Rational Speech Comprehension: Effects of Predictability and Background Noise

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

vorgelegt von
Marjolein van Os
aus Groningen, die Niederlande

Saarbrücken, 2023

Dekan der Fakultät P: Prof. Dr. Stefanie Haberzettl
Erstberichterstatterin: Prof. Dr. Vera Demberg
Zweitberichterstatterin: Prof. Dr. Jutta Kray
Tag der letzten Prüfungsleistung: 04. August 2023

In Memoriam

Johan van Os

Abstract

Having a conversation is something we are capable of without a second thought. However, this is not as simple as it seems. While there can be a myriad of difficulties arising, one common occurrence is the presence of background noise during every-day language use. This negatively affects speech recognition on the listener's part. One strategy the listener has to cope with this, is to rely on predictive processes, where the upcoming words of the speaker are predicted from the context. The present dissertation concerns the interplay of background noise and prediction.

This interplay can be summarized by the question whether listeners rely more on bottom-up information (the acoustic speech signal) or top-down information (for example context-based predictions). While previous studies have investigated how background noise and context predictability interact, and how different speech sounds are affected by (different types of) background noise, the three factors have so far not been studied together in a single experiment. These manipulations of the listening condition result in fine-grained differences in the intelligibility of the speech signal and subsequently affect to what extent listeners rely on the bottom-up information. This allows us to test the predictions of the Noisy Channel Model, a model that explains human speech comprehension in background noise. So far, the predictions of this model have been primarily tested using written stimuli and never with acoustical noise. In the word recognition experiments we collect confidence ratings to investigate false hearing effects. Additionally, we were interested in consequences of these adverse listening conditions and addressed the question how recognition memory is affected by background noise and contextual predictability, studying false memory effects in particular.

This dissertation presents results from three experiments that were set up to address these questions. The first two experiments used a word recognition paradigm in which participants listened to recordings of sentences that were embedded in background noise in some conditions. Our stimuli contained different speech sound contrasts. We varied the level as well as the type of background noise across the two experiments. Because they lead to varying degrees of relying on either bottom-up or top-down processes, these differences in listening conditions allowed us to test the predictions of the Noisy Channel Model. Across experiments, we also varied the tested population. In Experiment 2 we recruited older adults as well as younger adults. As older adults differ from younger adults in their trade-off between top-down and bottom-up information. This allowed us to test their recognition accuracy in listen-

ing situations with small-grained differences regarding the intelligibility (through the overlap between the speech signal and noise). We can further test the predictions of the Noisy Channel Model, and investigate false hearing effects. The third experiment investigated consequences of listening in background noise, as we tested the effects of both noise and predictability on recognition memory.

Taken together, the results from the three experiments lead us to four conclusions. First, we find that the three factors of noise type, speech sound, and predictability of the context interact during speech comprehension. Second, this leads to small-grained differences in the intelligibility of the stimuli, which in turn affects how the listener relies on either the bottom-up signal or top-down predictions. Across the experiments, our findings support the predictions made by the Noisy Channel Model, namely that the reliance on predictive processes is dependent on the effort that is required to process the speech signal and the amount of overlap between speech and background noise. Third, this was the case even for older adults, who did not show the expected false hearing effects. Instead, they behaved rationally, taking into account possible age-related hearing loss and a stronger reliance on prediction. Finally, we additionally showed that the difficulty of the listening condition affects metacognitive judgements, operationalised through confidence ratings: the more difficult the listening condition, the lower listeners' confidence was, both for word recognition and memory. Against our expectations, we did not find evidence of false memory. Future studies should investigate in exactly which situations this effect occurs, as our experimental design differed from those of previous studies.

In sum, the findings in this dissertation contribute to our understanding of speech recognition in adverse listening conditions, in particular background noise, and of how predictive processes can both help and hinder speech perception. Our results consistently support the predictions of the Noisy Channel Model, indicating that human listeners behave rationally. Their reliance on either the bottom-up acoustic signal or top-down predictions depends on the clarity of the speech signal, and here they take into account fine-grained differences. In this way, they can maximize the chance of successful communication while minimizing effort.

Zussamenfassung

Wenn wir in unserem täglichen Leben Sprache verwenden, geschieht dies sehr selten in einer ruhigen Umgebung. Die meiste Zeit, in der wir Sprache unter natürlichen Bedingungen hören, sind Hintergrundgeräusche vorhanden. Wir nehmen dies kaum wahr, und in der Regel ist der Geräuschpegel so gering, dass er nicht zu großen Störungen der Kommunikation führt. Denken Sie an das Brummen von vorbeifahrenden Autos, an Gespräche anderer Menschen, an die Arbeit von Maschinen oder an das Geräusch von Tellern und Besteck in einem Restaurant. Wir sind in der Lage, fast mühelos mit anderen zu kommunizieren, obwohl diese Geräusche mit dem Sprachsignal konkurrieren.

Wie schaffen Menschen das? Ist dieser Prozess des Zuhörens bei Hintergrundgeräuschen wirklich so mühelos, wie es scheint? Und können die Strategien, die zur Bewältigung der zusätzlichen kognitiven Belastung eingesetzt werden, auch Nachteile mit sich bringen? Ziel der vorliegenden Dissertation ist es, die Literatur zum Sprachverständnis bei Hintergrundgeräuschen weiterzuentwickeln, indem die folgenden Fragen beantwortet werden: Wie interagieren Vorhersagbarkeit, Hintergrundgeräusche und Sprachgeräusche in den Reizen? Gibt es bei dieser Interaktion Unterschiede zwischen verschiedenen Arten von Hintergrundgeräuschen? Wie lassen sich diese Unterschiede in der Erkennungsgenauigkeit erklären? Darüber hinaus untersuchen wir, wie sich diese Hörbedingungen auf ältere Erwachsene auswirken und welche Auswirkungen sie auf die Kommunikation über die Worterkennung hinaus haben.

Eine Strategie, die dazu beiträgt, die Belastung des Sprachverstehens durch Hintergrundgeräusche zu verringern, besteht darin, sich auf Vorhersagen zu stützen. Oft ist es möglich, z. B. Weltwissen, Wissen über den Sprecher oder den vorangegangenen Kontext in einem Dialog zu nutzen, um vorherzusagen, was als Nächstes gesagt werden könnte. Verschiedene Studien haben ergeben, dass ein hoher Vorhersagewert eines Satzes den Zuhörern in ruhiger Umgebung hilft, ihn zu verarbeiten, aber auch besonders in lauten Hörsituationen, wo er zu einer besseren Erkennung führt als ein niedriger Vorhersagewert (Boothroyd & Nittrouer, 1988; Dubno et al., 2000; Hutchinson, 1989; Kalikow et al., 1977; Pichora-Fuller et al., 1995; Sommers & Danielson, 1999). Verschiedene Spracherkennungstheorien berücksichtigen dies und erklären, wie Bottom-up-Informationen aus dem auditiven Sprachsignal mit Top-down-Informationen kombiniert werden, z. B. mit Vorhersagen, die auf dem Kontext basieren (Levy, 2008; Levy et al., 2009; Luce & Pisoni, 1998; Norris & McQueen,

2008; Oden & Massaro, 1978). Diese Interaktion von Bottom-Up- und Top-Down-Informationsströmen bildet die Grundlage dieser Dissertation.

Bislang wurde in empirischen Studien untersucht, wie sich die Variation der Vorhersagbarkeit des Zielwortes auf das Sprachverstehen im Störgeräusch auswirkt (Boothroyd & Nittrouer, 1988; Dubno et al., 2000; Hutchinson, 1989; Kalikow et al., 1977; Pichora-Fuller et al., 1995; Sommers & Danielson, 1999). Die Ergebnisse dieser Studien stimmen überein: Ein vorhersehbarer Kontext erleichtert das Sprachverstehen im Störgeräusch. Ein anderer Zweig der Literatur hat untersucht, wie Hintergrundgeräusche die Erkennung verschiedener Phoneme beeinflussen (Alwan et al., 2011; Cooke, 2009; Gordon-Salant, 1985; Phatak et al., 2008; Pickett, 1957). Diese Studien konzentrieren sich oft auf die isolierten Phoneme und testen feste Kontexte in Nonsense-Silben statt in Wörtern. Somit wird die Vorhersagbarkeit der Elemente nicht manipuliert, obwohl diese die Erkennung beeinflusst. Die drei Faktoren Kontextvorhersagbarkeit, Hintergrundgeräusche und Phoneme wurden bisher noch nicht gemeinsam in einer Studie untersucht. Es ist möglich, dass diese Faktoren zusammenwirken, z. B. der Effekt der Vorhersagbarkeit auf bestimmte Phoneme könnte stärker sein, wenn Hintergrundgeräusche die Erkennung behindern. Die vorliegende Dissertation soll diese Lücke in der Literatur füllen.

In diesem Zusammenhang wollen wir untersuchen, wie sich verschiedene Arten von Lärm auf die Spracherkennung auswirken. In früheren Arbeiten wurde diese Frage bereits untersucht, doch die Ergebnisse waren nicht eindeutig (Danhauer & Leppler, 1979; Gordon-Salant, 1985; Horii et al., 1971; Nittrouer et al., 2003; Taitelbaum-Swead & Fostick, 2016). Einige Studien kommen zu dem Ergebnis, dass weißes Rauschen zu größeren Interferenzen führt, während andere Studien feststellen, dass Babbelgeräusche oder sprachförmige Geräusche schwieriger sind. Diese unterschiedlichen Ergebnisse in Bezug auf die Art des Lärms lassen vermuten, dass hier andere Faktoren eine Rolle spielen, z. B. die Eigenschaften der getesteten Stimuli. Dieser offenen Frage wollen wir nachgehen.

Ein Modell, das vorgeschlagen wurde, um das menschliche Sprachverstehen bei Hintergrundgeräuschen zu erklären, ist das Noisy Channel Model (Levy, 2008; Levy et al., 2009; Shannon, 1949). Diesem Modell zufolge kombinieren Hörer auf rationale Weise Bottom-up-Informationen mit Top-down-Informationen, wobei der Rückgriff auf eine der beiden Arten von Informationen von der Klarheit der Hörsituation abhängt. Frühere Studien haben die Vorhersagen dieses Modells vor allem im schriftlichen Bereich getestet, indem sie syntaktische Änderungen verwendeten und die Interpretation unplausibler Sätze prüften (Gibson et al., 2013; Poppels &

Levy, 2016; Ryskin et al., 2018). Sie manipulierten den Grad des wahrgenommenen Lärms durch die Anzahl der Füllwörter mit syntaktischen Fehlern und qualifizierten den Abstand zwischen den plausiblen und unplausiblen Sätzen durch die Anzahl der eingefügten und gelöschten Wörter. In den Studien wurden Belege gefunden, die die Vorhersagen des Noisy Channel Models unterstützen. Ein Schritt in Richtung eines naturalistischeren Sprachverständnisses im Rauschen wurde unternommen, indem die gleichen Stimuli in gesprochener Form getestet wurden (Gibson et al., 2016; Gibson et al., 2017). Obwohl das Noisy Channel Model konstruiert wurde, um das menschliche Sprachverständnis bei Lärm zu erklären, wurden seine Vorhersagen bisher in keiner Studie bei akustischem Lärm untersucht. Die vorliegende Dissertation soll neue Erkenntnisse liefern. Unsere Stimuli unterscheiden sich von den bisher überwiegend getesteten und sind so konstruiert, dass die Vorhersagbarkeit des Zielwortes sowie die Überlappung zwischen Sprach- und Geräuschsignal variiert. Somit würden unsere Ergebnisse die Situationen erweitern, in denen die Vorhersagen des Noisy Channel Model zutreffen könnten. Wir werden dies in verschiedenen Hörsituationen mit unterschiedlichen Arten von Hintergrundgeräuschen, Stimuli-Charakteristika und Populationen testen. Außerdem setzen wir das Noisy Channel Model in Beziehung zu anderen Modellen der Sprachwahrnehmung.

Die Unterschiede zwischen jüngeren und älteren Erwachsenen erlauben uns, die Vorhersagen des Geräuschkanalmodells im Detail zu testen. Ältere Erwachsene haben einen anderen Trade-Off zwischen Top-down- und Bottom-up-Informationen als jüngere Erwachsene. Ihr Gehör ist durch altersbedingten Hörverlust beeinträchtigt, und diese Beeinträchtigung führt zu größeren Schwierigkeiten beim Verstehen von Sprache unter ungünstigen Hörbedingungen (Li et al., 2004; Pichora-Fuller et al., 1995; Pichora-Fuller et al., 2017; Schneider et al., 2005). Andererseits bleiben ihre prädiktiven Prozesse intakt, und es hat sich gezeigt, dass ältere Erwachsene sich stärker auf diese verlassen, um Hörprobleme zu überwinden (Stine & Wingfield, 1994; Wingfield et al., 1995; Wingfield et al., 2005). Diese Unterschiede im Vergleich zu jüngeren Erwachsenen machen die Population der älteren Erwachsenen theoretisch besonders interessant für Tests. Wir wollen untersuchen, ob es tatsächlich so ist, dass ältere Erwachsene in Fällen, in denen der Satzkontext irreführend ist, mehr Hörfehler zeigen als jüngere Erwachsene. Wir variieren die Hörbedingungen, um feinkörnige Unterschiede in der Überlappung zwischen dem Sprachsignal und dem Hintergrundgeräusch zu konstruieren, wodurch sich der Schwierigkeitsgrad der Hörbedingung ändert.

Einer der Unterschiede zwischen jüngeren und älteren Erwachsenen wurde in Bezug auf den Effekt des "false hearing" festgestellt (Failes et al., 2020; Failes & Sommers, 2022; Rogers et al., 2012; Rogers & Wingfield, 2015; Rogers, 2017; Sommers et al., 2015). Dabei handelt es sich um ein Phänomen, bei dem ein Hörer sehr sicher ist, ein bestimmtes Wort richtig erkannt zu haben, aber in Wirklichkeit falsch liegt. In diesen Fällen wurde das Wort oft durch Top-Down-Prozesse und nicht durch das akustische Sprachsignal selbst erkannt. Da sie im Allgemeinen stärker auf prädiktive Prozesse angewiesen sind, hat man festgestellt, dass der false-hearing-Effekt bei älteren Erwachsenen größer ist als bei jüngeren. Wir wollen dies untersuchen und erwarten, dass der false-hearing-Effekt stärker ist, wenn die Hörbedingungen schwieriger sind (aufgrund des Geräuschpegels oder einer größeren Überlappung zwischen den Sprachklängen in den Stimuli und dem Hintergrundgeräusch). Die Vertrauensbewertungen, die die Versuchspersonen abgeben, werden zusätzlich Aufschluss über ihre metakognitiven Prozesse während des Hörens geben (siehe unten).

In den meisten Studien, die das Sprachverständnis untersuchen, werden die Versuchspersonen gebeten, einfach zu berichten, was sie gehört haben. Anschließend wird die Genauigkeit ermittelt. Dies beantwortet zwar die Fragen nach der Verständlichkeit von Sprache und der Schwierigkeit der Hörbedingungen, lässt aber andere Punkte offen. Im Alltag wird Sprache zur Kommunikation genutzt, die mehr erfordert als die Wiedergabe des Gehörten. Daher sollte untersucht werden, wie sich unterschiedliche Hörbedingungen auf nachfolgende übergeordnete Prozesse auswirken, die in der Kommunikation häufig eine Rolle spielen, um festzustellen, wie sich das Hören im Lärm (oder unter anderen, möglicherweise ungünstigen Bedingungen) auf das Gespräch zwischen Gesprächspartnern auswirkt, das über das bloße Erkennen des Gesagten hinausgeht. In dieser Dissertation wollen wir die Folgen des Sprachverstehens unter verschiedenen Hörbedingungen testen, wobei wir den Hintergrundlärm und die Vorhersagbarkeit variieren. Einerseits bitten wir die Versuchspersonen uns nach jedem experimentellen Versuch mitzuteilen, wie sicher sie sich sind, die richtige Antwort gegeben zu haben. Auf diese Weise können wir nicht nur den false-hearing-Effekt untersuchen, sondern auch feststellen, wie sie die Hörbedingungen erlebt haben und wie sich dies auf ihre subjektive Empfindung das Wort verstanden zu haben. Dies kann damit zusammenhängen, ob sie die richtige Antwort geben oder nicht, muss es aber nicht. Andererseits legen wir den Versuchspersonen in einem unserer Experimente nach der Hörphase einen Gedächtnistest vor, um zu testen, wie sich die Schwierigkeit der Geräuschbedingung und die Vorhersagbarkeit des Zielworts auf die spätere Erinnerung auswirken. Zu wissen, wie das Gedächtnis durch die Hörbedingungen beeinflusst wird, ist wichtig, da es Aufschluss über den Umgang mit Situationen

geben kann, in denen Hintergrundgeräusche unvermeidlich sind, aber Anweisungen verstanden und erinnert werden müssen.

