in language comprehension. *Cognitive Neurodynamics*. https://doi.org/10.1007/s11571-023-09983-7

Aurnhammer, C., Delogu, F., Brouwer, H., & Crocker, M. W. (2023). The P600 as a continuous index of integration effort. *Psychophysiology*, *60*(9), e14302. https://doi.org/10.1111/psyp.14302

Aurnhammer, C., Delogu, F., Schulz, M., Brouwer, H., & Crocker, M. W. (2021). Retrieval (N400) and integration (P600) in expectation-based comprehension. *PLOS ONE*, *16*(9), e0257430. https://doi.org/10.1371/journal.pone.0257430

Aurnhammer, C., & Frank, S. L. (2019a). Comparing gated and simple recurrent neural network architectures as models of human sentence processing. *Proceedings of the 41st Annual Conference of the Cognitive Science Society*, 112–118.

Aurnhammer, C., & Frank, S. L. (2019b). Evaluating information-theoretic measures of word prediction in naturalistic sentence reading. *Neuropsychologia*, *134*, 107198. https://doi.org/10.1016/j.neuropsychologia.2019.107198

Baggio, G. (2018). *Meaning in the brain*. MIT Press.

Baggio, G., & Hagoort, P. (2011). The balance between memory and unification in semantics: A dynamic account of the N400. *Language and Cognitive Processes*, *26*(9), 1338–1367. https://doi.org/10.1080/01690965.2010.542671

Barber, H., Vergara, M., & Carreiras, M. (2004). Syllable-frequency effects in visual word recognition: Evidence from ERPs. *NeuroReport*, *15*(3), 545–548.

Bartholow, B. D., Pearson, M. A., Dickter, C. L., Sher, K. J., Fabiani, M., & Gratton, G. (2005). Strategic control and medial frontal negativity: Beyond errors and response conflict. *Psychophysiology*, *42*(1), 33–42. https://doi.org/10.1111/j.1469-8986.2005.00258.x

beim Graben, P., Gerth, S., & Vasishth, S. (2008). Towards dynamical system models of language-related brain potentials. *Cognitive Neurodynamics*, *2*(3), 229–255. https://doi.org/10.1007/s11571-008-9041-5

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Berger, H. (1929). Über das Elektroenkephalogramm des Menschen. *Archiv für Psychiatrie und Nervenkrankheiten*, *87*(1), 527–570.

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review*, *59*(1), 65–98. https://doi.org/10.1137/141000671

Bornkessel-Schlesewsky, I., Kretzschmar, F., Tune, S., Wang, L., Genç, S., Philipp, M., Roehm, D., & Schlesewsky, M. (2011). Think globally: Cross-linguistic variation in electrophysiological activity during sentence comprehension. *Brain and Language*, *117*(3), 133–152. https://doi.org/10.1016/j.bandl.2010.09.010

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2008). An alternative perspective on "semantic P600" effects in language comprehension. *Brain Research Reviews*, *59*(1), 55–73. https://doi.org/10.1016/j.brainresrev.2008.05.003

Boston, M. F., Hale, J., Kliegl, R., Patil, U., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*, *2*(1), 1–12. https://doi.org/10.16910/jemr.2.1.1

Boudewyn, M. A., Gordon, P. C., Long, D., Polse, L., & Swaab, T. Y. (2012). Does discourse congruence influence spoken language comprehension before lexical association? Evidence from event-related potentials. *Language and Cognitive Processes*, *27*(5), 698–733. https://doi.org/10.1080/01690965.2011.577980

Brothers, T., Swaab, T. Y., & Traxler, M. J. (2015). Effects of prediction and contextual support on lexical processing: Prediction takes precedence. *Cognition*, *136*, 135–149. https://doi.org/10.1016/j.cognition.2014.10.017

Brouwer, H., & Crocker, M. W. (2017). On the proper treatment of the N400 and P600 in language comprehension. *Frontiers in Psychology*, *8*, 1327. https://doi.org/10.3389/fpsyg.2017.01327

Brouwer, H., Crocker, M. W., Venhuizen, N. J., & Hoeks, J. C. J. (2017). A neurocomputational model of the N400 and the P600 in language processing. *Cognitive Science*, *41*(Suppl. 6), 1318–1352. https://doi.org/10.1111/cogs.12461

Brouwer, H., Delogu, F., & Crocker, M. W. (2021). Splitting event-related potentials: Modeling latent components using regression-based waveform estimation. *European Journal of Neuroscience*, *53*, 974–995. https://doi.org/10.1111/ejn.14961

Brouwer, H., Delogu, F., Venhuizen, N. J., & Crocker, M. W. (2021). Neurobehavioral correlates of surprisal in language comprehension: A neurocomputational model. *Frontiers in Psychology*, *12*, 615538. https://doi.org/10.3389/fpsyg.2021.615538

Brouwer, H., Fitz, H., & Hoeks, J. C. J. (2010). Modeling the noun phrase versus sentence coordination ambiguity in Dutch: Evidence from surprisal theory. *Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics*, 72–80.

Brouwer, H., Fitz, H., & Hoeks, J. C. J. (2012). Getting real about Semantic Illusions: Rethinking the functional role of the P600 in language comprehension. *Brain Research*, *1446*, 127–143. https://doi.org/10.1016/j.brainres.2012.01.055

Brouwer, H., & Hoeks, J. C. J. (2013). A time and place for language comprehension: Mapping the N400 and the P600 to a minimal cortical network. *Frontiers in Human Neuroscience*, *7*, 758. https://doi.org/10.3389/fnhum.2013.00758

Brown, C., & Hagoort, P. (1993). The processing nature of the N400: Evidence from masked priming. *Journal of Cognitive Neuroscience*, *5*(1), 34–44. https://doi.org/10.1162/jocn.1993.5.1.34

Brown, C., & Hagoort, P. (2000). On the electrophysiology of language comprehension: Implications for the human language system. In M. W. Crocker, M. Pickering, & C. J. Clifton (Eds.), *Architectures and mechanisms for language processing* (pp. 213–237). Cambridge University Press.