Frühere Arbeiten haben die Auswirkungen der Vorhersagbarkeit auf die Gedächtnisleistung untersucht und dabei ein interessantes Phänomen aufgedeckt. In den Studien wurde festgestellt, dass Wörter, die vorhergesagt, den Versuchspersonen aber nicht tatsächlich präsentiert wurden, im Gedächtnis bleiben und die Gedächtnisleistung später in Form von so genannten falschen Erinnerungen beeinträchtigen (Haeuser & Kray, 2022a; Hubbard et al., 2019). In diesen Studien geben die Versuchspersonen an, sich an diese Begriffe zu erinnern, obwohl sie sie gar nicht gesehen haben. Dieser Effekt wurde bisher noch nicht für Elemente untersucht, die in Hintergrundgeräusche eingebettet sind, und wir wollen testen, ob wir hier mehr falsche Erinnerungen finden werden, da sich die Versuchspersonen unter solch schwierigen Hörbedingungen stärker auf Vorhersageprozesse verlassen, was zu falschen Erinnerungen führt.

Wir haben drei Experimente durchgeführt, um diese Fragen zu beantworten. Im ersten Experiment manipulierten wir die Vorhersagbarkeit (hoch oder niedrig), die Art des Geräuschs (Babbel oder weißes Rauschen) und das akustische Signal (verschiedene Sprachlaute; Plosive, Vokale und Frikative) und untersuchten so die Interaktion der drei Faktoren. Diese Versuchsanordnung ermöglicht es uns auch, die Auswirkungen von Babbelgeräuschen auf der einen und weißem Rauschen auf der anderen Seite zu vergleichen, um die bisher nicht eindeutigen Ergebnisse in der Literatur zu berücksichtigen (Danhauer & Leppler, 1979; Horii et al., 1971; Gordon-Salant, 1985; Nittrouer et al., 2003; Taitelbaum-Swead & Fostick, 2016). Das Experiment testet die Vorhersagen des Noisy-Channel-Modells für das Verstehen gesprochener Sprache. Dieses Modell wurde vorgeschlagen, um das Sprachverstehen in verrauschten Umgebungen zu erklären (Levy, 2008; Levy et al., 2009; Shannon, 1949). Es geht davon aus, dass Versteher sich nicht ausschließlich auf den Bottom-Up-Signal verlassen, sondern das Eingangssignal rational mit Top-Down-Vorhersagen kombinieren. Bislang wurde diese Hypothese vor allem im schriftlichen Bereich getestet, wo die Top-Down-Vorhersagen mit dem Bottom-Up-Signal in Konflikt stehen. Unsere Hörbedingungen unterscheiden sich hinsichtlich ihrer Editierdistanz in einer Weise, die auf früheren Arbeiten über die Verwechselbarkeit von Sprachlauten im Lärm beruht und zu feinkörnigen Unterschieden führt. Unsere Ergebnisse stimmen mit den Vorhersagen des Noisy-Channel-Modells überein: Hörer kombinieren probabilistisch Top-Down-Vorhersagen, die auf dem Kontext basieren, mit verrauschten Bottom-Up-Informationen aus dem akustischen Signal, um gesprochene Sprache besser zu verstehen.

Das zweite Experiment testet das Sprachverstehen bei gleichzeitigen Störgeräuschen bei jüngeren und älteren Erwachsenen. Diese beiden Populationen unterscheiden sich hinsichtlich der Spracherkennung und der Art und Weise, wie Informationen aus dem Bottom-Up-Audiosignal und Top-Down-Vorhersagen kombiniert werden. So können wir die Vorhersagen des Noisy-Channel-Modells anhand dieser beiden Populationen weiter testen. Darüber hinaus haben frühere Studien, insbesondere bei älteren Erwachsenen, false-hearing-Effekte festgestellt, bei dem sie sehr zuversichtlich waren, während der Worterkennung eine korrekte Antwort zu geben, die aber tatsächlich falsch war (Failes et al., 2020; Failes & Sommers, 2022; Rogers et al., 2012; Rogers & Wingfield, 2015; Rogers, 2017; Sommers et al., 2015). Wir untersuchen dieses Phänomen und gehen der Frage nach, inwieweit die metakognitiven Entscheidungen unserer Versuchspersonen ihr Vertrauen in das akustische Signal oder den semantischen Kontext widerspiegeln. Auch hier variieren wir das Rauschen, konzentrieren uns aber jetzt auf den Rauschpegel (0 SNR dB oder -5 SNR dB) und nicht auf die Art des Rauschens (wir verwenden nur Babble Noise). Unsere Ergebnisse zeigen, dass das Ausmaß des Fehlhörens direkt mit der Wahrnehmbarkeit des Sprachsignals zusammenhängt und dass die eigen Einschätzung der Versuchspersonen, ob das Wort im Signal so gesprochen wurde, ebenfalls damit übereinstimmen: Je mehr Überschneidungen zwischen Sprachsignal und Störgeräusch, desto weniger waren die Versuchspersonen sich sicher, dass sie die richtige Antwort gegeben haben. Wir finden keine Hinweise auf false-hearing-Effekte, was auf die von uns getestete Population zurückzuführen sein könnte. Stattdessen hing die Höhe der Sicherheitsbewertungen von der Schwierigkeit der Hörbedingung ab.

Das dritte Experiment untersucht die Folgen des Hörens von mehr oder weniger vorhersehbarer Sprache im Hintergrundgeräusch, indem es die Versuchspersonen nicht nur fragt, was sie gehört haben. Wir testen, wie das spätere Wiedererkennungsgedächtnis der Versuchspersonen durch diese verschiedenen Hörbedingungen beeinflusst wird. In früheren Studien wurden Effekte falschen Erinnerns festgestellt, bei denen Elemente, die (in hohem Maße) vorhergesagt, aber nicht tatsächlich präsentiert wurden, im Gedächtnis verbleiben, so dass sie von den Veruschspersonen in Gedächtnistests wiedergegeben werden (Haeuser & Kray, 2022a; Hubbard et al., 2019). Wir wollen testen, ob dieser Effekt bei Hintergrundgeräuschen stärker ist, da sich die Hörer unter diesen Bedingungen stärker auf prädiktive Prozesse verlassen. In unserem Experiment hörten die Versuchspersonen zunächst Sätze, die entweder in Störgeräusche eingebettet oder in Ruhe präsentiert wurden. Die Vorhersagbarkeit des Zielworts wurde durch Änderung der Wortreihenfolge des Satzes beeinflusst, und wir variierten auch die Häufigkeit der Zielwörter. In einem Test zum Wiedererkennen von Über-

raschungen fragten wir die Versuchspersonen, ob sie das dargebotene Wort schon einmal gesehen hatten, wobei es drei Arten von Wörtern gab: alte Begriffe, die sie schon einmal gehört hatten, neue Begriffe, die sie noch nie gehört hatten und die in keinem Zusammenhang mit den Sätzen standen, und semantische Köder, die sie noch nie gehört hatten, die aber semantisch mit den alten Begriffen verbunden waren. Die Ergebnisse zeigten, dass die Gedächtnisleistung für die semantischen Köder nicht durch Faktoren auf Satzebene, wie Hintergrundgeräusche oder Vorhersagbarkeit, beeinflusst wurde. Während dies bei den alten Elemente der Fall war, zeigen diese Ergebnisse zusammengenommen, dass wir keine Hinweise auf falsche Erinnerneffekte bei den Köder-Elemente gefunden haben, sondern dass die Genauigkeit bei den alten Elemente, ähnlich wie in den beiden vorangegangenen Experimenten, von der Schwierigkeit der Hörbedingung abhing.

Wir erweitern die Literatur, indem wir drei Faktoren kombinieren, die beim Sprachverstehen im Lärm interagieren, nämlich die Art des Lärms, bestimmte Sprachlaute in den Stimuli und die Vorhersagbarkeit des Kontexts. Die ersten beiden Faktoren interagieren, was zu einer Variation des Anteils des Sprachsignals führt, der verdeckt wird. In früheren Studien wurde diese Wechselwirkung zwar untersucht, allerdings vorwiegend bei isolierten Silben (Alwan et al., 2011; Cooke, 2009; Gordon-Salant, 1985; Phatak et al., 2008; Pickett, 1957). Da Hörer im Alltag in der Lage sind, Vorhersagen auf der Grundlage des semantischen Kontexts zu treffen, um ihren Erkennungsprozess zu steuern, ist es wichtig, auch den Effekt eines mehr oder weniger vorhersagenden Satzkontexts zu berücksichtigen. Wir zeigen, dass die drei Faktoren zusammenwirken und zu geringfügigen Unterschieden in der Verständlichkeit der Stimuli führen, je nach Art des Hintergrundrauschens und der Phoneme im akustischen Signal. Daher ist der Effekt der Vorhersagbarkeit stärker, wenn das Hintergrundgeräusch stärker mit einem bestimmten Sprachklang interferiert. Dies wirkt sich darauf aus, wie sehr sich der Hörer bei der Spracherkennung auf Top-down- oder Bottom-up-Prozesse verlässt. Der Rückgriff auf prädiktive Prozesse kann zu falsch erkannten Wörtern führen, wenn der Satzkontext irreführend ist.

Wir zeigen, dass die Vorhersagen des Noisy-Channel-Modells auch dann zutreffen, wenn sie in einem anderen Versuchsaufbau als bisher getestet werden, nämlich mit einem anderen Satz von Stimuli und in akustischem Hintergrundrauschen. Wir verwendeten gesprochene Stimuli, die in Hintergrundgeräusche unterschiedlicher Art und Lautstärke eingebettet waren, und manipulierten die wahrgenommene Geräuschmenge durch die Überlappung des Sprachsignals und des akustischen Rauschens, indem wir Stimuli mit unterschiedlichen Klangkontrasten konstruierten (Minimalpaare mit

Plosiven, Frikativen und Affrikaten sowie Vokalen). Da unter diesen verschiedenen Hörbedingungen die Klarheit des Sprachsignals und damit der erforderliche Verarbeitungsaufwand variiert, macht das Noisy Channel Model feinkörnige Vorhersagen darüber, wie sehr sich Hörer entweder auf das akustische Signal von unten nach oben oder auf prädiktive Prozesse von oben nach unten verlassen. Wir haben zusätzlich zwei verschiedene Populationen getestet, nämlich jüngere und ältere Erwachsene. Bei älteren Erwachsenen wurde bereits festgestellt, dass sie sich stärker auf prädiktive Prozesse verlassen als jüngere Erwachsene, so dass wir die Vorhersagen des Noisy-Channel-Modells weiter testen konnten. Auch hier wurden die Vorhersagen durch unsere Daten bestätigt. Insgesamt bieten unsere Ergebnisse eine zusätzliche und übereinstimmende Unterstützung für das Noisy-Channel-Modell.

Wir replizieren den Befund, dass ältere Erwachsene dazu neigen, sich mehr auf den Satzkontext zu verlassen als jüngere Erwachsene. Während dies in der bisherigen Literatur häufig berichtet wurde, zeigen wir diesen Effekt in einer Online-Studie mit einer jüngeren Versuchspersonengruppe (50-65 Jahre) als in diesen Studien üblich (65+). Wir haben auch den false-hearing-Effekt untersucht. Bei diesem Effekt ist das Vertrauen in falsche Antworten hoch, und es wurde festgestellt, dass dieser Effekt bei älteren Erwachsenen stärker ist als bei jüngeren Erwachsenen. Dies wird auf den altersbedingten Abbau der kognitiven Kontrolle zurückgeführt. In unserem Experiment konnten wir den Befund des false hearings nicht replizieren. Stattdessen fanden wir eine Tendenz zu geringerer Zuversicht bei falschen Antworten, die mit der Schwierigkeit der Hörbedingung übereinstimmt: Je mehr Überschneidungen zwischen Sprachlauten und Geräuschen oder je höher der Geräuschpegel, desto geringer war die Zuversicht unserer Hörer, sowohl bei jüngeren als auch bei älteren Erwachsenen. Eine mögliche Erklärung für das Fehlen des false-hearing-Effekts ist das relativ junge Alter unserer Gruppe älterer Versuchspersonen.

Wir untersuchten die Folgen des Hörens bei Hintergrundgeräuschen, d.h. die Art und Weise, wie Prozesse höherer Ordnung durch unterschiedliche Hörbedingungen (unterschiedliche Geräusche und Vorhersagbarkeit) beeinflusst werden. Dies ist wichtig, um die Auswirkungen von Hintergrundgeräuschen und Vorhersagbarkeit auf die Kommunikation zu untersuchen, die viel mehr umfasst als nur das Erkennen von Wörtern. Wir zeigen, dass der Schwierigkeitsgrad der Hörbedingung die metakognitiven Urteile beeinflusst, operationalisiert durch Sicherheitsbewertungen: Je schwieriger die Hörbedingung war, desto geringer war die Sicherheit der Hörer, was auf ein Bewusstsein für die Veränderung der Bedingung hinweist. Darüber hinaus untersuchten wir, wie Hintergrundgeräusche und Veränderungen in der Vorhersagbarkeit

das spätere Wiedererkennungsgedächtnis für experimentelle Elemente beeinflussen. Wir zeigen, dass diese Faktoren bei zuvor präsentierten Elemente die Gedächtnisleistung beeinflussen. Was wir jedoch nicht beobachten können, ist ein falsche Erinnerung für nicht präsentierte, aber semantisch verwandte Elemente.

Zusammenfassend lässt sich sagen, dass die Ergebnisse dieser Dissertation zu unserem Verständnis der Spracherkennung unter ungünstigen Hörbedingungen, insbesondere bei Hintergrundgeräuschen, beitragen und zeigen, wie prädiktive Prozesse die Sprachwahrnehmung sowohl fördern als auch behindern können.

Acknowledgements

The past years were in multiple respects difficult ones, with many unforeseen circumstances. Still, little by little, this dissertation took shape. For that I have many people to thank, who helped me in one way or another.

First, I want to thank Vera Demberg and Jutta Kray for their supervision in the past five years. You provided me with invaluable guidance and advice that made this research so much better. You helped me through all the additional challenges of lockdowns, home office, and transferring lab experiments to an online setting. I am glad to have been supervised by both of you and receive both your perspectives on the research.

I thank all the people at university whose path I crossed. This dissertation would not have been possible without your knowledge, feedback and suggestions. I learned so much from all of you. But there is more than the academic side: thank you too for the lunches, the board game nights, the bike rides, and all the other get-togethers. You filled my free time in the best of ways. I also want to thank the Phonetics group for taking me in as an honorary member of the group, inviting me to their talks and group hikes. I enjoyed and valued all the time I spent with you.

I am grateful to SFB1102 for funding my research and giving me the opportunity to grow academically and personally by being a member of such an interdisciplinary research group. I also learned a lot from acting as a representative for the PhD students for four years. Thank you for giving me that chance!

My research would not have been possible without the research assistants who helped me create the stimuli, program the experiments, record the sentences, and conduct the data analyses. Thank you for all your hard work!

I want to thank my friends for spending their time with me, be it in person or online. Thank you for all the entertainment, reassurance, and confidence you gave me, thank you for sticking with me through everything. Thank you for all the real and virtual plans we made, the movie nights, the quick visits and the holidays. Thank you for listening to my worries and my struggles, thank you for helping me with my doubts and problems. My life is much brighter with all of you in it!

Last, but not least, I am grateful to my family for their support and belief in me during the past years. Through the ups and downs (quite some downs), you were always there for me, also when this necessarily had to be from a distance. Mama, dankjewel voor alles. Papa, ik weet dat je nu heel trots op me zou zijn.

Contents

Chapter 1

Introduction

When we use speech in our every-day life, it is very rarely in quiet surroundings. In fact, the majority of the time that we listen to speech in natural circumstances, there is background noise present. We hardly notice this, and usually the levels of the noise are low enough that they do not lead to large disruptions in communication. Think of the hum of cars passing by, other people talking, machines working, or the sound of plates and cutlery when in a restaurant. We are able to communicate almost effortlessly with others despite these sounds competing with the speech signal.

How do people manage this? Is this process of listening in background noise really as effortless as it seems? And what happens when the strategies used to overcome the added cognitive load backfire? The present dissertation aims to advance the literature studying speech comprehension in background noise.

One strategy that helps to alleviate the burden that background noise places on speech comprehension is to rely on prediction. Often, it is possible to use for example world knowledge, knowledge about the speaker, or the preceding context in a dialogue to predict what might be said next. Many studies have found that a high predictability level of a sentence helps listeners in quiet to process it, but also especially in noisy listening situations, where it leads to better recognition compared to low predictability. Many theories of speech recognition incorporate this, and explain how bottom-up information from the auditory speech signal is combined with top-down information, such as predictions made based on context. This interaction of bottom-up and top-down information streams forms the basis of this dissertation.

1.1 Research Goals

We aim to address various questions in this dissertation regarding this interaction of bottom-up and top-down information, which we test in different listening conditions. Concretely, our research goals are the following:

1. **Investigating speech recognition by combining predictability, background noise, and speech sounds**
   To date, many empirical studies have investigated how varying the predictability of the target word affects speech comprehension in background noise (Boothroyd & Nittrouer, 1988; Dubno et al., 2000; Hutchinson, 1989; Kalikow et al., 1977; Pichora-Fuller et al., 1995; Sommers & Danielson, 1999; Wingfield et al., 1995; Wingfield et al., 2005). The findings of these studies concur: a predictable context facilitates speech comprehension in background noise. A different branch of literature has investigated how background noise affects the recognition of different phonemes (Alwan et al., 2011; Cooke, 2009; Gordon-Salant, 1985; Miller & Nicely, 1955; Phatak & Allen, 2007; Phatak et al., 2008; Pickett, 1957; Weber & Smits, 2003). These studies often focus on the phonemes in isolation, testing fixed contexts in nonsense syllables rather than existing words. Thus, they do not manipulate the predictability of the item, even though this affects recognition. The three factors, context predictability, background noise, and phonemes, have not been investigated together in a single study. It is possible that these factors interact: for example, the facilitatory effect of predictability might be stronger for certain phonemes, when background noise hinders their recognition.