Bulkes, N. Z., Christianson, K., & Tanner, D. (2020). Semantic constraint, reading control, and the granularity of form-based expectations during semantic processing: Evidence from ERPs. *Neuropsychologia*, *137*, 107294. https://doi.org/10.1016/j.neuropsychologia.2019.107294

Burkhardt, P. (2006). Inferential bridging relations reveal distinct neural mechanisms: Evidence from event-related brain potentials. *Brain and Language*, *98*(2), 159–168. https://doi.org/10.1016/j.bandl.2006.04.005

Burkhardt, P. (2007). The P600 reflects cost of new information in discourse memory. *NeuroReport*, *18*(17), 1851–1854. https://doi.org/10.1097/WNR.0b013e3282f1a999

Camblin, C. C., Gordon, P. C., & Swaab, T. Y. (2007). The interplay of discourse congruence and lexical association during sentence processing: Evidence from ERPs and eye tracking. *Journal of Memory and Language*, *56*(1), 103–128. https://doi.org/10.1016/j.jml.2006.07.005

Cheyette, S. J., & Plaut, D. C. (2017). Modeling the N400 ERP component as transient semantic over-activation within a neural network model of word comprehension. *Cognition*, *162*, 153–166. https://doi.org/10.1016/j.cognition.2016.10.016

Chow, W.-Y., & Phillips, C. (2013). No semantic illusions in the "Semantic P600" phenomenon: ERP evidence from Mandarin Chinese. *Brain Research*, *1506*, 76–93. https://doi.org/10.1016/j.brainres.2013.02.016

Christiansen, M. H., Conway, C. M., & Onnis, L. (2012). Similar neural correlates for language and sequential learning: Evidence from event-related brain potentials. *Language and Cognitive Processes*, *27*(2), 231–256. https://doi.org/10.1080/01690965.2011.606666

Cohn, N., & Kutas, M. (2015). Getting a cue before getting a clue: Event-related potentials to inference in visual narrative comprehension. *Neuropsychologia*, *77*, 267–278. https://doi.org/10.1016/j.neuropsychologia.2015.08.026

Connolly, J. F., Stewart, S. H., & Phillips, N. A. (1990). The effects of processing requirements on neurophysiological responses to spoken sentences. *Brain and Language*, *39*(2), 302–318. https://doi.org/10.1016/0093-934X(90)90016-A

Coulson, S., Federmeier, K. D., Van Petten, C., & Kutas, M. (2005). Right hemisphere sensitivity to word- and sentence-level context: Evidence from event-related brain potentials. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(1), 129–147. https://doi.org/10.1037/0278-7393.31.1.129

Dambacher, M., Kliegl, R., Hofmann, M., & Jacobs, A. M. (2006). Frequency and predictability effects on event-related potentials during reading. *Brain Research*, *1084*(1), 89–103. https://doi.org/10.1016/j.brainres.2006.02.010

Debruille, J. B. (2007). The N400 potential could index a semantic inhibition. *Brain Research Reviews*, *56*(2), 472–477. https://doi.org/10.1016/j.brainresrev.2007.10.001

Delogu, F., Brouwer, H., & Crocker, M. W. (2019). Event-related potentials index lexical retrieval (N400) and integration (P600) during language comprehension. *Brain and Cognition*, *135*, 103569. https://doi.org/10.1016/j.bandc.2019.05.007

Delogu, F., Brouwer, H., & Crocker, M. W. (2021). When components collide: Spatiotemporal overlap of the N400 and P600 in language comprehension. *Brain Research*, *1766*, 147514. https://doi.org/10.1016/j.brainres.2021.147514

Delogu, F., Crocker, M. W., & Drenhaus, H. (2017). Teasing apart coercion and surprisal: Evidence from eye-movements and ERPs. *Cognition*, *161*, 46–59. https://doi.org/10.1016/j.cognition.2016.12.017

Delogu, F., Drenhaus, H., & Crocker, M. W. (2018). On the predictability of event boundaries in discourse: An ERP investigation. *Memory & Cognition*, *46*(2), 315–325. https://doi.org/10.3758/s13421-017-0766-4

DeLong, K. A., Quante, L., & Kutas, M. (2014). Predictability, plausibility, and two late ERP positivities during written sentence comprehension. *Neuropsychologia*, *61*, 150–162. https://doi.org/10.1016/j.neuropsychologia.2014.06.016

DeLong, K. A., Urbach, T. P., Groppe, D. M., & Kutas, M. (2011). Overlapping dual ERP responses to low cloze probability sentence continuations. *Psychophysiology*, *48*(9), 1203–1207. https://doi.org/10.1111/j.1469-8986.2011.01199.x

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature neuroscience*, *8*, 1117–1121. https://doi.org/10.1038/nn1504

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193–210. https://doi.org/10.1016/j.cognition.2008.07.008

Dimitrova, D. V., Stowe, L. A., Redeker, G., & Hoeks, J. C. J. (2012). Less is not more: Neural responses to missing and superfluous accents in context. *Journal of Cognitive Neuroscience*, *24*(12), 2400–2418. https://doi.org/10.1162/jocn_a_00302

Ditman, T., Holcomb, P. J., & Kuperberg, G. R. (2007). An investigation of concurrent ERP and self-paced reading methodologies. *Psychophysiology*, *44*(6), 927–935. https://doi.org/10.1111/j.1469-8986.2007.00593.x

Drummond, A. (2012). Ibex: A web interface for psycholinguistic experiments [Accessed July 7, 2023]. https://adrummond.net/ibexfarm

Eddine, S. N., Brothers, T., & Kuperberg, G. R. (2022). Chapter Four - The N400 in silico: A review of computational models. In K. D. Federmeier (Ed.), *Psychology of learning and motivation* (pp. 123–206). Academic Press. https://doi.org/10.1016/bs.plm.2022.03.005

Erickson, T. D., & Mattson, M. E. (1981). From words to meaning: A semantic illusion. *Journal of Verbal Learning and Verbal Behavior*, *20*(5), 540–551. https://doi.org/10.1016/S0022-5371(81)90165-1

Ettinger, A. (2020). What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, *8*, 34–48. https://doi.org/10.1162/tacl_a_00298

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language*, *41*(4), 469–495. https://doi.org/10.1006/jmla.1999.2660

Federmeier, K. D., & Laszlo, S. (2009). Time for meaning: Electrophysiology provides insights into the dynamics of representation and processing in semantic memory. In B. H. Ross (Ed.), *Psychology of Learning and Motivation* (pp. 1–44). Academic Press. https://doi.org/10.1016/S0079-7421(09)51001-8

Federmeier, K. D., Wlotko, E. W., De Ochoa-Dewald, E., & Kutas, M. (2007). Multiple effects of sentential constraint on word processing. *Brain Research*, *1146*, 75–84. https://doi.org/10.1016/j.brainres.2006.06.101

Fernandez Monsalve, I., Frank, S. L., & Vigliocco, G. (2012). Lexical surprisal as a general predictor of reading time. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 398–408.

Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology*, *47*(2), 164–203. https://doi.org/10.1016/S0010-0285(03)00005-7

Ferreira, F., Bailey, K. G. D., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science*, *11*(1), 11–15. https://doi.org/10.1111/1467-8721.00158

Ferreira, F., & Patson, N. D. (2007). The 'Good Enough' approach to language comprehension. *Language and Linguistics Compass*, *1*(1), 71–83. https://doi.org/10.1111/j.1749-818X.2007.00007.x

Fischler, I., Bloom, P. A., Childers, D. G., Roucos, S. E., & Perry, N. W. (1983). Brain potentials related to stages of sentence verification. *Psychophysiology*, *20*(4), 400–409. https://doi.org/10.1111/j.1469-8986.1983.tb00920.x

Fitz, H., & Chang, F. (2019). Language ERPs reflect learning through prediction error propagation. *Cognitive Psychology*, *111*, 15–52. https://doi.org/10.1016/j.cogpsych.2019.03.002

Frank, S. L. (2009). Surprisal-based comparison between a symbolic and a connectionist model of sentence processing. *Proceedings of the 31st Annual Meeting of the Cognitive Science Society*, 1139–1144.

Frank, S. L. (2017). Word embedding distance does not predict word reading time. *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, 385–390.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1–11. https://doi.org/10.1016/j.bandl.2014.10.006

Frank, S. L., & Thompson, R. (2012). Early effects of word surprisal on pupil size during reading. *Proceedings of the 34th Annual Meeting of the Cognitive Science Society*, 1554–1559.

Frank, S. L., & Willems, R. M. (2017). Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. *Language, Cognition and Neuroscience*, *32*(9), 1192–1203. https://doi.org/10.1080/23273798.2017.1323109

Franklin, M. S., Dien, J., Neely, J. H., Huber, E., & Waterson, L. D. (2007). Semantic priming modulates the N400, N300, and N400RP. *Clinical Neurophysiology*, *118*(5), 1053–1068. https://doi.org/10.1016/j.clinph.2007.01.012

Friederici, A. D. (1995). The time course of syntactic activation during language processing: A model based on neuropsychological and neurophysiological data. *Brain and Language*, *50*(3), 259–281. https://doi.org/10.1006/brln.1995.1048

Friederici, A. D., & Mecklinger, A. (1996). Syntactic parsing as revealed by brain responses: First-pass and second-pass parsing processes. *Journal of Psycholinguistic Research*, *25*(1), 157–176. https://doi.org/10.1007/BF01708424

Frisson, S., Harvey, D. R., & Staub, A. (2017). No prediction error cost in reading: Evidence from eye movements. *Journal of Memory and Language*, *95*, 200–214. https://doi.org/10.1016/j.jml.2017.04.007

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 815–836. https://doi.org/10.1098/rstb.2005.1622

Fritz, I., & Baggio, G. (2020). Meaning composition in minimal phrasal contexts: Distinct ERP effects of intensionality and denotation. *Language, Cognition and Neuroscience*, *35*(10), 1295–1313. https://doi.org/10.1080/23273798.2020.1749678

Fritz, I., & Baggio, G. (2022). Neural and behavioural effects of typicality, denotation and composition in an adjective–noun combination task. *Language, Cognition and Neuroscience*, *37*(5), 537–559. https://doi.org/10.1080/23273798.2021.2004176

Frost, M., Lynch, D., Peyton, H., & Deschanel, C. (1991, January 12). The black widow (Season 2, Episode 19) [TV Series Episode. In M. Frost & D. Lynch (Executive Producers). *Twin Peaks*. Propaganda Films; Spelling Entertainment; Lynch/Frost Productions.].