2. **Comparing the effects of babble noise and white noise**
   Related to the previous goal, we aim to study how different types of noise affect speech recognition. While previous work has investigated the same question (Danhauer & Leppler, 1979; Gordon-Salant, 1985; Horii et al., 1971; Nittrouer et al., 2003; Taitelbaum-Swead & Fostick, 2016), the results have been inconclusive. Some studies find that white noise leads to larger amounts of interference, while other studies find that babble noise or speech-shaped noise is more difficult. These varying findings with regards to noise type suggest that other factors play a role here, for example characteristics of the tested stimuli. We aim to address this open question.

3. **Testing the predictions of the Noisy Channel Model**
   One model that has been proposed to explain human speech comprehension in background noise is the Noisy Channel Model (Levy, 2008; Levy et al., 2009; Shannon, 1949). According to this model, listeners rationally combine bottom-up information with top-down information, where the reliance on either type depends on the clarity of the listening condition. Previous studies have tested the predictions made by this model primarily in the written domain, using syntactic alternations and testing the interpretation of implausible sentences (Gibson et al., 2013; Poppels & Levy, 2016; Ryskin et al., 2018). They manipulated the level of perceived noise through the number of fillers with syntactic errors and qualified the distance between the plausible and implausible sentences in terms of the number of insertions and deletions of words. The studies found evidence supporting the predictions made by the Noisy Channel Model. One step towards more naturalistic language comprehension in noise was made by testing the same stimuli in spoken form (Gibson et al., 2016; Gibson et al., 2017). However, even though the Noisy Channel Model is constructed to explain human speech comprehension in noise, so far no studies have investigated its predictions in acoustic noise. The present dissertation aims to provide new insights. Our stimuli are different from those so far predominantly tested, constructed so that the predictability of the target word varies, as well as the overlap between speech and noise signal. Thus, our findings would be adding to the situations in which the predictions of the Noisy Channel Model might hold. We will test this in various listening situations, with different types of background noise, stimuli characteristics, and populations. Additionally, we set the Noisy Channel Model in relation to other models of speech perception.

4. **Examining the interaction of top-down and bottom-up processes in older adults**
   Differences between younger and older adults in their speech comprehension process allow us to test in detail the predictions of the Noisy Channel Model. Older adults have a different trade-off between top-down and bottom-up information compared to younger adults. On the one hand, their hearing is affected by age-related hearing loss and these declines lead to greater difficulty understanding speech in adverse listening conditions (Hnath-Chisholm et al., 2003; Gates & Mills, 2005; Gordon-Salant et al., 2010; Helfer et al., 2020; Li et al., 2004; Pichora-Fuller et al., 1995; Pichora-Fuller et al., 2017; Schneider et al., 2005; Schuknecht & Gacek, 1993; Tun et al., 2012). On the other hand, their predictive processes remain intact, and older adults have been found to rely

more on these to overcome hearing difficulties (Benichov et al., 2012; Dubno et al., 2000; Hutchinson, 1989; Pichora-Fuller et al., 1995; Rogers et al., 2012; Sheldon et al., 2008; Stine & Wingfield, 1994; Sommers & Danielson, 1999; Wingfield et al., 1995; Wingfield et al., 2005). These differences compared to younger adults make the population of older adults theoretically particularly interesting to test. We aim to study if it is indeed the case that older adults show more mishearing than younger adults in cases where the sentence context is misleading. We vary the listening conditions to construct fine-grained differences in overlap between the speech signal and the background noise, which changes the difficulty of the listening condition. Additionally, the population of older adults is often overlooked in linguistics research, which tends to rely on younger (student) populations.

5. **Investigating false hearing**
   One of the differences between younger and older adults has been found regarding the effect of *false hearing* (Failes et al., 2020; Failes & Sommers, 2022; Rogers, 2017; Sommers et al., 2015). This is the phenomenon where a listener is highly confident they are accurate in recognizing a given word, but in fact are incorrect. Often in these cases, the word has been recognized by relying on top-down processes instead of on the acoustic speech signal itself. Due to their generally stronger reliance on predictive processes, false hearing effects have been found to be larger for older adults compared to younger adults. We aim to investigate this and expect to find that the false hearing effect is stronger when the listening conditions are more difficult (due to the level of noise or increased overlap between the speech sounds in the stimuli and the background noise). The confidence ratings that participants give, will additionally shed light on their meta-cognitive processes during listening (see Research Goal 6).

6. **Testing consequences of speech comprehension in noise**
   In most studies investigating speech comprehension, participants are asked to simply report what they have heard. Afterward, their accuracy is determined. While this does answer questions regarding intelligibility of speech and difficulty of listening conditions, it leaves other points open. In every-day life, speech is used for communication, which requires more than reporting back what one was just told. Therefore, it should be investigated how different listening conditions affect subsequent higher-level processes that often play a role in communication, in order to determine how listening in noise (or other, possibly adverse, conditions) affects the conversation between interlocutors beyond mere recognition

of what is being said. In this dissertation we aim to test consequences of speech comprehension in different listening conditions, varying background noise and predictability. On the one hand, we ask participants to rate their confidence in giving the correct response after each experimental trial. This allows us not only to investigate false hearing (see Research Goal 5), but also to determine how they experienced the listening condition and how it affected their certainty. This might align with whether or not they give the correct response, but does not have to. On the other hand, in one of our experiments we present participants with a memory test after the listening phase, to test how the difficulty of the noise condition and the predictability of the target word affect subsequent memory. Knowing how memory is affected by the listening condition is important, as it can inform the management of situations where background noise is inevitable but instructions need to be understood and remembered.

7. **Investigating false memory**
   Previous work has studied the effects of predictability on memory performance, and uncovered an interesting phenomenon. Studies have found that words that are predicted but not actually presented to participants, linger in memory and affect memory performance later on, in the form of so-called false memories (Deese, 1959; Haeuser & Kray, 2022a; Hubbard et al., 2019; Roediger & McDermott, 1995; Roediger et al., 2001). In these studies, participants report remembering these predicted items, although they did not see them in the first place. This effect has not previously been studied for items embedded in background noise, but we expect to find larger amounts of false memory, as participants have been found to rely more on predictive processes in such difficult listening conditions, which could cause false memory.

1.2 Contributions of the Research

In addressing the research goals described above and answering the research questions in the three experimental chapters, this dissertation makes the following contributions:

- We extend the literature by combining three factors that interact in speech comprehension in background noise, namely the type of noise, particular speech sounds present in the stimuli, and the predictability of the context. The first two factors interact, which leads to variation in the amount of the speech signal that is obscured. While previous studies did investigate this interaction, this

was predominantly done in isolated syllables. Because in every-day life listeners are able to make use of predictions based on the semantic context to guide their recognition process, it is important to take into account the effect of a more or less predictive sentence context too. We show that the three factors interact and lead to small-grained differences in the intelligibility of the stimuli given the type of background noise and phonemes in the acoustic signal. This affects how the listener relies on either top-down or bottom-up processes during speech recognition. Reliance on predictive processes can lead to incorrectly recognised words when the sentence context is misleading. This contribution follows from Research Goals 1 and 2.

- We show that the predictions made by the Noisy Channel Model hold also when tested in a different experimental setup than what has so far been done, namely with a different set of stimuli and in acoustic background noise. We used spoken stimuli embedded in background noise of various types and levels, and manipulated the amount of perceived noise through the overlap of the speech signal and the acoustic noise by constructing stimuli with different sound contrasts (minimal pairs with plosives, fricatives and affricates, and vowels). As in these different listening conditions the clarity of the speech signal and thus the effort required for processing varies, the Noisy Channel Model makes fine-grained predictions regarding how much listeners rely on either the bottom-up acoustic signal or top-down predictive processes. We additionally tested two different populations, namely younger and older adults. Older adults have previously been found to rely more strongly on predictive processes compared to younger adults, thus allowing us to further test the predictions of the Noisy Channel Model. Again, the predictions were confirmed by our data. Taken together, our results provide additional and concurrent support for the Noisy Channel model. This contribution corresponds with Research Goal 3.

- We replicate the finding that older adults tend to rely more on the sentence context than younger adults. While this has often been reported in previous literature, we show this effect in an online study testing a younger group of participants (50-65 years old) than common in these studies (65+). We also studied the effect of false hearing. In this effect, confidence in incorrect responses is high, and it has been found to be stronger for older adults compared to younger adults. This has been attributed to age-related decays in cognitive control. In our experiment, we were unable to replicate the finding of false hearing. Instead, we find a tendency for lower confidence for incorrect responses,

in line with the difficulty of the listening condition: the more overlap between speech sounds and noise, or the higher the noise level, the lower our listeners' confidence was, both for younger and older adults. One possible explanation for the lack of false hearing effect is the relatively young age of our group of older participants. This contribution corresponds to Research Goals 4 and 5.

- We investigated the consequences of listening in background noise, or in other words, the way higher-order processes are affected by different listening conditions (varying noise and predictability). This is important to be able to investigate the effects of background noise and predictability on communication, which involves much more than simply recognizing words. We show that the difficulty of the listening condition affects meta-cognitive judgements, operationalised through confidence ratings: the more difficult the listening condition, the lower listeners' confidence was, thus showing awareness of the change in condition. Furthermore, we examined how background noise and changes in predictability affect subsequent recognition memory of experimental items. We show that for previously presented items, these factors affect memory performance. However, contrary to our predictions we do not see effects of false memory for not presented but semantically related items. This contribution follows from Research Goals 6 and 7.

1.3 Overview of the Dissertation

To address the research goals that are central to this dissertation, we conducted three experiments. Each will be described in a separate chapter. This dissertation is structured as follows.

Chapter 2 provides a literature review regarding all relevant topics for this dissertation. First, it gives a general introduction of the process of speech comprehension, and then explains how the presence of background noise and predictability of the target word affect speech comprehension. To provide context for our research questions, we briefly explain recognition memory and how this is affected by predictability, as well as age-related changes in speech comprehension. We end with a discussion of rational models of speech comprehension in noise (in particular the Noisy Channel Model), which form the basis of many of the experimental hypotheses. The state of the field and open questions are summarised.

Chapter 3 presents an experiment that addresses Research Goals 1, 2, and 3. In a word recognition task, we manipulated predictability (high or low), type of noise (babble or white noise) and the acoustic signal (different speech sounds: plosives, vowels, and fricatives), thus investigating the interaction of the three factors. This design also allowed us to compare the effects of babble noise on one hand and white noise on the other, addressing the inconclusive results in the literature so far. The experiment tested the predictions of the Noisy Channel Model for spoken language comprehension. This model has been proposed to account for language comprehension in noisy environments, suggesting that comprehenders do not solely rely on the bottom-up input, but rationally combine the input signal with top-down predictions. So far, this hypothesis has been tested mainly in the written domain, in settings where the top-down predictions are in conflict with the bottom-up signal. Our listening conditions differed in terms of their edit distance in a way that is grounded in prior work on the confusability of speech sounds in noise, leading to small-grained differences. Our findings are in line with the Noisy Channel Model's predictions: listeners probabilistically combine top-down predictions based on context with noisy bottom-up information from the acoustic signal, leading to a trade-off between the different types of information.

Chapter 4 presents an experiment that is aimed at Research Goals 3, 4, 5, and part of 6. It tested speech comprehension in babble noise in both younger and older adults. These two populations differ from each other regarding speech recognition and how information from the bottom-up auditory signal and top-down predictions are combined. Thus, we can further test predictions of the Noisy Channel Model using these two populations. Additionally, previous studies found *false hearing* effects, especially for older adults, in which they were highly confident of making a correct response during word recognition, but were in fact incorrect. We studied this phenomenon and addressed how our participants' meta-cognitive decisions reflect their reliance on either the acoustic signal or semantic context. Again, we varied the noise, but now focused on the noise level (0 dB SNR or -5 dB SNR) rather than noise type (using only babble noise). Our results show that the amount of mishearing is directly related to the perceptibility of the speech signal, and that participants' confidence ratings are in line with this too: the more overlap between speech signal and noise, the lower the confidence ratings were. We do not find evidence for false hearing effects.

Chapter 5 presents an experiment that addressed Research Goals 6 and 7, and investigated the consequences of listening to more or less predictable speech in background noise, beyond asking participants simply what they heard. We tested

how participants' subsequent recognition memory was affected by these various listening conditions. Previous studies found effects of *false memory*, in which items that are (highly) predicted but not in fact presented, linger in memory so that they are reported by participants in memory tests. We expected that this effect is stronger in background noise, as listeners rely more strongly on predictive processes in these conditions. In our experiment, participants first listened to sentences either embedded in babble noise or presented in quiet. The predictability of the target word differed by changing the word order of the sentence, and we varied the frequency of the target words as well. In a surprise recognition memory test, we asked participants whether they had seen the presented word before, and these words occurred in three types: *old items*, that were heard before, *new items*, that were not heard and unrelated to the sentences, and *semantic lures*, which were not heard, but semantically related to the old items. Findings show that memory performance for the semantic lures is not affected by any sentence-level factors, such as background noise or predictability. This is the case for old items, and these results taken together show that we do not find evidence for false memory effects for the lure items, but that, similar to the previous two experiments, accuracy for the old items depends on the difficulty of the listening condition.

Finally, **Chapter 6** summarises the findings of all three studies and discusses the outcomes. Additionally, it covers implications, open questions, and directions for future research.

1.4 Relevant Publications and Presentations

This dissertation has been partially based on the following peer-reviewed publications:

- Van Os, M., Kray, J., & Demberg, V. (2021). Mishearing as a side effect of rational language comprehension in noise. *Frontiers in Psychology, 12*:679278, 1-17.
- Van Os, M., Kray, J., & Demberg, V. (2021). Recognition of Minimal Pairs in (un) predictive Sentence Contexts in two Types of Noise. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*(43), 2943-2949.
- Van Os, M., Kray, J., & Demberg, V. (2022). Rational speech comprehension: Interaction between predictability, acoustic signal, and noise. *Frontiers in Psychology*, *13*:914239, 1-18.
- Van Os, M., Kray, J., Mecklinger, A., & Demberg, V. Effects of Noise and Predictability on (false) Memory (in prep.).

Additionally, the research has been presented at various conferences and workshops, in short (online) talks or poster presentations:

- Van Os, M., Kray, J., Demberg, V. (2020). Effect of noise on recognition of minimal pairs in (un)predictive sentence contexts. Poster presentation at the 26th Architectures and Mechanisms for Language Processing Conference (AMLaP)
- Van Os, M., Kray, J., Demberg, V. (2021). Recognition of Minimal Pairs in (un)predictive Sentence Context in Two Types of Noise. Short oral presentation at the 34th CUNY Conference on Human Sentence Processing (HSP).
- Van Os, M., Kray, J., Demberg, V. (2021). Age Differences in Mishearings During Language Comprehension under Noise. Poster presentation at the 14th Conference of the European Cognitive Aging Society (EUCAS).
- Van Os, M., Kray, J., Demberg, V. (2021). Effect of noise on recognition of minimal pairs in (un)predictive sentence contexts. Poster presentation at the Psycholinguistics in Flanders 2020/2021 conference.
- Van Os, M., Kray, J., Demberg, V. (2021). Recognition of Minimal Pairs in (un)predictive Sentence Contexts in Noise. Oral presentation at the 3rd SFB Networking Workshop.
- Van Os, M., Kray, J., Demberg, V. (2022). Rational Speech Comprehension: Interaction between Predictability, Acoustic Signal, and Noise. Poster presentation at the 28th Architectures and Mechanisms for Language Processing Conference (AMLaP).

Chapter 2

Theoretical Background

This chapter introduces the theoretical background that is relevant to the experimental studies in this thesis and which will be discussed in later chapters. The aim is to give an overview of the relevant notions and state of the literature, before building on top of that in the three experiments in this dissertation. The current chapter will first introduce how speech comprehension functions in general (Section 2.1). It will then cover the effects that background noise has on these comprehension processes (Section 2.2), as well as the role of prediction (Section 2.3). We aim to investigate also consequences of background noise and predictability, and examine the effects on recognition memory. The related literature will be covered in Section 2.4. Due to changes that occur with aging, testing older adults allows us to further examine how background noise and predictability interact in context, beyond the results we find for younger adults. Therefore, age-related changes in speech comprehension will be covered in Section 2.5. Models of language processing will be discussed (Section 2.6), with a focus on rational models, and in particular the Noisy Channel Model, whose predictions will be tested in two of the three experiments in this dissertation. Finally, a summary will be given (Section 2.7) that discusses open questions and how this dissertation aims to provide new insights in these areas. Parts of this chapter have been adapted from, or are identical to Van Os et al. (2021) and Van Os et al. (2022).

2.1 Speech Comprehension

This dissertation investigates how people understand speech in three experiments, in which we vary the listening conditions. First, this section briefly explains how speech comprehension functions in general, before turning to other factors that affect recognition of speech in the subsequent sections, where we are moving away from ideal listening situations.