Geyer, A., Holcomb, P., Kuperberg, G., & Perlmutter, N. (2006). Plausibility and sentence comprehension. An ERP study [Abstract]. *Journal of Cognitive Neuroscience, Supplement*.

Gouvea, A. C., Phillips, C., Kazanina, N., & Poeppel, D. (2010). The linguistic processes underlying the P600. *Language and Cognitive Processes*, *25*(2), 149–188. https://doi.org/10.1080/01690960902965951

Hagoort, P., Baggio, G., & Willems, R. M. (2009). *Semantic unification*. MIT Press.

Hagoort, P., Brown, C., & Groothusen, J. (1993). The syntactic positive shift (SPS) as an ERP measure of syntactic processing. *Language and Cognitive Processes*, *8*(4), 439–483. https://doi.org/10.1080/01690969308407585

Hagoort, P., & Brown, C. M. (2000). ERP effects of listening to speech: Semantic ERP effects. *Neuropsychologia*, *38*(11), 1518–1530. https://doi.org/10.1016/S0028-3932(00)00052-X

Hagoort, P., Brown, C. M., & Osterhout, L. (1999). The neurocognition of syntactic processing. In C. M. Brown & P. Hagoort (Eds.), *The neurocognition of language* (pp. 273–316). Oxford University Press.

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science*, *304*(5669), 438–441. https://doi.org/10.1126/science.1095455

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*, 159–166.

Hale, J. (2003). The information conveyed by words in sentences. *Journal of Psycholinguistic Research*, *32*(2), 101–123. https://doi.org/10.1023/A:1022492123056

Hennighausen, E., Heil, M., & Rösler, F. (1993). A correction method for DC drift artifacts. *Electroencephalography and Clinical Neurophysiology*, *86*(3), 199–204. https://doi.org/10.1016/0013-4694(93)90008-J

Hoeks, J. C. J., & Brouwer, H. (2014). Electrophysiological research on conversation and discourse processing. In T. M. Holtgraves (Ed.), *The Oxford handbook of language and social psychology*. Oxford University Press.

Hoeks, J. C. J., Stowe, L. A., & Doedens, G. (2004). Seeing words in context: The interaction of lexical and sentence level information during reading. *Cognitive Brain Research*, *19*(1), 59–73. https://doi.org/10.1016/j.cogbrainres.2003.10.022

Hoeks, J. C. J., Stowe, L. A., Hendriks, P., & Brouwer, H. (2013). Questions left unanswered: How the brain responds to missing information. *PLOS ONE*, *8*(10), e73594. https://doi.org/10.1371/journal.pone.0073594

Holcomb, P. J., Grainger, J., & O'Rourke, T. (2002). An electrophysiological study of the effects of orthographic neighborhood size on printed word perception. *Journal of Cognitive Neuroscience*, *14*(6), 938–950. https://doi.org/10.1162/089892902760191153

Jachmann, T. K., Drenhaus, H., Staudte, M., & Crocker, M. W. (2019). Influence of speakers' gaze on situated language comprehension: Evidence from event-related potentials. *Brain and Cognition*, *135*, 103571. https://doi.org/10.1016/j.bandc.2019.05.009

Kaan, E. (2007). Event-related potentials and language processing: A brief overview. *Language and Linguistics Compass*, *1*(6), 571–591. https://doi.org/10.1111/j.1749-818X.2007.00037.x

Kaan, E., Harris, A., Gibson, E., & Holcomb, P. (2000). The P600 as an index of syntactic integration difficulty. *Language and Cognitive Processes*, *15*(2), 159–201. https://doi.org/10.1080/016909600386084

Kaan, E., & Swaab, T. Y. (2003). Repair, revision, and complexity in syntactic analysis: An electrophysiological differentiation. *Journal of Cognitive Neuroscience*, *15*(1), 98–110. https://doi.org/10.1162/089892903321107855

Keller, F., Gunasekharan, S., Mayo, N., & Corley, M. (2009). Timing accuracy of web experiments: A case study using the WebExp software package. *Behavior Research Methods*, *41*(1), 1–12. https://doi.org/10.3758/BRM.41.1.12

Khachatryan, E., Vliet, M. v., Deyne, S. D., Storms, G., Manvelyan, H., & Hulle, M. M. V. (2014). Amplitude of N400 component unaffected by lexical priming for moderately constraining sentences. *4th International Workshop on Cognitive Information Processing (CIP)*, 1–6. https://doi.org/10.1109/CIP.2014.6844516

Khachatryan, E., Hnazaee, M. F., & Van Hulle, M. M. (2018). Effect of word association on linguistic event-related potentials in moderately to mildly constraining sentences. *Scientific Reports*, *8*(1), 7175. https://doi.org/10.1038/s41598-018-25723-y

Kim, A., & Osterhout, L. (2005). The independence of combinatory semantic processing: Evidence from event-related potentials. *Journal of Memory and Language*, *52*(2), 205–225. https://doi.org/10.1016/j.jml.2004.10.002

Kim, S. (2015). *ppcor: Partial and semi-partial (part) correlation* [R package version 1.1]. https://CRAN.R-project.org/package=ppcor

Kolk, H. H. J., Chwilla, D. J., van Herten, M., & Oor, P. J. W. (2003). Structure and limited capacity in verbal working memory: A study with event-related potentials. *Brain and Language*, *85*(1), 1–36. https://doi.org/10.1016/S0093-934X(02)00548-5

Kos, M., Vosse, T. G., Van Den Brink, D., & Hagoort, P. (2010). About edible restaurants: Conflicts between syntax and semantics as revealed by ERPs. *Frontiers in Psychology*, *1*. https://doi.org/10.3389/fpsyg.2010.00222