Speech comprehension is a process that happens very rapidly, and yet involves multiple subprocesses to go from acoustic information to the speaker's intended meaning. In order to understand what is being said, a listener has to find the optimal mapping of the acoustic signal onto words that are stored in the mental lexicon. In the past decades, there have been many different models, both computational and theoretical, that tried to explain the process of speech recognition (see also Section 2.6). Often these models focus on the question how single, isolated words are recognized, without yet turning the attention to larger units of language, such as sentences. While the models differ in the details, most do agree that as the acoustic information is processed, it is mapped onto the representations of the words in the mental lexicon. This happens in a process of *multiple activation*, where all the words that overlap even partly with the input, are being activated at the same time (e.g., Allopenna et al., 1998; Gow & Gordon, 1995; Luce & Pisoni, 1998; Slowiaczek et al., 1987; Zwitserlood, 1989). Because a given language consists of a limited set of phonemes that form all the words in that language (Maddieson, 1984), words can be highly similar, sometimes differing only in a single phoneme, or with shorter words embedded in longer words. The activation of the word is graded, meaning that the activation level corresponds to how well it fits with the perceived input (McQueen, 2005). There is competition between all words that are activated, in which candidate words that do not match the unfolding acoustic input are inhibited. Lexical competition makes recognition more efficient by exaggerating the differences in the activation level of candidate words that differ only in very fine-grained acoustic details. At the end of the lexical competition process, all but the most optimal word candidate should be inhibited, leaving the word to be recognized (Marslen-Wilson, 1993).

When we have selected a candidate word, the process of speech recognition is not complete, however. Every-day speech occurs in sentences and paragraphs, not in isolated words. The word and all its semantic and syntactic information thus need to be integrated in the ongoing sentence and larger discourse. The language input gets mapped onto a discourse model that contains all communicative acts and

contextual information (Clark, 1996; Hagoort & Van Berkum, 2007), and this model is then used to determine the speaker's message. With all its sub components, it is astonishing that we humans can understand (and produce) speech so quickly and seemingly effortlessly. This is the case even in circumstances that add difficulty to the task, like understanding speech in the presence of background noise. This, listening in background noise, is what the current dissertation aims to investigate. The next section will explain how noise affects the process of speech comprehension as just described.

2.2 Background Noise

While psycholinguistic experiments often focus on clean speech in optimal laboratory settings, in every-day listening situations the presence of background noise is extremely common. You might be talking to someone in a busy room, with many other speakers around you. In this case, understanding speech is more difficult, the so-called cocktail party effect (Arons, 1992; Bronkhorst, 2000; Cherry, 1953). There also might be the noise of vehicles, machines, or appliances that interfere with the speech signal. All this affects the speech recognition process in a negative way, as the noise masks the speech signal, making it harder to understand.

In this dissertation, we aim to investigate the effect of background noise on speech comprehension, and in particular how the noise interacts with speech sound in predictive and unpredictive contexts. In Experiment 1 (see Chapter 3) we are interested in the masking properties of different types of background noise, and contrast white noise and multi-speaker babble noise. Therefore, we will also explain how these types of noise affect speech comprehension, in particular how they interact with different speech sounds. This latter interaction is the point of interest studied in Experiments 1 and 2 (Chapters 3 and 4). Background noise is also used in Experiment 3 (Chapter 5) to investigate its effect on higher-level processes, beyond word recognition.

2.2.1 Types of Masking

Some types of noisy signal affect the speech directly. Examples are reverberation or vocoded speech. In the case of reverberation, the speech signal occurs in an enclosed environment and gets reflected, thus overlapping with itself. In this way, reverberation blurs the temporal and spectral cues in the acoustic signal, and flattens formant

transitions (Nábělek, 1988), lowering speech intelligibility. In vocoded speech, the acoustic signal gets divided into different frequency bands, which are then filtered. The amplitude envelopes are extracted and used to modulate noise (Loizou, 1999; Shannon et al., 1995). Once the bands are recombined, the result is a distorted speech sound in which the supra-segmental features (variations in pitch, loudness, and timing; prosody) are preserved, and in which intelligibility depends on the number of bands. The higher the number of bands, the easier it is to understand what was being said.

In other cases, the noisy signal occurs through *additive noise*, where the speech signal is not directly distorted, but rather where there is an extra source of sound in the environment. This is the type of noise this dissertation will focus on. There are different types of additive noise, with different types of effects. These effects can be distinguished by the type of masking they cause, energetic masking or informational masking (Shinn-Cunningham, 2008). In the case of energetic masking, the background noise interacts directly with the speech signal outside the listener (Pollack, 1975), which leads to portions of the target speech being imperceptible. The masking signal (i.e., the background noise) overlaps in time and frequency with the target (Brungart, 2001; Culling & Stone, 2017). Due to this overlap, listeners can no longer effectively identify the cues that are needed to recognize sounds, which negatively affects speech recognition. However, listeners are able to fill in missing auditory information based on so-called *glimpses*, short regions in which the speech signal is least affected by the background noise (Ciocca & Bregman, 1987; Cooke, 2006). Therefore, energetic masking is often not the main factor leading to difficulties in speech comprehension (Shinn-Cunningham, 2008).

Informational masking includes all forms of masking that are not energetic (Cooke et al., 2008; Durlach et al., 2003; Kidd et al., 2008; Mattys et al., 2009), and thus accounts for most of the difficulties caused by background noise. It occurs when both the target speech and noise are audible, but the listener is unable to disentangle which elements belong to the target. Unlike energetic masking, informational masking is due to interference with speech perception inside the listener (Lidestam et al., 2014; Pollack, 1975). It often refers to higher-level effects of masking (Mattys et al., 2009), such as divided attention, higher cognitive load, and interference from competing speech. The presence of the masking noise takes attention away from the target speech, as there needs to be selective attention to ignore the masker. This is related to the assumption that processing resources are limited (Kahneman, 1973), and especially when the masking noise needs to be attended to for some dual task,

there will be increased informational masking of the target. If the masking noise contains speech as well, this can lead to additional informational masking, more so in the case of few speakers ($N \leq 3$; Carhart et al., 1975; Freyman et al., 2004) than of many ($N \geq 6$). Acoustic cues from competing speech can "attach" themselves to the target speech (Cooke, 2009), leading to interference. The more similar the masking speech is to that of the target, the larger the informational masking effect, for example when the speakers are of the same sex (Durlach et al., 2003). There is additional lexical-semantic interference in cases when the masking speech is intelligible to the listener (Van Engen & Bradlow, 2007), leading to informational masking even in babble with six speakers. In this dissertation, we will be making use of either white noise or a babble noise with a sufficiently large number of speakers that informational masking is reduced as much as possible. The aim of this dissertation is to investigate how speech comprehension is affected by noise and predictability, rather than test masking effects of different sources. Thus, we will focus on the effect of energetic masking of an additive noise source to isolate the effects directly caused by background noise, without the confound of additional interference caused by informational masking.

2.2.2 White Noise vs Babble Noise

The two types of noise used to mask speech in the present dissertation are white noise and babble noise. These types of noise were chosen because they differ in distinctive ways, one of which is the spectral amplitude (Gordon-Salant, 1985). Babble noise continuously varies in amplitude, while white noise is stationary (Weber & Smits, 2003). Multi-speaker babble noise approximates the average long-term spectrum of the speech of a single speaker (Garcia Lecumberri et al., 2010; Simpson & Cooke, 2005), whereas white noise has a flat spectral density with the same amplitude throughout the audible frequency range (20 – 20,000 Hertz). Previous studies have used both babble noise and white noise to test speech intelligibility in background noise. Gordon-Salant (1985) used multi-speaker babble noise, testing 57 consonant-vowel sequences, and comparing results to previous similar studies using white noise maskers (e.g., Soli & Arabie, 1979). The results showed that the interference effects of both noise types differ from each other, particularly in louder levels of noise. Taitelbaum-Swead and Fostick (2016) found lower accuracy for white noise than babble noise and speech noise when testing the intelligibility of both meaningful and nonsense words. Nittrouer et al. (2003) found a recognition advantage in speech-shaped noise compared to white noise with phonetically balanced monosyllabic words for adults, while children show impaired recognition. Other studies found opposite ef-

fects of noise type, with worse performance in speech-shaped noise maskers compared to white noise maskers (Horii et al., 1971), which Carhart et al. (1975) attributed to difficulty to separate target speech from a speech competitor. Danhauer and Leppler (1979) found similar masking results for white noise and cocktail party noise. These varying findings with regards to noise type effects suggest that the type of task and exact stimuli used, as well as the tested population, affect these effects. The present dissertation will help shed light on how the stimuli might matter and affect the results that studies report, by testing both types of noise and moving away from syllables and words, but testing sentences instead.

As briefly touched upon above, there are different types of background noise that have different effects on speech comprehension, depending on how they mask the target signal. Besides energetic masking caused by different types of noise interacting with different speech sounds, the level of the noise also matters. The level of background noise is commonly measured in Signal to Noise Ratio (SNR). It quantifies the relation between the amplitude of the speech signal and the amplitude of the background noise. A negative SNR means that the background noise is louder than the speech signal (which is thus more difficult to understand), and a positive SNR means that the speech signal is louder than the background noise. In the case of 0 dB SNR, both the noise and the speech are equally loud.

The focus of this dissertation is investigating speech recognition in noise and people's reliance on semantic context, while we vary the listening conditions. In Experiment 1 (Chapter 3) we vary the type of background noise, comparing babble noise and white noise, in part to address the question how these two types of background noise affect word recognition. This is done at a single SNR level (-5 dB SNR). In Experiment 2 (Chapter 4) we instead test only one type of noise (babble noise), but vary the SNR level, testing both 0 dB SNR and -5 dB SNR. This way, we can examine both the effects of the type of background noise and the level of the noise relative to the target speech.

2.2.3 Speech Sounds

Not only the type of noise and the noise level affect speech recognition, but also which kind of sounds occur in the speech matters (in interaction with the noise). This interaction forms the basis of our hypotheses regarding the Noisy Channel Model and will be tested in the first two experiments in this dissertation (see Chapters 3 and 4).

In order to be able to find the clearest effects of noise type on speech comprehension, we test different sets of sounds that differ from each other. These differences should cause the type of noise to have varying effects on each of the speech sounds. We chose pairs of plosives, fricatives, and vowels. First we will explain how these sounds are made and differ from each other, and subsequently we will give a review of previous studies that investigated the recognition of these speech sounds.

Plosives consist of a closure of some part of the vocal tract, followed by a short burst of energy (Ladefoged & Maddieson, 1996). In this study, the plosive pairs will differ only in the place of articulation, while voicing is kept the same across pairs. Spectral frequency information and formant transitions have been found to be particularly important for identifying the place of articulation in plosives (Alwan et al., 2011; Edwards, 1981; Liberman et al., 1954). This information can easily be lost in noise, making correct recognition difficult. We also test pairs of fricatives and affricates. These sounds are made by forcing air through a narrow channel in the speech tract, causing a turbulent airflow. They have a greater constancy of shape in different phonetic contexts compared to plosives (Ladefoged & Maddieson, 1996). The turbulent noise in fricatives is irregular and random, like in white noise, albeit less flat (Johnson, 2004). Important cues for the recognition of place of articulation in fricatives are, like for plosives, the relative spectral amplitudes, as well as other characteristics like duration (Alwan et al., 2011; You, 1979). Vowels are defined by having no major constrictions in the vocal tract (Ladefoged & Maddieson, 1996). The exact position of the tongue in terms of height and backness determines the vowel sound, also known as the first and second formants. While different features are being distinguished to describe vowels, for the present study the distinction between tense and lax is most important. The difference in a tense/lax vowel pair can generally be described in the same terms of height and backness as all other vowels, and the difference might be in the force input of the articulation (Hoole & Mooshammer, 2002). Generally, lax vowels tend to be more centralized, while tense vowels are more peripheral in the vowel space.

Some early studies (Miller & Nicely, 1955; Pickett, 1957) investigated the perception of English consonants and vowels, respectively, and how there might be confusions between the phonemes in noise. Both these studies presented a set of consonants and a set of vowels in fixed contexts (*Ca* for the consonants, *bVb* for the vowels). They were embedded in different types of noise with varying spectra, thus masking different frequency regions of speech, and at different SNRs. Participants were asked to write down what syllable they had heard, and their responses were combined and ordered

into a confusion matrix. Such a table shows how often one phoneme gets mistaken for another. The results showed that the confusions depended on what feature was masked by noise. For example, in high-frequency maskers, there is more confusion of the second formant in the recognition of vowels, while the unmasked formant is perceived correctly (Pickett, 1957). Miller and Nicely (1955) also found consistent patterns of confusions in noise depending on the type of background noise and features of the speech sounds.

Most studies investigated the effect of noise type on the recognition of consonants, usually in Consonant - Vowel (CV) or Vowel - Consonant (VC) syllables with fixed vowel contexts. Gordon-Salant (1985) found that in louder levels of multi-talker babble noise, fricatives are identified more accurately than plosives, while in severe levels of white noise the recognition of fricatives is reduced (Miller & Nicely, 1955; Phatak et al., 2008). Using a multi-speaker babble noise, Weber and Smits (2003) tested all possible CV and VC syllables in English. They found that vowels were recognized better than consonants in general, while plosives and fricatives led to the lowest number of correct responses. Mistakes that were made in the recognition of plosives were mostly errors regarding the place of articulation. Fricatives had a larger variety of errors, which were in manner, place, and voicing. Phatak et al. (2008) compared the effect of white noise as a masker to a previous study that used speech weighted noise (Phatak & Allen, 2007). They found that the tested consonants (English voiced and voiceless plosives, fricatives, and nasals) were recognized more poorly in white noise compared to speech weighted noise, but that the recognition of sibilant fricatives in particular is most reduced in white noise. Alwan et al. (2011) compared paired plosives and fricatives that differed only in their place of articulation. In a two-alternative forced choice task when listening in white noise, they found that fricatives were more robust than plosives. When it comes to the recognition of vowels in background noise, Parikh and Loizou (2005) have found that in relatively loud speech-shaped noise, the second formant is heavily masked, while the first formant was reliably detected in noise.

These previous studies often investigated the effect of noise and subsequent confusions in isolated stimuli. In every-day listening situations, speech occurs in larger units than meaningless syllables. Sounds occur in words, that form sentences, that are embedded in a larger discourse. This leads to redundancy in the information, giving the listener a chance to detect and correct perceptual errors. In the present dissertation, we test stimuli that contain speech sounds of interest (plosives, fricatives, and vowels) and study the effect of noise on their perception in a sentence context.

The listener can also use the information from the context to make predictions that aid (or hinder) the speech recognition process. We are interested in how the three factors interact and can inform us about the speech comprehension process. In the next section, the effect of predictability on language comprehension will be reviewed.

2.3 Predictability

In this dissertation, we aim to investigate how predictability affects speech recognition in noise: how are the two sources of information combined to understand the speech, depending on the listening condition? We manipulate the predictability of our experimental stimuli in two ways, namely through cloze values and surprisal, which will be described in more detail below. Changing the predictability level of the sentence affects the (ease of) processing on the part of the participants, thus allowing us to investigate different aspects of speech comprehension.

Predictability affects language comprehension, so that generally the more predictable a word is, the easier it is to process and integrate. This effect has been shown in different aspects of language processing. Responses on a cloze task are faster in sentences with a predictability as measured by cloze value (Nebes et al., 1986; Staub et al., 2015). Words are read faster or even skipped when they have a high predictability rather than a lower predictability (Ehrlich & Rayner, 1981; Kliegl et al., 2006; Smith & Levy, 2013). Conversely, low predictability and violations of plausibility lead to processing difficulties as shown by longer reading times (Rayner et al., 2004; Staub et al., 2007; Warren & McConnell, 2007). Highly predictable words show reduced processing effort as measured by pupil size compared to less predictable words (Häuser et al., 2018). Also neural responses are affected by predictability, with both the N400 and P600 components being modulated by context (Aurnhammer et al., 2021; DeLong et al., 2005; Kutas & Hillyard, 1984; Van Berkum et al., 2005; Van Petten & Luka, 2012). Taken together, these results suggest that a word's context can facilitate comprehension of that word if it is predictable, while in cases of a low predictability context comprehension is hindered.

In psycholinguistics, the predictability of a word in a sentence context is usually measured using a cloze procedure (Taylor, 1953). In such a task, a group of participants is asked to continue a sentence frame with the first word that comes to mind. This word often occurs in sentence-final position, but this does not have to be the case. The cloze value of the obtained responses corresponds to the proportion of participants that came up with this particular word. For example, when one

hundred participants are asked for the next word in the sentence "The children go outside to...", eighty might respond with "play", which then has a cloze probability of 0.8. Ten participants might respond with "run", which has a cloze value of 0.1. The remaining ten participants might have different endings to the sentence, which then have lower cloze probabilities (< 0.1). Words that have not been mentioned have a cloze value of zero. In this way, cloze predictability signifies the likelihood of a sentence ending in a particular word. What is regarded as high cloze versus low cloze can vary wildly and depend on the purposes of a particular study. Cloze values have been found to be a good predictor of self-paced reading time (Smith & Levy, 2011) and influence eye fixation durations during reading as well (Frisson et al., 2005). We used cloze ratings when constructing the stimuli used in two of our experiments (see Chapters 3 and 4).