Kounios, J., Green, D. L., Payne, L., Fleck, J. I., Grondin, R., & McRae, K. (2009). Semantic richness and the activation of concepts in semantic memory: Evidence from event-related potentials. *Brain Research*, *1282*, 95–102. https://doi.org/10.1016/j.brainres.2009.05.092

Kounios, J., & Holcomb, P. J. (1994). Concreteness effects in semantic processing: ERP evidence supporting dual-coding theory. *Journal of Experimental Psychology:*

Learning, Memory, and Cognition, *20*(4), 804–823. https://doi.org/10.1037/0278-7393.20.4.804

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86. https://doi.org/10.1214/aoms/1177729694

Kumar, M., Federmeier, K. D., & Beck, D. M. (2021). The N300: An index for predictive coding of complex visual objects and scenes. *Cerebral Cortex Communications*, *2*(2), tgab030. https://doi.org/10.1093/texcom/tgab030

Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research*, *1146*, 23–49. https://doi.org/10.1016/j.brainres.2006.12.063

Kuperberg, G. R., Brothers, T., & Wlotko, E. W. (2020). A tale of two positivities and the N400: Distinct neural signatures are evoked by confirmed and violated predictions at different levels of representation. *Journal of Cognitive Neuroscience*, *32*(1), 12–35. https://doi.org/10.1162/jocn_a_01465

Kuperberg, G. R., Holcomb, P. J., Sitnikova, T., Greve, D., Dale, A. M., & Caplan, D. (2003). Distinct patterns of neural modulation during the processing of conceptual and syntactic anomalies. *Journal of Cognitive Neuroscience*, *15*(2), 272–293. https://doi.org/10.1162/089892903321208204

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31*(1), 32–59. https://doi.org/10.1080/23273798.2015.1102299

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203–205. https://doi.org/10.1126/science.7350657

Kutas, M. (1993). In the company of other words: Electrophysiological evidence for single-word and sentence context effects. *Language and Cognitive Processes*, *8*(4), 533–572. https://doi.org/10.1080/01690969308407587

Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences*, *4*(12), 463–470. https://doi.org/10.1016/S1364-6613(00)01560-6

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, *62*, 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*(5947), 161–163. https://doi.org/10.1038/307161a0

Kutas, M., Lindamood, T. E., & Hillyard, S. A. (1984). Word expectancy and event-related brain potentials during sentence processing. In S. Kornblum & J. Requin (Eds.), *Preparatory states and processes* (pp. 217–237). Erlbaum.

Laszlo, S., & Armstrong, B. C. (2014). PSPs and ERPs: Applying the dynamics of post-synaptic potentials to individual units in simulation of temporally extended event-related potential reading data. *Brain and Language*, *132*, 22–27. https://doi.org/10.1016/j.bandl.2014.03.002

Laszlo, S., & Federmeier, K. D. (2008). Minding the PS, queues, and PXQs: Uniformity of semantic processing across multiple stimulus types. *Psychophysiology*, *45*(3), 458–466. https://doi.org/10.1111/j.1469-8986.2007.00636.x

Laszlo, S., & Federmeier, K. D. (2009). A beautiful day in the neighborhood: An event-related potential study of lexical relationships and prediction in context. *Journal of Memory and Language*, *61*(3), 326–338. https://doi.org/10.1016/j.jml.2009.06.004

Laszlo, S., & Federmeier, K. D. (2011). The N400 as a snapshot of interactive processing: Evidence from regression analyses of orthographic neighbor and lexical associate effects. *Psychophysiology*, *48*(2), 176–186. https://doi.org/10.1111/j.1469-8986.2010.01058.x

Laszlo, S., & Plaut, D. C. (2012). A neurally plausible parallel distributed processing model of event-related potential word reading data. *Brain and Language*, *120*(3), 271–281. https://doi.org/10.1016/j.bandl.2011.09.001

Lau, E., Almeida, D., Hines, P. C., & Poeppel, D. (2009). A lexical basis for N400 context effects: Evidence from MEG. *Brain and Language*, *111*(3), 161–172. https://doi.org/10.1016/j.bandl.2009.08.007

Lau, E., Namyst, A., Fogel, A., & Delgado, T. (2016). A direct comparison of N400 effects of predictability and incongruity in adjective-noun combination. *Collabra: Psychology*, *2*(1), 13. https://doi.org/10.1525/collabra.40

Lau, E., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (De)constructing the N400. *Nature Reviews Neuroscience*, *9*(12), 920–933. https://doi.org/10.1038/nrn2532

Leckey, M., & Federmeier, K. D. (2020). The P3b and P600(s): Positive contributions to language comprehension. *Psychophysiology*, *57*(7), e13351. https://doi.org/10.1111/psyp.13351

Leckey, M., Troyer, M., & Federmeier, K. D. (2023). Patterns of hemispheric asymmetry provide evidence dissociating the semantic and syntactic P600. *Neuropsychologia*, *179*, 108441. https://doi.org/10.1016/j.neuropsychologia.2022.108441

Ledoux, K., Camblin, C. C., Swaab, T. Y., & Gordon, P. C. (2006). Reading words in discourse: The modulation of lexical priming effects by message-level context. *Behavioral and Cognitive Neuroscience Reviews*, *5*(3), 107–127. https://doi.org/10.1177/1534582306289573

Lelekov-Boissard, T., & Dominey, P. F. (2002). Human brain potentials reveal similar processing of non-linguistic abstract structure and linguistic syntactic structure. *Neurophysiologie Clinique/Clinical Neurophysiology*, *32*(1), 72–84. https://doi.org/10.1016/S0987-7053(01)00291-X

Levelt, W. J. M. (2013). *A history of psycholinguistics: The pre-Chomskyan era*. Oxford University Press.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Li, J., & Ettinger, A. (2023). Heuristic interpretation as rational inference: A computational model of the N400 and P600 in language processing. *Cognition*, *233*, 105359. https://doi.org/10.1016/j.cognition.2022.105359

Luck, S. J. (2005). *An introduction to the event-related potential technique*. The MIT Press.