Cloze relates to the notion of constraint. A sentence can be more or less limiting in the number of continuations that can easily follow. A highly constraining sentence, for example, would lead to many of the same continuations in a cloze test. The example above is a highly constraining sentence: A large majority of participants, namely 80%, responded with the same word. When a sentence has low constraint, it does not lead to the same kind of agreement. For example, there might be twenty different response words, each of which then has a cloze value of 5%. There is not one candidate that is most likely, but instead, many words that are good continuations of the sentence. Even when a certain word has a cloze value of 10%, the constraint of the sentence matters for its processing. In the case of a high-constraining sentence this word competes with a single candidate word with 90% cloze, while in the case of a low-constraining sentence there can be nine competitors also with a cloze of 10%.

Additionally, there is the notion of plausibility. This is often measured in a rating study, where participants are asked to rate the plausibility (or meaningfulness, or naturalness) of a sentence on a scale. As such, there is not an objective measure of plausibility (like there is for predictability). Still, plausibility can be of importance. It is correlated with predictability (Nieuwland et al., 2020): highly plausible words are often more predictable; but unpredictable words can be either plausible or implausible. This correlation makes it difficult to separate the effects of both phenomena, and makes it unclear which effect might be due to differences in predictability, and which are caused by plausibility differences. Many studies investigating the effect of predictability did not control plausibility, which makes direct comparisons between studies difficult. Papers investigating only effects of plausibility found that implausible target words lead to a disruption in processing compared to plausible target

words, as shown by reading behaviour (Staub et al., 2007; Warren & McConnell, 2007), which was delayed compared to anomalous target words (Rayner et al., 2004). This suggests that implausible items are harder to process and integrate, similar to unpredictable items. In two of the studies in this dissertation, we did not control for plausibility. We carefully selected the target words in our sentences to match certain criteria (see Chapters 3 and 4), which led to our unpredictable sentences being in some cases plausible, but in other cases implausible. In our experiments, this might have led to larger amounts of mishearing, as the sentence context and the implausibility of the target word both strongly favoured the distractor. In Chapter 5 we took a different approach to construct our stimuli and ensured that all items were plausible, also those in which the target word had a lower predictability. Thus, in this experiment, plausibility no longer was a confound.

Finally, a term that is related to predictability and often used particularly in the computational domain is surprisal. This is a concept from information theory that is defined as the negative logarithm of the probability of a word in a given context. As such, it is inversely related to predictability: highly predictable words have a low surprisal value. Computational models of sentence processing have made use of the concepts of surprisal and predictability to determine processing difficulty (e.g. Hale, 2001; Levy, 2008). The exact nature of this relationship is still unclear. While some papers report a linear relationship between word predictability and processing difficulty (measured in reading times, where longer reading times signal increased effort, Brothers & Kuperberg, 2021), other studies find a super-linear relationship (tested on a sentence-level, Meister et al., 2021), possibly logarithmic (Smith & Levy, 2013). These latter findings, of a super-linear relationship between surprisal and processing effort, are in line with predictions made by rational models of language comprehension that are based on information theory (Shannon, 1949). We will cover (rational) language models in more detail in Section 2.6. We used surprisal as a measure of predictability in the stimuli of our third experiment (see Chapter 5).

2.3.1 Predictability and Noise

The facilitatory effect of a predictable context is present also in background noise, leading to improved recognition in these usually difficult conditions (Boothroyd & Nittrouer, 1988; Kalikow et al., 1977). Kalikow and colleagues (1977) set out to construct a controlled test of speech intelligibility in noise where they varied the predictability of the target words' context. The sentences were either highly predictable (The watchdog gave a warning *growl*) or had a neutral carrier phrase to be low pre-

dictable (The old man discussed the *dive*). Normal hearing subjects were presented with the sentences embedded in multi-speaker babble noise at different SNRs. They performed differently on the items with high versus low predictability, with higher accuracy for the predictable items. Similar results were found by Boothroyd and Nittrouer (1988) testing CVC syllables (existing words vs nonsense words) and four-word sentences (high predictable, low predictable, and random sequences of words).

To date, most studies investigating predictability effects in noise did not carefully control the tested speech sounds, while the literature investigating the effect of background noise on the recognition of phonemes did not manipulate sentence predictability. Additionally, studies on the effect of background noise show conflicting results regarding which noise type affects speech comprehension the most. We address this gap in the literature in this dissertation, and designed our experiments to investigate the effect of predictability (measured by cloze values or surprisal) on comprehension when listening to speech in noise. By combining these two factors, we can test certain predictions of a model of speech comprehension (see details in Section 2.6.1). Finally, we will investigate the effects of predictability and background noise on subsequent memory performance. In the next section, we will cover the literature on recognition memory and how this is affected by the predictability of items.

2.4 Recognition Memory

In this dissertation, we aim to investigate not only how predictability and background noise interact in speech comprehension, we also want to investigate what consequences they have on further processing. In Chapter 5, we test the effects of surprisal on the participants' subsequent memory. Here we are interested in particular in recognition memory, which is the ability to (correctly) recognize that one has experienced something before. Recognition memory consists of two sub-components: familiarity and recollection (Yonelinas et al., 2010). Familiarity is a relatively automatic, fast process that does not rely on much attentional control (Coane et al., 2011; Mandler, 1980). It rather reflects a global, quantitative measure of memory strength (Yonelinas et al., 2010) in which no details about the previous experience can be recalled (Medina, 2008). On the other hand, recollection is a slower process that involves more control (Coane et al., 2011; Mandler, 1980). It includes being able to remember discrete details about the previous experience (qualitative information such as the source or temporal information; Coane et al., 2011; Yonelinas et al., 2010).

We will focus on the recollection part of recognition memory, rather than familiarity. In empirical research, a common method of testing recognition memory is by presenting a participant with an item and asking them if they have seen (or heard, depending on the test used) the item before. Recognition memory is then quantified in the proportions of correct hits (responding with 'old' when the item was old) and false alarms (responding with 'old' when the item was new). We will make use of a similar design in our experiment.

Previous studies have shown conflicting results for the effect of predictability on memory. Some studies have found that predictable words are remembered better than unpredictable words (Haeuser & Kray, 2021; Höltje et al., 2019; Perry & Wingfield, 1994; Riggs et al., 1993; Staresina et al., 2009), while other studies found that unpredictable words are remembered better (Corley et al., 2007; Federmeier et al., 2007; Rommers & Federmeier, 2018). These conflicting findings might be due to the plausibility of the target word: unpredictable words are remembered better when violating plausibility (Haeuser & Kray, 2022b; Kuperberg, 2021; Kuperberg et al., 2020). In our stimuli for the memory experiment, we manipulated the predictability of the item through the word order, using the (computational) measure of surprisal. In this way, we were able to keep the content of the items, and thus its plausibility, the same across different predictability conditions. This would not have been possible if we had used the same manner of constructing stimuli as in the other two experiments, where the sentence context has to change to vary the predictability of the target word, which is fixed in its position. Thus, we are able to test the effect of predictability on memory while keeping the content and plausibility of the sentences the same across conditions, avoiding these variables that have been identified as confounds in previous studies.

2.4.1 False Memory

Relying on predictive processes might lead to the effect of *false memory*, which is when people remember having seen or heard a certain word that was not presented to them. We investigate this effect in Chapter 5.

The first studies investigating false memory in a linguistic sense (rather than a psychological sense, in which false memories refer to the remembering of events that did not occur, see for example Brainerd et al. (2008) and Laney and Loftus (2013)) made use of word lists containing twelve to fifteen items (Deese, 1959; Roediger & McDermott, 1995; Roediger et al., 2001). All items on a list were semantically related,

and crucially, all were highly associated with the "topic" of the list, which was not presented. For example, for the list with the topic *sleep*, items on the study list were *rest*, *bed*, *dream*, and so on. Participants were asked to study the list, and then completed both a free recall test and a recognition test. Results show a large number of intrusions of the topic word, meaning that participants report remembering the word *sleep*, even though it was not presented. While the amount of intrusions varies with the exact list used (between 0% and 44% in the study by Deese (1959); and 40% to 55% found by Roediger & McDermott (1995)), the finding of false memories for the not presented topic word is replicated in many studies. Roediger et al. (2001) aimed to test what factors give rise to false memories. In a regression analysis, they found that the two factors that contribute most to explaining false recall are the average probability of recall of the words in a study list, and the strength of associations between the study words and the topic word.

One explanation that has been given for the phenomenon of false memories is that activation spreads through the semantic network (Anderson & Bower, 1974; Collins & Loftus, 1975). False memories occur, according to this theory, through residual activation that non-presented items received through the associated words on the lists. Other evidence for a spreading activation account comes from Sommers and Lewis (1999), who investigated false memories generated by lists of phonological neighbours, rather than semantically associated words. They relate this to the Neighborhood Activation Model (Luce & Pisoni, 1998).

More recently, these (semantic) false memories have been studied from a different perspective, namely for words that are predicted from a sentence context, but not actually presented (Hubbard et al., 2019). Participants are presented with a set of constraining sentences and subsequently tested in a recognition memory test. Results show that they are more likely to false alarm on items that were predicted but not presented, compared to new items. Similar effects were found in a self-paced reading study using longer retention intervals (Haeuser & Kray, 2022a). In this kind of paradigm, the false memory effect occurs even after a single exposure to a sentence, in contrast to the accumulated associations of all items in a word list in the early studies, showing a strong effect.

Because noisy listening conditions force listeners to rely more on contextual information, we expect to find a stronger effect of predictive processes in the memory test, as shown by larger amounts of false memory for items presented in noise compared to those presented in quiet. Additionally, we vary the predictability of the

target word to change how much information participants have to predict the target word.

2.5 Speech Comprehension and Aging

One of the research goals of this dissertation is to test the predictions of the Noisy Channel Model, and we test both younger and older adults to address this. These two populations differ from each other in several ways when it comes to the comprehension process, ranging from auditory processing in "ideal", quiet listening conditions to the phenomenon of *false hearing* that has been found for older adults in background noise. Based on the Noisy Channel Model, we can make different predictions regarding the recognition results for younger adults on one hand and older adults on the other. This section explains these differences between younger and older adults.

There are differences between younger and older adults even in quiet listening situations, although these differences tend to increase in noisy conditions. With increasing age, there are changes in auditory processing (Gordon-Salant et al., 2010; Helfer et al., 2020). In particular, changes in the inner ear and neural pathways can lead to age-related hearing loss, presbycusis, in which the highest frequencies (4 - 8 kHz) are most affected and continue to get worse in older adults (Gates & Mills, 2005). When the hearing loss progresses to frequencies of 2 - 4 kHz, this affects speech comprehension, and in particular understanding of voiceless consonants. Older adults also often have a reduced ability to differentiate between different frequencies, to discriminate spectral and temporal transitions in the speech signal, and to localize sound sources (Hnath-Chisholm et al., 2003; Helfer et al., 2020; Schuknecht & Gacek, 1993; Tun et al., 2012). These declines lead to greater difficulty understanding speech in adverse listening conditions (Li et al., 2004; Pichora-Fuller et al., 1995; Pichora-Fuller et al., 2017; Schneider et al., 2005). Additionally, there are cognitive changes with increasing age. Older adults have been found to show decreased attention, working memory, executive functions, and processing speed (Lindenberger & Ghisletta, 2009; Salthouse, 1990; Salthouse, 1996; Tucker-Drob et al., 2019; Tun et al., 2012). These abilities all play a role in speech comprehension, which will thus be negatively impacted as well (see also Cohen, 1987), primarily the higher level processes (Federmeier et al., 2003). Older adults struggle to divide attention between the different levels of analysis that are needed in the complete speech comprehension process (Cohen, 1987).

General language abilities are well-preserved in old age, and older adults are able to compensate for their reduced auditory and cognitive abilities by using knowledge-based factors such as a supportive sentence context (Nittrouer & Boothroyd, 1990; Stine & Wingfield, 1994; Wingfield et al., 1995; Wingfield et al., 2005). Studies compared groups of younger adults with groups of older adults, to determine how noisy environments and informative contexts might affect the groups differently (Benichov et al., 2012; Dubno et al., 2000; Hutchinson, 1989; Pichora-Fuller et al., 1995; Sommers & Danielson, 1999). The results showed that older adults are generally more adversely affected by background noise than younger adults, and that older adults rely more heavily on the provided sentence context than younger adults. In fact, older adults have been shown to rely on contextual prediction to such an extent that the predictions can make up for the adverse effect of noise (Wingfield et al., 2005) and other adverse listening conditions (Lash et al., 2013; Wingfield et al., 1995). However, online measures of brain activity (e.g. ERPs) show age-related changes in the brain response to constraining contexts, with delays in time and lower amplitudes (Federmeier et al., 2003; Federmeier et al., 2010; Federmeier & Kutas, 2005).

Still, these differences do not necessarily lead to a difference between younger and older adults on behavioural measures, especially when the task is not timed. Older adults might be particularly adept at using contextual information as a compensation mechanism, because every day they are exposed to challenging listening situations. They may have come to rely on using contextual cues to support speech comprehension processes, so that with age and experience, increased attention is allocated to higher-order knowledge structures (Steen-Baker et al., 2017). Koeritzer et al. (2018) investigated how background noise and ambiguous words in sentences affect recognition memory for spoken sentences. They presented the sentences in SNRs of +5 and +15 dB, thus with an increased acoustic challenge, but with intelligible speech. Results showed that recognition memory was worse for acoustically challenging sentences and sentences containing ambiguous words, and older adults performed worse than younger adults in the ambiguous sentences in noise. Koeritzer and colleagues concluded that in particular older listeners rely on domain-general cognitive processes in challenging listening conditions, even when the speech is highly intelligible. Rogers et al. (2012) concluded that older adults are more biased to respond consistently with the context than younger adults, due to general deficits in cognitive control. However, other studies have argued that older adults' reliance on context is instead due to a habit of making predictions and having more language experience (Sheldon et al., 2008; Wingfield et al., 2005). In the present dissertation, we aim to weigh in on this discussion and investigate what might cause this finding for older

adults. Additionally, due to the differences between younger adults and older adults, we can obtain a wider range of recognition performance to test the predictions of the Noisy Channel Model. Finally, we investigate the effect of false hearing, which will be explained in the following subsection.

2.5.1 False Hearing

While usually helpful, predictions made based on context might come at a cost. Older adults have been found to show higher rates of "false hearing" than younger adults (Rogers et al., 2012). Here, false hearing is defined as a "mistaken high confidence in the accuracy of perception when a spoken word has been misperceived" (p. 33). In other words, a participant is very certain that a word they actually misperceived, was correctly identified. In their study, Rogers and colleagues used a priming paradigm in which they paired semantically related words (*barn* / *hay*). In a training phase, participants were familiarized with these associations. In a following testing phase, the cue word (*barn*) was presented in clear listening conditions, and subsequently the target word was presented in noise. There were three conditions: (1) congruent, where the target word was the same as in the training phase (e.g., *hay*); (2) incongruent, where the target word was a phonological neighbor that formed a minimal pair with the word in training (e.g., *pay*); and (3) baseline, where the target word was unrelated to the training word (e.g., *fun*). Both younger and older adults indicated which words they had heard and how confident they were that they had identified the word correctly. The results of the study showed that older adults made use of the trained context more often and with more confidence than younger adults, even when the presented words were not matched in the training phase. Thus, older adults showed a larger false hearing effect than the younger adults.

Comparable results using a similar priming paradigm have been found by Rogers (2017) and Rogers and Wingfield (2015). Rogers (2017) additionally investigated whether the false hearing effect is caused by semantic priming or repetition priming, by manipulating the number of exposures to the training cue-target pairs. The results showed that an increased number of exposures did not increase the effect of false hearing, but that this effect was strongest when the cue-target pair was not presented at all during the training phase. These observations indicate that the false hearing effect is caused by semantic priming rather than repetition priming, suggesting that false hearing relies on top-down semantic associations in the context. More recent studies have investigated false hearing using a more naturalistic paradigm than the priming paradigm used in previous studies. Failes et al. (2020); Failes and Sommers

(2022) and Sommers et al. (2015) used sentences rather than word pairs, in three conditions. A neutral carrier phrase formed the baseline condition, and there were congruent (e.g., "The shepherd watched his sheep.") and incongruent ("The shepherd watched his sheath.") sentences. Here the sentence-final target items differed in the first or last phoneme, while controlling for frequency and neighborhood density. Participants listened to the sentence in quiet, and the target item embedded in babble noise. Identification accuracy as well as confidence ratings were analyzed, showing that older adults performed better than younger adults on congruent trials, but had a higher false alarm rate for the incongruent trials. Older adults were more confident of these false alarms than younger adults, showing the increased false hearing effect for older participants. Eye-tracking results by Failes and Sommers (2022) suggest that false hearing is caused by a decline in the ability to inhibit the activation of a response predicted by the context on the part of older adults.

Besides false hearing, larger effects for older adults compared to younger adults have been found for false memories (Hay & Jacoby, 1999, see also Section 2.4.1) and false seeing (Jacoby et al., 2012). These processes seem to share a common mechanism, as Failes et al. (2020) found that participants who showed more false hearing, also were more likely to have false memories, and Jacoby et al. (2012) link false seeing to false hearing. In all cases, there seem to be top-down processes that lead to the false perceptions by overriding bottom-up signals (Balcetis & Dunning, 2010; Bruner, 1957; Roediger & McDermott, 1995).

One of the aims of the present dissertation is to investigate this false hearing effect. We collected participants' confidence ratings in their reported answers and used this to see whether we can replicate previous findings (see Chapter 4). Our findings will also shed light on how the reliance on the acoustic signal versus the semantic context changes in older adults compared to younger adults, and how this is reflected in meta-cognitive decisions (as measured in the confidence ratings).

2.6 Models of language processing

As mentioned above, when listening to speech, one can combine information from different sources, for example from the speech itself and the given context. This idea forms the basis of many formalizations and models that have been proposed to explain human speech comprehension. While it is outside the scope of this dissertation to discuss all models that have been suggested in the past, we will briefly describe some of the more influential ones that incorporate some similar ideas as the model whose

predictions we are testing in part of this dissertation, the Noisy Channel Model, which is described below (Section 2.6.1). Many notions go back to decades old ideas, in one form or another.

Most models of speech comprehension capture the idea that information from several sources (the speech signal, lexical frequency, or context, for example) are combined in the recognition process, and use Bayesian principles to simulate a rational listener. For example, the Fuzzy Logical Model of Perception (FLMP; Oden & Massaro, 1978) assumes that speech recognition should be optimal by independently evaluating different sources of information. There is a trade-off in the model where context plays a larger role when the phonetic information is ambiguous. The Neighborhood Activation Model (NAM; Luce & Pisoni, 1998) is based on word frequencies as well as the concept of similarity neighborhoods: a set of words that are phonetically similar to a target word. It also takes into account stimulus word intelligibility as the input of the model is data from an experiment where listeners identified CVC words in background noise. A limitation of this model is that it can only account for isolated monosyllabic words, rather than a continuous speech stream. Shortlist B (Norris & McQueen, 2008) is a model that is capable of this. It is a Bayesian model that integrates bottom-up and top-down signals to simulate an optimal listener. Its input is the data of a gating study of CV and VC syllables. This data set provides perceptual confusions, like in the NAM, but in quiet rather than background noise, and also provides time-course information of the confusions. According to these latter two models, optimal word recognition depends on bottom-up evidence (from the acoustic signal) and prior lexical probabilities (based on frequencies). Shortlist B additionally considers contextual information to affect the priors. In particular, both contextual information and word frequency will influence recognition when the perceptual evidence is poor and decrease as the perceptual evidence improves.

Rational models of language comprehension also posit that humans make use of all the information that is available to them during language processing, and combine the information from different sources rationally to come to a final decision. This involves processing on an abstract, or computational level (Marr, 1982), and is founded on the assumption that human language processing is optimized (Anderson, 1991). It has been proposed as a solution to the performance paradox: despite the fact that language is very complex and full of ambiguity, people are able to understand it (and produce it) very effectively (Crocker, 2005). A rational system increases the likelihood of successful communication, while minimizing effort, and taking into account all limitations that might be present in the communication partners and the

environment (Frank & Jaeger, 2008). These rational models often involve complex computations of probabilities on various levels (discourse, sentences, lexical items, phonemes) that are based on Bayesian statistics. In these cases inferences are based on prior expectations that are made from previous (language) experience with noisy circumstances.

2.6.1 Noisy Channel Model

One particular model of rational language processing that this dissertation will focus on is the Noisy Channel Model (Levy, 2008; Levy et al., 2009; Shannon, 1949). This model proposes that speech comprehension in background noise is a rational process, where we make use of all available sources of information. Bottom-up information from the speech signal is supplemented with top-down predictions of what the speaker is likely to say. Combining these two sources of information is a sensible strategy to maximize comprehension. Bayesian decoding of the speaker's intended message is used to deal with the background noise. The speech comprehension process according to the Noisy Channel Model relies on prior knowledge in the form of linguistic and world knowledge (which meanings are more plausible; what is the base-rate frequency of certain grammatical constructions), which can be denoted by a probability distribution P_L. It also uses knowledge about what the most likely corruptions due to the noise might be, denoted by a probability distribution P_N. Using these two types of information and Bayes' Rule, the probability that a sentence s_i was intended by the speaker given the perceived sentence s_p is equal to:

$$P(s_i|s_p) = \frac{P_L(s_i)P_N(s_i \rightarrow s_p)}{P(s_P)} \tag{2.1}$$

To maximize comprehension, a listener should rationally combine bottom-up information from the acoustic speech signal with the top-down predictions based on context, situation, and world knowledge. The listener aims to identify the candidate word that is most probable given the context and given the perceived acoustic signal, following probability distributions as in Formula 2.1. Let's take, for example, the sentence "He buys the bar," In background noise, the listener might comprehend this as "He buys the car," where *car* might be more probable given the context of buying than *bar*, while sounding similar. The task of the listener consists of identifying a candidate word for which this probability given context and probability of signal fitting that word is maximal. This means that there is a trade-off between top-down and bottom-up information, where the probability distribution is shaped

differently depending on the clarity of the acoustic signal. A noisier signal leads to a flatter distribution: There are more words for which the perceived signal s has a relatively high probability, compared to a situation in which the signal is clearly intelligible. In cases where we therefore have a relatively flat probability distribution, the top-down probability will dominate what comes out as the most likely word in the calculation (besides words like *car*, also other words that frequently occur in a context of buying that share some overlap with the signal are probable based on the context). Under high noise, the top-down information will hence count more than the uncertain bottom-up information due to the stronger peaks in its distribution, leading to a stronger reliance on prediction. In everyday listening situations, relying on predictions in noisy conditions can be beneficial (Dubno et al., 2000; Hutchinson, 1989; Pichora-Fuller et al., 1995; Sommers & Danielson, 1999), and the adverse effect of noise can even be overcome by using the context to guide predictions (Wingfield et al., 1995; Wingfield et al., 2005). However, as we have seen above, it can also lead to mishearings and false hearing when the predictions are wrong or the context is misleading (Failes et al., 2020; Failes & Sommers, 2022; Rogers, 2017; Rogers et al., 2012; Sommers et al., 2015; Van Os et al., 2021).

2.7 Summary

In this chapter we have provided a theoretical background for this dissertation. The major findings and research gaps in the literature will be summarized here.

During speech comprehension, words are recognized by activating representations in the mental lexicon. Multiple representations get activated, and this happens in a graded fashion, depending on how well this candidate matches the auditory input. Through activation, inhibition, and lexical competition, the word that fits the input best is recognized.

Background noise negatively impacts the speech comprehension process, and causes information and/or energetic masking. Different types of noise have different masking effects, in particular in interaction with characteristics of the embedded speech. Often, the research investigating how certain speech sounds or phonemes are affected by (different types of) background noise, presents listeners with meaningless syllables in isolation. However, during everyday listening, there is context available that can help listeners overcome the adverse effect of background noise. The present dissertation tests how different types of background noise affect different speech sounds, in a predictable or unpredictable sentence context, thus extending the

existing literature by examining more naturalistic stimuli and combining three factors that previously have been investigated only in combinations of two (testing the effect of background noise type on the recognition of various speech sounds; testing the interaction of background noise and sentence predictability).

More predictable words have been found to be easier to process and integrate. Predictability can even help overcome adverse effects of background noise. Of course, this only holds if the context is actually supportive of the target word, and when it is not, predictive processes can lead to the wrong candidate. Predictability can be measured in different ways, for example through cloze testing or by using the computational notion of surprisal.

As people get older, there are changes in auditory and cognitive processing that affect speech comprehension, while general language abilities are well-preserved. Older adults have been found to rely more than younger adults on predictive processing to guide recognition. In this way they can overcome age-related changes. In many cases these predictive processes aid and improve speech recognition, but they can also lead to *false hearing*. False hearing is defined as mishearing a word, but being convinced that this (falsely) identified word is in fact correct, and has been found to a larger extent for older adults. The present dissertation investigates the differences in speech comprehension between younger and older adults, and in particular the effect of background noise and reliance on prediction based on context. We additionally aimed to replicate the false hearing effect and test participants' meta-cognitive judgement regarding their speech recognition process.

We relate our findings to the Noisy Channel Model, a rational model of speech comprehension, in which multiple sources of information are integrated. Bottom-up information from the speech signal is combined with top-down predictions of what the speaker is likely to say. Its predictions have been tested by Gibson et al. (2013) and follow-up work (see Chapter 3 for details). While their results are in line with the predictions of the Noisy Channel Model, their experiments made use of primarily written stimuli, in which the perceived noise level was manipulated through filler sentences that contained syntactic errors. In the present dissertation we test the predictions of the Noisy Channel Model, but now using more naturalistic stimuli: recordings of sentences in which the noise level is manipulated through the presence of background noise and the interaction between noise and speech sounds. We also test two different populations, namely younger and older adults, to further test the Noisy Channel Model's predictions.

In this dissertation, we present three experiments that aim to further our understanding of speech comprehension in general, and that in particular try to provide new insights to the literature as mentioned above. In the first experiment (Chapter 3), we test the predictions made by the Noisy Channel Model using auditory stimuli. In contrast to previous studies we embedded the stimuli in acoustic noise to manipulate the perceived level of noise, asking participants to type in the word they heard. Additionally, we controlled how much the speech signal and the noise signal overlapped by using different speech sounds (pairs of plosives, vowels, and fricatives) and two types of background noise (white noise and multi-speaker babble noise). This allowed us to test both specific predictions of the Noisy Channel Model regarding the amount of overlap between the speech signal and background noise, and to investigate the effect of both types of noise on the recognition of the different speech sounds, while participants can rely on the sentence context. Previous literature has either investigated the effects of predictability and noise on sentence comprehension without taking into account specific speech sounds, or studied the effect of background noise on specific phonemes, without including predictability as a factor. The interaction of predictability and noise on speech sounds is of interest in this dissertation, as previous findings in one of the two domains might not hold for the interaction. Additionally, the results may inform machine-generated speech and improve intelligibility in this domain, as alternatives with lower risk of misunderstanding can be selected (Chingacham et al., 2021).

In the second experiment (Chapter 4) we conduct a similar experiment, but change the sampled population and some stimuli characteristics. In this experiment, we test both younger and older adults, to compare changes in word recognition and reliance on sentence context between these two groups. Again, we relate these results to the predictions of the Noisy Channel Model, building on the findings from the first experiment. Additionally, we aim to study the false hearing effect, which we expect, based on the literature, to be stronger for older adults than younger adults.

Finally, in the third experiment (Chapter 5) we shift gears: We no longer simply ask participants to report what they heard in the sentence recording. Instead, we are interested in the consequences of background noise and predictability on higher-level processes. Thus, in this experiment we test how noise and surprisal affect recognition memory some time after the sentences have been heard. The results will tell us more about what the effects might be of adverse listening conditions and different predictability levels on communication at a higher level, beyond the mere recognition of the words that are said.

Chapter 3

Exp. 1: Background Noise and Speech Sound Contrasts

In this dissertation, we investigate how people recognize speech in noise, and how they make use of the sentence context to aid them. One model that has been proposed to capture the effects of listening to speech in background noise is the Noisy Channel Model (Levy, 2008; Levy et al., 2009; Shannon, 1949). This model posits that speech comprehension in noise is a rational process, where different sources of information are combined. Thus, people make use of both bottom-up auditory information and top-down predictions based on for example context. The Noisy Channel Model makes predictions of human behaviour that can be tested in experiments. Results from previous studies (Gibson et al., 2013; Gibson et al., 2016; Gibson et al., 2017; Poppels & Levy, 2016; Ryskin et al., 2018) confirmed these hypotheses for stimuli using syntactic alternations. However, these studies used filler sentences with syntactic errors as an operationalisation of noise. The predictions of the Noisy Channel Model have not been investigated using auditory stimuli embedded in acoustic background noise, as the present experiment does.

The rest of this chapter is structured as follows. In the next section, we will briefly summarize the relevant literature on speech comprehension in background noise and predictability (Section 3.1), which was covered extensively in Chapter 2. In Section 3.1.2 we define the research goals and the hypotheses of the current experiment. The method and stimuli creation are detailed in Section 3.2. The results of the experiment are reported in Section 3.3 and discussed in Section 3.4. Finally, the

chapter is summarised in Section 3.5. This chapter is based on, and in some places identical with Van Os et al. (2022).

3.1 Introduction

Listening in background noise is more difficult than listening in quiet circumstances, because the noise masks the speech signal. In this dissertation, the focus is on additive noise (rather than distortions of the signal), which leads to energetic masking (Shinn-Cunningham, 2008) The noise overlaps with the speech signal in time and frequency (Brungart, 2001; Culling & Stone, 2017), which means that the cues needed for speech recognition can no longer be identified. Different types of noise can have different types of masking effects, depending on their characteristics (Gordon-Salant, 1985; Weber & Smits, 2003). For example, while white noise is stationary, babble noise varies in amplitude over time. Previous studies have found conflicting results on how these two types of noise affect speech comprehension, with some studies finding that white noise leads to worse performance (Nittrouer et al., 2003; Taitelbaum-Swead & Fostick, 2016), while other studies found that babble noise was more detrimental (Danhauer & Leppler, 1979; Horii et al., 1971). These inconsistent findings suggest that other factors than the noise itself might influence the effects.

One of these factors is the particular kind of sounds that are being masked. Phonemes can be divided into different categories, which are in different ways affected by noise. For example, white noise shares a similar signal with fricatives, and thus affects these more negatively than other speech sounds (Miller & Nicely, 1955; Phatak et al., 2008). Vowels are more affected by babble noise (although this effect does depend on the exact vowel and its formant values; Parikh & Loizou, 2005), and the short signal of plosives is lost easily in noise in general. In our experiment, we use these differences in overlap between the speech signal and noise as a way to test the predictions of the Noisy Channel Model.

Predictability affects language comprehension, and generally has a positive effect: when an utterance is more predictable, it is easier to process and integrate (see for example DeLong et al., 2005; Ehrlich & Rayner, 1981; Kliegl et al., 2006; Kutas & Hillyard, 1984; Rayner et al., 2004; Smith & Levy, 2013; Staub et al., 2007, 2015; Van Berkum et al., 2005; Van Petten & Luka, 2012; Warren & McConnell, 2007). In background noise, this facilitatory effect also aids comprehension, helping people overcome the adverse effects of background noise (Boothroyd & Nittrouer, 1988; Kalikow et al., 1977).

3.1.1 Noisy Channel Model and Syntactic Alternations

The Noisy Channel Model makes certain predictions that can be tested in experimental settings. The Noisy Channel Model is a rational model of human speech comprehension in noise. It states that listeners rationally combine all information available to them, so that, for example, they use both the acoustic speech signal and contextual information from the sentence. These two sources of information are combined based on the availability of the information: in background noise the speech signal is less reliable and thus the word recognition process relies more on predictive processes instead. The Noisy Channel Model makes additional predictions following the same logic, which can be tested in experiments. One of the studies that most systematically tested the predictions made by the Noisy Channel Model is that by Gibson et al. (2013). We will describe this study here in some detail to provide a baseline of the work that has been done to verify the Noisy Channel Model's predictions, which we will build on further in our experiments.

In their study, Gibson et al. (2013) investigated the interpretation of syntactic alternations in English. In particular, they used different English constructions (active / passive, subject / object locative, transitive / intransitive, double object / prepositional phrase object) in a plausible and implausible version. The difference between these two versions was expressed in the number of insertions and deletions of words. The exact four predictions they tested were: (1) Alternatives that are more similar (in terms of edit distance) are more often interpreted as the plausible version than alternatives that have a larger distance between them; (2) Not all changes should be treated equally: Plausibility interpretations are stronger for items with deletion than insertion; (3) Increasing the perceived noise level should lead to more plausible interpretations; and (4) If the base rate of implausible sentences is higher, listeners stick to the literal meaning.

Gibson and colleagues (2013) then presented these items in written form to participants with comprehension questions that were used to determine the participants' interpretation of the experimental items. The researchers checked whether the participants had interpreted the implausible sentences in their literal, implausible meaning, or whether participants had "edited" the sentence to fit a more plausible version instead. The results showed that participants preferred a plausible interpretation over the literal one, and this was dependent on the number of edits (insertions and deletions). The larger the number of edits, the more likely it was that the sentence was interpreted literally. When the perceived noise rate increased, participants more often

interpreted the sentence in its plausible meaning rather than literally. Thus, all four predictions were confirmed by the results of the experiment.

Taken together, the results of this study provide strong evidence in favour of the Noisy Channel Model in sentence comprehension. These conclusions have been replicated using the same method by Poppels and Levy (2016), who also found that comprehenders consider positional exchange of function words when interpreting implausible sentences, as well as using a different method where participants were asked to retype the experimental item and edit if they thought there were mistakes (Ryskin et al., 2018). However, these studies made use of written materials that participants could read multiple times if they wanted, and as such is different from speech comprehension. Here the short-lived speech signal is presented only once and can be processed only incrementally. Using the same materials but now presented in auditory form, Gibson et al. (2016) and Gibson et al. (2017) found similar results providing evidence for the Noisy Channel Model. However, even when the items were presented in spoken form, the perceived noise levels were, like in previous studies, manipulated through the number of filler sentences that contained syntactic errors, rather than through actual acoustic noise. The present dissertation aims to extend these results by using auditory stimuli of a different kind than in the previous studies, which mainly investigated the interpretation of syntactic alternations. By using different types of speech sound contrasts and background noise, we manipulate the distance between the literal interpretation of the experimental item and the meaning inferred based on sentence context. In this way, we will test the predictions of the Noisy Channel Model in a more naturalistic setting. If our results are in line with the predictions made by the Noisy Channel Model, this would provide additional support for the theory.