Luke, S. G., & Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology*, *88*, 22–60. https://doi.org/10.1016/j.cogpsych.2016.06.002

Merkx, D., & Frank, S. L. (2021). Human sentence processing: Recurrence or attention? *Proceedings of the 2021 Workshop on Cognitive Modeling and Computational Linguistics*, 12–22. https://doi.org/10.18653/v1/2021.cmcl-1.2

Metusalem, R., Kutas, M., Urbach, T. P., Hare, M., McRae, K., & Elman, J. L. (2012). Generalized event knowledge activation during online sentence comprehension. *Journal of Memory and Language*, *66*(4), 545–567. https://doi.org/10.1016/j.jml.2012.01.001

Michaelov, J. A., Bardolph, M. D., Van Petten, C. K., Bergen, B. K., & Coulson, S. (2023). Strong prediction: Language model surprisal explains multiple N400 effects. *Neurobiology of Language*, 1–29. https://doi.org/10.1162/nol_a_00105

Michaelov, J. A., & Bergen, B. K. (2020). How well does surprisal explain N400 amplitude under different experimental conditions? *Proceedings of the 24th Conference on Computational Natural Language Learning*, 652–663. https://doi.org/10.18653/v1/2020.conll-1.53

Michalon, O., & Baggio, G. (2019). Meaning-driven syntactic predictions in a parallel processing architecture: Theory and algorithmic modeling of ERP effects. *Neuropsychologia*, *131*, 171–183. https://doi.org/10.1016/j.neuropsychologia.2019.05.009

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. https://doi.org/10.48550/arXiv.1301.3781

Münte, T. F., Heinze, H.-J., Matzke, M., Wieringa, B. M., & Johannes, S. (1998). Brain potentials and syntactic violations revisited: No evidence for specificity of the syntactic positive shift. *Neuropsychologia*, *36*(3), 217–226.

Nieuwland, M. S., Barr, D. J., Bartolozzi, F., Busch-Moreno, S., Darley, E., Donaldson, D. I., Ferguson, H. J., Fu, X., Heyselaar, E., Huettig, F., Husband, M. E., Ito, A., Kazanina, N., Kogan, V., Kohút, Z., Kulakova, E., Mézière, D., Politzer-Ahles, S., Rousselet, G., … Von Grebmer Zu Wolfsthurn, S. (2020). Dissociable effects of prediction and integration during language comprehension:

Evidence from a large-scale study using brain potentials. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1791), 20180522. https://doi.org/10.1098/rstb.2018.0522

Nieuwland, M. S., & Kuperberg, G. R. (2008). When the truth is not too hard to handle: An event-related potential study on the pragmatics of negation. *Psychological Science*, *19*(12), 1213–1218. https://doi.org/10.1111/j.1467-9280.2008.02226.x

Nieuwland, M. S., & van Berkum, J. J. A. (2005). Testing the limits of the semantic illusion phenomenon: ERPs reveal temporary semantic change deafness in discourse comprehension. *Cognitive Brain Research*, *24*(3), 691–701. https://doi.org/10.1016/j.cogbrainres.2005.04.003

Núñez-Peña, M. I., & Honrubia-Serrano, M. L. (2004). P600 related to rule violation in an arithmetic task. *Cognitive Brain Research*, *18*(2), 130–141. https://doi.org/10.1016/j.cogbrainres.2003.09.010

Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language*, *31*(6), 785–806. https://doi.org/10.1016/0749-596X(92)90039-Z

Osterhout, L., Holcomb, P. J., & Swinney, D. A. (1994). Brain potentials elicited by garden-path sentences: Evidence of the application of verb information during parsing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 786–803. https://doi.org/10.1037/0278-7393.20.4.786

Osterhout, L., McKinnon, R., Bersick, M., & Corey, V. (1996). On the language specificity of the brain response to syntactic anomalies: Is the syntactic positive shift a member of the P300 family? *Journal of Cognitive Neuroscience*, *8*(6), 507–526. https://doi.org/10.1162/jocn.1996.8.6.507

Osterhout, L., & Mobley, L. A. (1995). Event-related brain potentials elicited by failure to agree. *Journal of Memory and Language*, *34*(6), 739–773. https://doi.org/10.1006/jmla.1995.1033

Otten, M., & van Berkum, J. J. A. (2008). Discourse-based word anticipation during language processing: Prediction or priming? *Discourse Processes*, *45*(6), 464–496. https://doi.org/10.1080/01638530802356463

Palaz, B., Rhodes, R., & Hestvik, A. (2020). Informative use of "not" is N400-blind. *Psychophysiology*, *57*(12), e13676. https://doi.org/10.1111/psyp.13676

Parviz, M., Johnson, M., Johnson, B., & Brock, J. (2011). Using language models and Latent Semantic Analysis to characterise the N400m neural response. *Proceedings of the Australasian Language Technology Association Workshop 2011*, 38–46.