3.1.2 Research Goals and Hypotheses

The aim of the present study was to test predictions made by the Noisy Channel Model (Levy, 2008; Levy et al., 2009; Shannon, 1949) in the auditory channel, varying the amount of background noise through the interaction of noise type and speech sound. We examined mishearings that occur when listening in background noise, depending on the predictability of the context and certain sound characteristics of the target word. We carefully controlled the target word so that it formed a minimal pair that differed in a medial sound contrast, controlling the specific pairs of these sound contrasts. In this experiment, we presented participants with a written sentence context on the screen that could be used to guide prediction while listening. In the

high predictability condition, these predictions would lead to the correct response, but in the low predictability condition, relying on the context would give an incorrect response.

We expected to find a main effect of noise: overall, there will be a lower number of correct responses in noise compared to quiet. This result acts as a control condition: finding fewer correct responses in quiet than noise would point to a problem in the experimental design. We additionally investigated whether there is a difference between the two noise types we use, babble noise and white noise. Overall, studies have found conflicting results when comparing white noise and babble noise (e.g., Gordon-Salant, 1985; Horii et al., 1971; Taitelbaum-Swead & Fostick, 2016). Because of these conflicting results, we hypothesize that this effect of noise type most likely depends on other factors, such as the exact task, population, and most importantly the characteristics of the stimuli, such as the presence of predictive context and the occurring phonemes (details of this interaction of noise type and speech sound contrast will be discussed below).

We expected an interaction of noise and predictability based on the semantic context. The Noisy Channel Model predicts that participants will rely more on the sentence context when listening in noise, rather than the acoustic signal, as they use the information from the context to compensate for the processing difficulties of the speech signal in background noise (Failes et al., 2020; Gibson et al., 2013; Wingfield et al., 1995; Wingfield et al., 2005). In the low predictability sentences, this will lead to incorrect responses, as the context is misleading by predicting a different word than the target. As such, we expected that the low predictability condition leads to more mishearing than the high predictability condition for exactly this reason. In contrast, in the high predictability condition, the target word is supported by both the audio signal and the context, leading to high accuracy rates independent of the noise condition responses, as predicted by the Noisy Channel Model.

According to the Noisy Channel Model, if there is more noise that obscures the speech signal, listeners should rely more on other sources of information such as sentence context. This prediction has been confirmed on stimuli testing syntactic alternations (Gibson et al., 2013; Gibson et al., 2016; Gibson et al., 2017; Poppels & Levy, 2016), but has not been tested using auditory stimuli of a different type than syntactic alterations. In the current study, we manipulate the amount of noise on the signal by using different types of background noise and speech sound contrasts. The interaction of noise type (babble and white noise) and sound contrast (plosives, vowels, and fricatives) should lead to different amounts of signal masking, defined

by the amount of interference the background noise has on the speech sound. With their burst, plosives have the shortest and least clear signal out of our three tested sound contrasts. Therefore, we expected that plosives will show more mishearing than fricatives and vowels, overall, because the perceived noise is greater. We do not predict any differences in the degree of mishearing in plosives depending on the type of noise, as the signal of the plosive is easily lost in general, but does not overlap in particular with a specific type of noise tested here. We do expect this interaction with noise type for fricatives and vowels. Fricative sounds have their energy at the same frequencies as white noise, and therefore fricatives should be harder to identify correctly in white noise than babble noise. There would be more noise in the form of (energetic) masking in the case of white noise, lowering performance for fricatives (Miller & Nicely, 1955; Phatak et al., 2008). Results for fricatives in white noise might show a low performance with a large amount of mishearing, possibly on the level of the plosives (which are difficult in general). In babble noise, with energy mainly in different frequencies (Garcia Lecumberri et al., 2010; Simpson & Cooke, 2005), recognition of fricatives should not be majorly affected, as the perceived noise is lower. For vowels, we hypothesized that these items generally tend to be more difficult to be identified correctly in babble noise, but easier in white noise. In fluctuating babble noise, the particular formant values that determine the vowel might be lost, while in the steady signal of white noise they can be recovered (Benkí, 2003; Pickett, 1957; Weber & Smits, 2003).

In sum, we aimed to investigate how listeners integrate different types of information when listening in noise, combining different factors like background noise, characteristics of the speech, and context. We were interested in interactions between noise and predictability on one hand, and noise and sound contrast on the other, as specific predictions of the Noisy Channel Model can be tested here. In short, these predictions are that participants' interpretations will be based more on sentence context in background noise, leading to incorrect responses in the low predictability noise conditions, and that these effects should be modulated by how much the type of noise and the speech sound characteristics overlap in their acoustic signal.

3.2 Method

3.2.1 Participants

Fifty native speakers of German were recruited for the experiment via the recruitment platform Prolific (prolific.co). Data from two participants was excluded due to technical problems. The mean age of the final group of 48 participants was 23 years (age range = 18-30 years), 25 were male. All participants gave informed consent before the experiment, and the study was approved by the Deutsche Gesellschaft für Sprachwissenschaft (DgfS) ethics committee. The experiment lasted approximately thirty minutes and all participants received €4,75 as compensation for their participation.

3.2.2 Materials and Task

Our stimuli consisted of 180 minimal pairs, around which sentences were constructed. These included both high predictability sentences and low predictability sentences. Items were recorded and embedded in background noise at -5 dB SNR. We used both white noise and a multi-speaker café babble noise. Details of the stimuli construction can be found in the sections below.

Selection of Minimal Pairs

In the first step of creating our stimuli, we made an inventory of the number of minimal pairs in the German part of the CELEX lexical database (Baayen et al., 1995), based on their phonetic transcription. We collected counts for existing minimal pairs that differed at the beginning of the word, in the middle, or at the end, for pairs of consonants in German (for example /p/-/t/ or /m/-/z/), as well as for the pairs of vowels (for example /ɪ/-/i/ or /ɛ/-/a/). Additionally, we coded whether these pairs of sounds differed in one or two phonetic features. We did not look up items where the pairs of sounds differed in more than two phonetic features.

We chose only words that differed in the same position, which was in the middle of the word. Here we wanted to avoid co-articulation cues coming from the previous word, which would not be controlled. Additionally, to facilitate item construction, most minimal pairs were available with the contrast in medial position. We chose coherent categories of sounds to compare, namely the following plosive contrasts, all differing in place of articulation only (thus one phonetic feature): /p/-/t/, /p/-

Table 3.1: Counts of sound contrast pairs

Vowels: 58 pairs					
ɛ/œ: 26	ɪ/i: 13	ɐ/ə: 8	ʊ/u: 6	ɔ/o: 3	ʏ/u: 2
Plosives: 62 pairs					
p/t: 17	k/p: 16	k/t: 9	b/g: 9	b/d: 7	d/g: 4
Fricatives: 60 pairs					
s/∅: 13	x/∅: 5	ts/∅: 5	f/∅: 5	ʃ/∅: 5	f/s: 5
s/ts: 5	s/x: 4	f/h: 3	f/ʃ: 3	f/x: 3	h/ts: 1
pf/ʃ: 1	pf/ts: 1	ʃ/s: 1	ʃ/x: 1	ʃ/ts: 1	

/k/, /t/-/k/, /b/-/d/, /b/-/g/, /d/-/g/ (62 pairs); the following vowel contrasts, having a tense/lax contrast: /i/-/ɪ/, /y/-/ʏ/, /u/-/ʊ/, /ɛ/-/œ/, /o/-/ɔ/, /ɐ/-/ə/ (58 pairs); and the following voiceless fricative and affricate pairs: /f/-/h/, /f/-/ʃ/, /f/-/s/, /f/-/x/, /h/-/ts/, /pf/-/ʃ/, /pf/-/ts/, /ʃ/-/s/, /ʃ/-/x/, /ʃ/-/ts/, /s/-/x/, /s/-/ts/ (29 pairs). To increase the number of fricative and affricate pairs, we also included minimal pairs that consisted of a fricative or affricate and a deletion of the sound: /f/-/∅/, /ʃ/-/∅/, /s/-/∅/, /x/-/∅/, /ts/-/∅/ (31 pairs). Once the sound pairs had been selected, we turned our attention to the word pairs that contained these sounds. Due to noise in the CELEX database, not all of the extracted pairs were true minimal pairs (standard pronunciation differs from the CELEX transcription, for example *Kasse* kɑsə/*Case* keːs). We excluded these. Furthermore, we made sure that all minimal pairs that were selected, matched in part of speech as well as gender in case of nouns. This was done with later sentence construction in mind. We also excluded word pairs with one or two too infrequent words (technical terms or words used only in certain regions of Germany) or pairs based on the same two verbs in different tenses. This left us with a total of 264 (82 for plosives; 75 for vowels; 107 for fricatives and affricates) word pairs to use as a basis for our stimuli. After pretesting our stimuli, we had a final set of 180 word pairs, 62 of which had a plosive contrast, 58 had a vowel contrast, and 60 had a fricative or affricate contrast. Table 3.1 shows the division across the specific contrasts. Controlling the previously mentioned factors meant we were not able to control for lexical frequency or neighborhood effects.

Sentence Construction

With the help of native German-speaking research assistants, sentences were constructed around the minimal pairs, so that the target word appeared in sentence-final

position and the word would be predictable from the sentence context. We tried to keep the sentences of a minimal pair of similar length, but had no strict criteria. We ensured that the target words always occurred at the end of the sentence. All stimuli were subjected to cloze testing by asking native German speakers on the Prolific platform to complete the items with a single word. Cloze probabilities for each item were calculated based on the answers of ten participants. We aimed for high cloze probabilities. Therefore, all stimuli that were still scoring below 0.5 on cloze probability were revised. In these revised versions, we tried to guide participants' predictions to the target word we had in mind, and changed the items based on their previous responses. We changed or narrowed down the context, or included the alternative candidates in the sentence to get more participants to converge on the target word. Three rounds of cloze testing were completed, until we had 360 high predictability sentences. An example of the high predictability sentences of a pair can be found in Table 3.2, sentences 1A and 1B. In this final set, we had 242 items with cloze values of 0.5 or higher. These had a mean cloze of 0.75 (SD = 0.17). For the remaining 118 items, cloze values were under 0.5. The cloze values ranged from 0.5 to 1 (mean = 0.72) for the 136 items constructed under strict conditions. In 104 cases, the cloze was still quite low. We relaxed the high cloze requirement when even after multiple revisions, there was a high cloze competitor that differed only in the prefix (*laden* vs *aufladen* for 'to charge') or that was very semantically similar and too highly frequent to allow us to improve the sentence (*sieden* vs more frequent *kochen* for 'to boil'). We included these items even though they had a lower cloze probability than 0.5. The average cloze for all items, including those with the relaxed requirements was 0.59 (SD = 0.27). None of the participants took part in more than one of the rounds of cloze testing, and none of them participated in the main experiment.

To make the low predictability stimuli, we swapped the two sentence-final target words, aiming for unpredictable but grammatically correct swaps wherever possible. Examples of low predictability sentences can be found in Table 3.2, sentences 1C and 1D. In practice, this meant that most swapped sentences were both unpredictable and implausible. Almost all sentences were still grammatically correct after swapping the target word, but two out of 240 swapped sentences became grammatically incorrect (for example, an argument was missing for a transitive verb). The addition of the low predictability stimuli resulted in 120 sets of four sentences, with two predictable and two unpredictable sentences of the minimal pair (total N = 480). Plausibility ratings were collected for all 480 items, again using the Prolific environment. Each item was rated ten times, and ratings were averaged. Again, none of the participants took part in the main experiment. Plausibility was rated on a scale from 1 (completely implau-

Table 3.2: Example Stimuli (Experiment 1)

1A	Am Pool im Hotel gab es nur noch eine freie **Liege**. *At the pool in the hotel there was only one free **lounger** left.*	HP
1B	Nach vier Jahren heiratete Paul seine große **Liebe**. *After four years, Paul married his big **love**.*	HP
1C	Am Pool im Hotel gab es nur noch eine freie **Liebe**. *At the pool in the hotel there was only one free **love** left.*	LP
1D	Nach vier Jahren heiratete Paul seine große **Liege**. *After four years, Paul married his big **lounger**.*	LP

Note. Highly predictable sentences (HP) were made based on minimal pairs (Liebe / Liege, target words in **bold**) in 1A and 1B), then sentence-final target words were swapped to make low predictability items (LP) with the sentence frames of 1A and 1B, resulting in 1C and 1D. English translations have been given in *italics*.

sible) to 5 (completely plausible). The predictable sentences had a mean plausibility rating of 4.60 (SD = 0.41), and the unpredictable sentences had a mean plausibility rating of 1.73 (SD = 0.59). A comprehensive list of all items as well as their cloze and plausibility ratings can be found in Appendix A.

Recordings and Preprocessing

Recordings were made of all predictable sentences (240 in total). The sentences were read by a female speaker, who was a native speaker of German. The speaker was standing in a sound-treated booth. She was instructed to read slowly, and to pay attention to not include any slips of the tongue or hesitations. Sentences that were not read as intended or included slips of the tongue were repeated until each sentence was recorded in a clean version suitable for testing.

All items were segmented from the continuous recording with no leading or trailing silence. Unpredictable sentences were constructed via cross-splicing of the recordings of predictable sentences. This was done in order to make sure that the intonation and stress patterns were identical across conditions and not indicative of the unpredictable items. The splicing was performed using Praat (Boersma & Weenink, 2009, Version 6.1.05), and resulted in a total of 480 sentences. All cross-spliced unpredictable items were listened to carefully, to identify any problems related to cross-splicing, and corrected by adapting the slicing boundary or adapting the pitch contours. This was done by the first author as well as two native German

student assistants. The final 480 sentences all sounded natural for the purposes of the experiment.

Adding Noise

All items were normalized to 65 dB and 300ms of leading and trailing silence were added. Then, all sentences were embedded in background noise. We used two types, a multi-speaker babble noise and a white noise. The babble noise was café noise (BBC Sound Effects Library, Crowds: Interior, Dinner-Dance, http://bbcsfx.acropolis.org.uk/). The noise was added with a Signal to Noise (SNR) ratio of -5 dB, meaning that the intensity of the background noise was five dB stronger than the target sound. This led to challenging, but not impossible listening conditions. As we were interested in the effect of background noise and sentence context on the intelligibility of the sentence-final target word, we took the mean intensity of each target word and calculated the SNR values based on this value, rather than the mean intensity of the sentence. Because the intensity of a spoken sentence tends to drop towards the end (Vaissière, 1983), it would mean the SNRs were actually lower for the target word, in case the mean sentence intensity were to be used. We calculated the level of background noise separately for all items, both in the HP and LP condition. For each item, the intensity of the target word was measured in Praat (Boersma & Weenink, 2009, Version 6.1.05) and the corresponding noise level calculated and automatically mixed in using a Python script. The noise was the same level throughout the sentence. It started 300 ms before sentence-onset and continued for 300 ms after sentence-offset, so that it would not be too abrupt, and to give participants a chance to focus on the speech in the noise. For the same reason, we decided against keeping the sentence context clear and only embedding the target word in noise. We feared participants would not have time to get used to the added noise, leading to worse performance, and it would be a less natural way of presenting the stimuli. Besides the noise conditions, we also presented participants with the items without added background noise, thus adding a quiet condition.

In the experiment, the participants' task was to listen to the recording of the sentence, and then type in the sentence-final target word they had heard. For this, a free-response text box was available. Subsequently, they rated their certainty in giving the correct answer on a four-point scale.

3.2.3 Design

All experimental items were arranged in a Latin Square design to make twenty-four experimental lists of ninety sentences each. This length was chosen to be manageable in a single experimental session. On each list, all noise conditions as well as quiet were presented. This was done blocked by noise with thirty items per block, starting with either babble noise or white noise. This order was counterbalanced across participants. Half of them started with white noise, the other half with babble noise. Quiet was presented last to make the manipulations in the experiment less obvious to participants. In each block of noise, there was the same number (N = 15) of high- and low predictability items, presented in random order. Participants heard only one item of a pair, and only in one of the noise conditions. Each list started with a short practice block of four items, during which all types of noise as well as quiet were presented.

3.2.4 Procedure

The experiment was run online, and hosted on Lingoturk, a crowd-sourcing client (Pusse et al., 2016). Participants were instructed to complete the experiment on a computer (not a tablet or smartphone) in a quiet room. The experiment started with on-screen instructions of the task (see Appendix D for all (German) instructions). These instructions included a sound check so that the participant could make sure the audio was working correctly before the experiment started. They were instructed to set the audio to a comfortable level. Due to the online setting of the experiment, we were unable to control the type of audio hardware participants used. In the post-experimental questionnaire we did include questions on how loud the participants' testing surroundings were and if they were doing any secondary tasks (watching TV, texting, etc.), to get an idea of the conditions during the experiment. Participants first listened to a recording of a sentence while looking at a fixation cross. The length of the audio recordings ranged from 1932 ms to 9632 ms. After the item played, the screen automatically moved from the fixation cross to the next screen without delay. Participants then were asked to type in the final word they had heard on the next screen. Here, we presented the sentence (minus the sentence-final target word) in written form on the screen to ensure that the participant could use the contextual information even in difficult noisy conditions. Participants typed their response in a text box, and could start as soon as the screen with the sentence context and text box appeared. On the same screen, there was a question regarding their confidence in

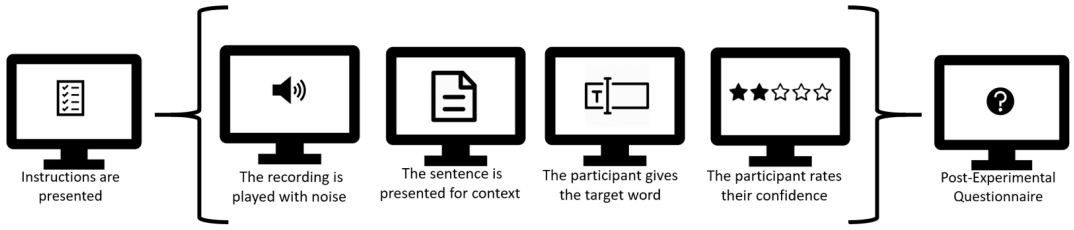

Figure 3.1: This figure shows the different stages of the experiment, with a single trial between brackets. Participants completed four practice trials and ninety experimental trials.

giving the correct response on a four-point scale. The next item's recording started playing automatically as soon as the participant had clicked on 'Next' to go to the next trial. The task was not timed, so participants could take as long as they needed to make their responses. Figure 3.1 presents a schematic overview of the experiment.