Patel, A. D., Gibson, E., Ratner, J., Besson, M., & Holcomb, P. J. (1998). Processing syntactic relations in language and music: An event-related potential study. *Journal of Cognitive Neuroscience*, *10*(6), 717–733. https://doi.org/10.1162/089892998563121

Payne, B. R., & Federmeier, K. D. (2017). Pace yourself: Intraindividual variability in context use revealed by self-paced event-related brain potentials. *Journal of Cognitive Neuroscience*, *29*(5), 837–854. https://doi.org/10.1162/jocn_a_01090

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Prolific. (2021). Prolific [Accessed March 30, 2021]. https://www.prolific.co/

Quante, L., Bölte, J., & Zwitserlood, P. (2018). Dissociating predictability, plausibility and possibility of sentence continuations in reading: Evidence from late-positivity ERPs. *PeerJ*, *6*, e5717. https://doi.org/10.7717/peerj.5717

Rabovsky, M., Hansen, S. S., & McClelland, J. L. (2018). Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour*, *2*(9), 693–705. https://doi.org/10.1038/s41562-018-0406-4

Rabovsky, M., & McClelland, J. L. (2020). Quasi-compositional mapping from form to meaning: A neural network-based approach to capturing neural responses during human language comprehension. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1791), 20190313. https://doi.org/10.1098/rstb.2019.0313

Rabovsky, M., & McRae, K. (2014). Simulating the N400 ERP component as semantic network error: Insights from a feature-based connectionist attractor model of word meaning. *Cognition*, *132*(1), 68–89. https://doi.org/10.1016/j.cognition.2014.03.010

Regel, S., Gunter, T. C., & Friederici, A. D. (2010). Isn't it ironic? An electrophysiological exploration of figurative language processing. *Journal of Cognitive Neuroscience*, *23*(2), 277–293. https://doi.org/10.1162/jocn.2010.21411

Rich, S., & Harris, J. (2021). Unexpected guests: When disconfirmed predictions linger. *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*, 2246–2252.

Roark, B., Bachrach, A., Cardenas, C., & Pallier, C. (2009). Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 324–333.

Rugg, M. D. (1985). The effects of semantic priming and word repetition on event-related potentials. *Psychophysiology*, *22*(6), 642–647. https://doi.org/10.1111/j.1469-8986.1985.tb01661.x

Rugg, M. D. (1990). Event-related brain potentials dissociate repetition effects of high-and low-frequency words. *Memory & Cognition*, *18*(4), 367–379. https://doi.org/10.3758/BF03197126

Ryskin, R., Stearns, L., Bergen, L., Eddy, M., Fedorenko, E., & Gibson, E. (2021). An ERP index of real-time error correction within a noisy-channel framework of

human communication. *Neuropsychologia*, *158*, 107855. https://doi.org/10.1016/j.neuropsychologia.2021.107855

Schacht, A., Sommer, W., Shmuilovich, O., Martíenz, P. C., & Martín-Loeches, M. (2014). Differential task effects on N400 and P600 elicited by semantic and syntactic violations. *PLOS ONE*, *9*(3), e91226. https://doi.org/10.1371/journal.pone.0091226

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-prime: User's guide*. Psychology Software Incorporated.

Schumacher, P. B. (2011). The hepatitis called...: Electrophysiological evidence for enriched composition. In M. J & S. M (Eds.), *Experimental pragmatics/semantics* (pp. 199–2019). John Benjamins Publishing.

Schumacher, P. B. (2013). When combinatorial processing results in reconceptualization: Toward a new approach of compositionality. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00677

Schwanenflugel, P. J., & LaCount, K. L. (1988). Semantic relatedness and the scope of facilitation for upcoming words in sentences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(2), 344–354. https://doi.org/10.1037/0278-7393.14.2.344

Schwartz, T. J., Federmeier, K. D., Van Petten, C., Salmon, D. P., & Kutas, M. (2003). Electrophysiological analysis of context effects in Alzheimer's disease. *Neuropsychology*, *17*(2), 187–201. https://doi.org/10.1037/0894-4105.17.2.187

Shapiro, S. S., & Francia, R. (1972). An approximate analysis of variance test for normality. *Journal of the American Statistical Association*, *67*(337), 215–216. https://doi.org/10.1080/01621459.1972.10481232

Sitnikova, T., Holcomb, P. J., Kiyonaga, K. A., & Kuperberg, G. R. (2008). Two neurocognitive mechanisms of semantic integration during the comprehension of visual real-world events. *Journal of Cognitive Neuroscience*, *20*(11), 2037–2057. https://doi.org/10.1162/jocn.2008.20143

Smith, N. J., & Kutas, M. (2015a). Regression-based estimation of ERP waveforms: I. The rERP framework. *Psychophysiology*, *52*(2), 157–168. https://doi.org/10.1111/psyp.12317

Smith, N. J., & Kutas, M. (2015b). Regression-based estimation of ERP waveforms: II. Nonlinear effects, overlap correction, and practical considerations. *Psychophysiology*, *52*(2), 169–181. https://doi.org/10.1111/psyp.12320

Smith, N. J., & Levy, R. (2008). Optimal processing times in reading: A formal model and empirical investigation. *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*, 595–600.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319. https://doi.org/10.1016/j.cognition.2013.02.013

Spotorno, N., Cheylus, A., Henst, J.-B. V. D., & Noveck, I. A. (2013). What's behind a P600? Integration operations during irony processing. *PLOS ONE*, *8*(6), e66839. https://doi.org/10.1371/journal.pone.0066839

Spychalska, M., Kontinen, J., & Werning, M. (2016). Investigating scalar implicatures in a truth-value judgement task: Evidence from event-related brain potentials. *Language, Cognition and Neuroscience*, *31*(6), 817–840. https://doi.org/10.1080/23273798.2016.1161806

Stone, K., Nicenboim, B., Vasishth, S., & Rösler, F. (2023). Understanding the effects of constraint and predictability in ERP. *Neurobiology of Language*, *4*(2), 221–256. https://doi.org/10.1162/nol_a_00094

Stowe, L. A., Kaan, E., Sabourin, L., & Taylor, R. C. (2018). The sentence wrap-up dogma. *Cognition*, *176*, 232–247. https://doi.org/10.1016/j.cognition.2018.03.011

Szewczyk, J. M., & Federmeier, K. D. (2022). Context-based facilitation of semantic access follows both logarithmic and linear functions of stimulus probability. *Journal of Memory and Language*, *123*, 104311. https://doi.org/10.1016/j.jml.2021.104311

Szewczyk, J. M., & Schriefers, H. (2011). Is animacy special?: ERP correlates of semantic violations and animacy violations in sentence processing. *Brain Research*, *1368*, 208–221. https://doi.org/10.1016/j.brainres.2010.10.070

Tales, A., Newton, P., Troscianko, T., & Butler, S. (1999). Mismatch negativity in the visual modality. *Neuroreport*, *10*(16), 3363–3367. https://doi.org/10.1097/00001756-199911080-00020

Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly*, *30*(4), 415–433. https://doi.org/10.1177/107769905303000401

Thornhill, D. E., & Van Petten, C. (2012). Lexical versus conceptual anticipation during sentence processing: Frontal positivity and N400 ERP components. *International Journal of Psychophysiology*, *83*(3), 382–392. https://doi.org/10.1016/j.ijpsycho.2011.12.007

Troyer, M., & Kutas, M. (2020). To catch a Snitch: Brain potentials reveal variability in the functional organization of (fictional) world knowledge during reading. *Journal of Memory and Language*, *113*, 104111. https://doi.org/10.1016/j.jml.2020.104111

Troyer, M., Urbach, T. P., & Kutas, M. (2020). Lumos!: Electrophysiological tracking of (wizarding) world knowledge use during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(3), 476–486. https://doi.org/10.1037/xlm0000737

Urbach, T. P., DeLong, K. A., Chan, W.-H., & Kutas, M. (2020). An exploratory data analysis of word form prediction during word-by-word reading. *Proceedings of the National Academy of Sciences*, *117*(34), 20483–20494. https://doi.org/10.1073/pnas.1922028117

Van Petten, C. (1993). A comparison of lexical and sentence-level context effects in event-related potentials. *Language and Cognitive Processes*, *8*(4), 485–531. https://doi.org/10.1080/01690969308407586

Van Petten, C., & Kutas, M. (1990). Interactions between sentence context and word frequency in event-related brain potentials. *Memory & Cognition*, *18*(4), 380–393. https://doi.org/10.3758/BF03197127

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*, *83*(2), 176–190. https://doi.org/10.1016/j.ijpsycho.2011.09.015

Van Petten, C., Weckerly, J., McIsaac, H. K., & Kutas, M. (1997). Working memory capacity dissociates lexical and sentential context effects. *Psychological Science*, *8*(3), 238–242. https://doi.org/10.1111%2Fj.1467-9280.1997.tb00418.x

van Berkum, J. J. A. (2009). The neuropragmatics of 'simple' utterance comprehension: An ERP review. In U. Sauerland & K. Yatsushiro (Eds.), *Semantics and pragmatics: From experiment to theory* (pp. 276–316). Palgrave Macmillan.

van Berkum, J. J. A. (2010). The brain is a prediction machine that cares about good and bad - Any implications for neuropragmatics? *Italian Journal of Linguistics*, *22*, 181–208.

van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(3), 443. https://doi.org/10.1037/0278-7393.31.3.443

van Herten, M., Chwilla, D. J., & Kolk, H. H. J. (2006). When heuristics clash with parsing routines: ERP evidence for conflict monitoring in sentence perception. *Journal of Cognitive Neuroscience*, *18*(7), 1181–1197. https://doi.org/10.1162/jocn.2006.18.7.1181

van Herten, M., Kolk, H. H. J., & Chwilla, D. J. (2005). An ERP study of P600 effects elicited by semantic anomalies. *Cognitive Brain Research*, *22*(2), 241–255. https://doi.org/10.1016/j.cogbrainres.2004.09.002

Venhuizen, N. J., Crocker, M. W., & Brouwer, H. (2019). Expectation-based comprehension: Modeling the interaction of world knowledge and linguistic experience. *Discourse Processes*, *56*(3), 229–255. https://doi.org/10.1080/0163853X.2018.1448677

Vissers, C. T., Chwilla, D. J., & Kolk, H. H. (2006). Monitoring in language perception: The effect of misspellings of words in highly constrained sentences. *Brain Research*, *1106*(1), 150–163. https://doi.org/10.1162/jocn.2008.21170

Võ, M. L. H., & Wolfe, J. M. (2013). The interplay of episodic and semantic memory in guiding repeated search in scenes. *Cognition*, *126*(2), 198–212. https://doi.org/10.1016/j.cognition.2012.09.017

West, W. C., & Holcomb, P. J. (2000). Imaginal, semantic, and surface-level processing of concrete and abstract words: An electrophysiological investigation. *Journal of Cognitive Neuroscience*, *12*(6), 1024–1037. https://doi.org/10.1162/08989290051137558

Xu, X., & Zhou, X. (2016). Topic shift impairs pronoun resolution during sentence comprehension: Evidence from event-related potentials. *Psychophysiology*, *53*(2), 129–142. https://doi.org/10.1111/psyp.12573

Zehr, J., & Schwarz, F. (2018). PennController for Internet Based Experiments (IBEX). https://doi.org/10.17605/OSF.IO/MD832