3.2.5 Analyses

All participants' responses were automatically classified on whether it was the target (the word that was played in the audio, e.g., in example 1A in Table 1 "Liege" / "lounger"), the similar sounding distractor (e.g., in 1A "Liebe" / "love"), or a different word entirely (e.g., in 1A "Platz" / "space", wrong). The list of responses that were classified as wrong was then manually checked by the first author and a native German-speaking student assistant to correct misclassifications because of typos or spelling mistakes. In our statistical analyses, we added the trial number within each noise block (1-30 for thirty trials per block) as a variable to check for learning effects in the experiment.

To get a better idea of what information participants relied on when making their wrong responses, we coded the semantic fit of the incorrect responses (fitting or not fitting). Fitting responses resulted in a grammatical and meaningful sentence, as judged by a native German student-assistant. We also coded the phonetic distance between the incorrect responses and target items. We made phonetic transcriptions based on the Deutsches Aussprachewörterbuch (German Pronunciation Dictionary; Krech et al., 2009) and calculated the weighted feature edit distance using the Python package Panphon (Mortensen et al., 2016). This distance was normalized by dividing it by the longest of the two compared words. The normalized distance fell between 0 and 1.

3.3 Results

In our statistical analyses, we used generalized linear mixed models with logit link (GLMMs), implemented in the lme4 package (Bates et al., 2014) in R (R Core Team, 2022). These models allow both fixed and random effects, letting us control for variation on the participant- and item-level (Baayen et al., 2008; Barr et al., 2013). To improve convergence, all models were run using the bobyqa optimizer and increased iterations to $2{\cdot}10^5$. Model comparisons were made to guide model selection based on the Akaike Information Criterion (AIC), models with the lowest AIC are reported below. We used forwards Helmert contrast coding for the Noise variable, so that the first contrast showed the difference between the Quiet condition and the mean of both types of noise (with weights of -1, 0.5, 0.5), and the second contrast showed the difference between Babble Noise and White Noise (using weights of 0, -0.5, 0.5). The other categorical predictor variables were treatment coded.

3.3.1 Interaction Noise and Predictability

We first ran a model that included Noise and Predictability, as well as their interaction. Noise was a categorical predictor with three levels that was contrast coded using forwards Helmert coding, as explained above. Trial Number was included as a continuous predictor that was scaled. In order to use logistic regression, we collapsed wrong and distractor responses in our models and compared them to the target responses. Running all models with targets vs distractor responses (leaving out wrong responses, which occur least), led to very similar results. The model included random intercepts for Participant and Item, with random slopes of Noise and Predictability for both Participant and Item. The model revealed a significant effect of Predictability (β = -4.19, SE = 0.50, z = -8.43, $p < .001$), showing fewer target responses in the low predictability items. There was a significant effect for the first contrast of Noise (β = -1.91, SE = 0.61, z = -3.16, $p < .01$), showing a lower amount of target responses in noise compared to quiet. The interaction of Predictability and Noise was also significant (β = -2.02, SE = 0.41, z = -4.87, $p < .001$), showing a larger negative effect of noise on correct target responses in low predictability items than high predictability items. The other effects were not significant (the other effects were not significant; ps $> .12$), as can be seen in Table 3.3.

Table 3.3: Model Outcomes for the Overall Model (Interaction Predictability & Noise)

	Estimate	*SE*	*Z*-value	*p*-value	
Intercept	5.07	0.52	9.52	$< .001$	***
Predictability (LP)	-4.19	0.50	-8.43	$< .001$	***
Noise Contrast 1	-1.91	0.61	-3.16	$<.01$	**
Noise Contrast 2	-0.03	0.35	-0.09	.93	
Trial Number	-0.10	0.07	-1.56	.12	
Predictability (LP) : Noise Contrast 1	-2.02	0.41	-4.87	$< .001$	***
Predictability (HP) : Noise Contrast 2	0.17	0.40	0.42	.67	

3.3.2 Low Predictability Subset

In the following analyses, we will analyze the subset of Low Predictability items in detail, as here we see interesting effects. In the High Predictability condition we find ceiling effects (see Figure 3.2). Reducing the size of the model also has the benefit of reducing the number of comparisons, which eases the interpretation of the model results. This means that per noise condition, a total of fifteen LP trials was analyzed for each participant. Figure 3.3 shows the responses for the low predictability items. We expect to find an interaction of Noise and Sound Contrast, which would indicate that the type of background noise has a different effect on intelligibility for the different speech sound contrasts in our stimuli.

We included three predictors in the LP model: Noise, Sound Contrast, and Trial Number. Sound Contrast was a categorical predictor with three levels, with Plosives as the base level. We additionally included the interaction of Noise and Sound Contrast. Trial Number was a continuous predictor that was scaled like before. The model included random intercepts for Participant and Item, with a random slope of Noise for Participants. Including additional random slopes resulted in non-convergence.

The model revealed a significant main effect of Noise, but only for the first contrast indicating the difference between Quiet and both noise conditions together, showing that the noise conditions led to fewer target responses than Quiet (β = -3.29, SE = 0.24, z = -13.53, $p < .001$). There was no difference in the second contrast indicating the difference between Babble and White Noise ($p = .80$). However, recall that Plosives is the base level and that thus these main effects only hold for Plosives. We additionally found interactions of Noise and Sound Contrast when comparing Plosives to Vowels (β = 0.62, SE = 0.29, z = 2.13, $p < .05$ for Noise contrast 1

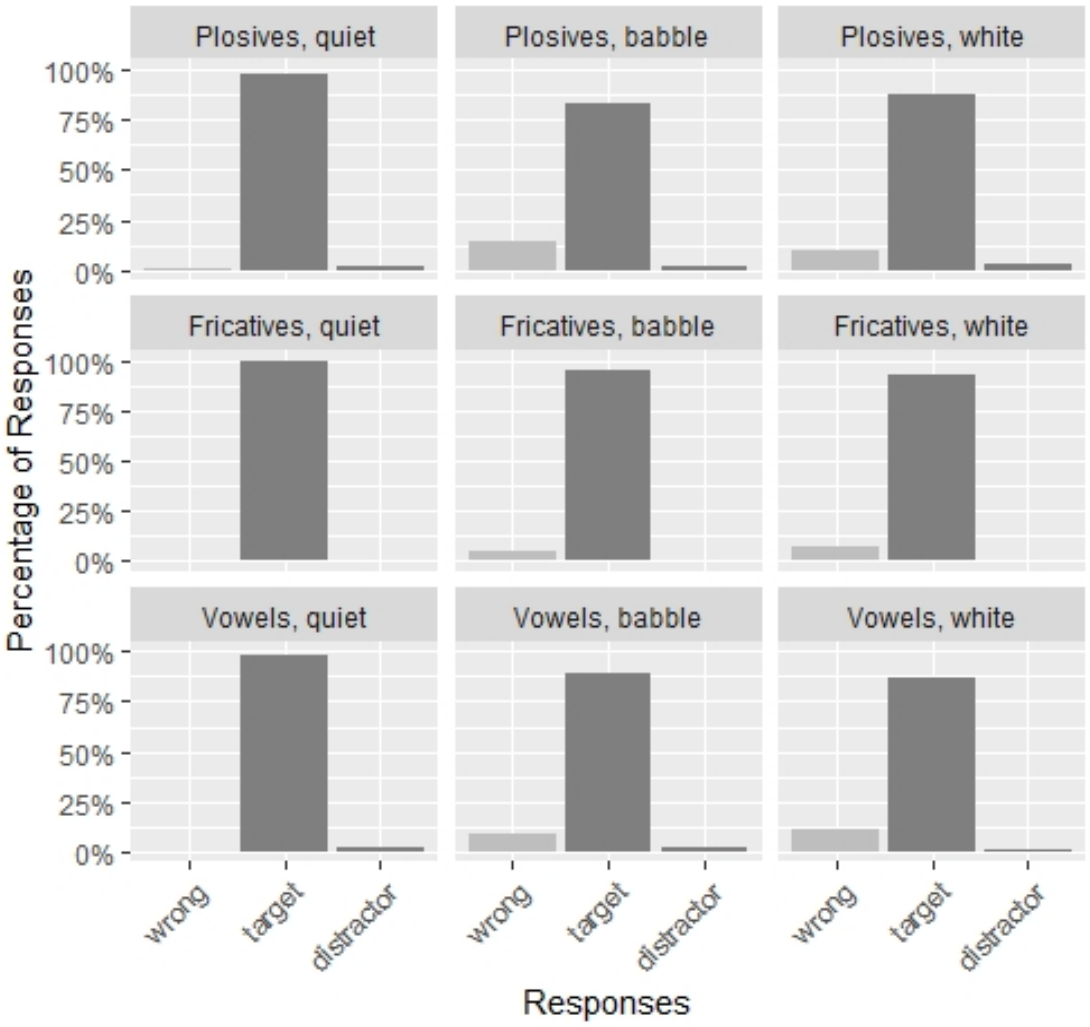

Figure 3.2: This figure shows the proportion of participants' responses (wrong, target, and distractor) for each of the three noise conditions (quiet, babble, white noise) and three sound types (plosives, fricatives, vowels) for the high predictability condition.

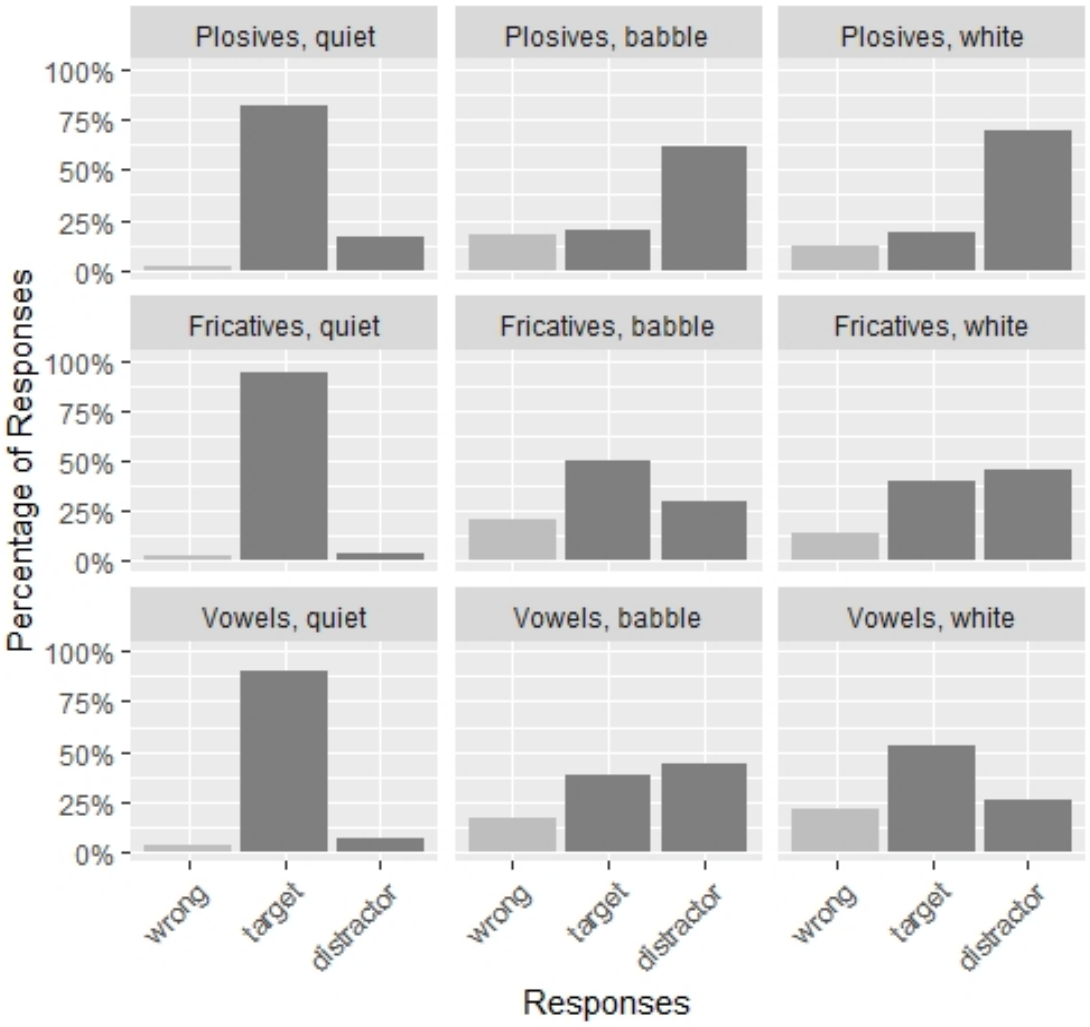

Figure 3.3: This figure shows the proportion of participants' responses (wrong, target, and distractor) for each of the three noise conditions (quiet, babble, white noise) and three sound types (plosives, fricatives, vowels) for the low predictability condition.

Table 3.4: Model Outcomes for the Overall Model (Low Predictability Subset)

	Estimate	*SE*	*Z*-value	*p*-value	
Intercept (Sound Contrast Plosives)	-0.71	0.25	-2.8	.01	**
Noise Contrast 1	-3.29	0.24	-13.53	$< .001$	***
Noise Contrast 2	-0.09	0.34	-0.26	.80	
Sound Contrast Vowels	1.88	0.29	6.6	$< .001$	***
Sound Contrast Fricatives	1.81	0.28	6.36	$< .001$	***
Trial Number	0	0.07	0.07	.95	
Noise 1 : Sound Contrast Fricatives	0.2	0.3	0.65	.52	
Noise 2 : Sound Contrast Fricatives	-0.6	0.39	-1.53	.12	
Noise 1 : Sound Contrast Vowels	0.62	0.29	2.13	$< .05$	*
Noise 2 : Sound Contrast Vowels	0.95	0.4	2.39	$< .05$	*
Sound Contrast Vowels vs Fricatives	-0.07	0.28	-0.26	.80	

and $\beta = 0.95$, $SE = 0.40$, $z = 2.39$, $p < .05$ for Noise contrast 2), but not when comparing Plosives to Fricatives (ps $> .12$). These results suggested different effects of Noise depending on the Sound Contrast. The model also showed a main effect of Sound Contrast, with more target responses for both Fricatives and Vowels compared to Plosives ($\beta = 1.88$, $SE = 0.29$, $z = 6.60$, $p < .001$ for Fricatives, and $\beta = 1.81$, $SE = 0.28$, $z = 6.36$, $p < .001$ for Vowels). As can be seen in Table 3.4, the other effects were not significant (ps $> .12$).

In order to compare Fricatives to Vowels, we reran the same model with Fricatives as the base level for Sound Contrast. This model showed no significant difference between Fricatives and Vowels ($p = .80$).

3.3.3 Interaction Sound Contrast and Noise

To get a better insight into the interaction of Sound Contrast and Noise, we next turn to the low predictability subsets of Plosives, Vowels, and Fricatives. These analyses will show exactly how the two types of background noise affected each of the contrasts. We expect in particular an adverse effect of white noise for fricatives, and an adverse effect of babble noise for vowels. For each subset, we ran a GLMM that only differed in its random structure, as indicated below. Again, Noise is contrast coded using forward Helmert coding. The first contrast shows the difference between Quiet on one hand and both noise types on the other (with weights of -1, 0.5, 0.5), and the second contrast shows the difference between Babble Noise and White Noise (with weights of 0, -0.5, 0.5). Results for the three models are presented in Tables 3.5, 3.6, and 3.7 for the subsets of plosives, fricatives, and vowels, respectively.

Table 3.5: Model Outcomes for the Subset of Plosives

	Estimate	*SE*	*Z*-value	*p*-value	
Intercept	-0.7	0.25	-2.82	$< .01$	**
Noise Contrast 1	-3.15	0.28	-11.29	$< .001$	***
Noise Contrast 2	-0.05	0.29	-0.16	.87	
Trial Number	-0.11	0.13	-0.86	.39	

The model for the LP subset of Plosives (all LP trials in which the minimal pair had a plosive contrast: N = 744) included two predictors, Noise and Trial Number (scaled as before). There were random intercepts for Participant and Item, but the inclusion of random slopes led to non-convergence and singular fit. We found only a significant effect of the first Noise contrast, so between Quiet and both types of Noise, with fewer target responses in Noise than Quiet (β = -3.15, SE = 0.28, z = -11.29, p < .001). The lack of a significant difference for the second contrast (p = .87) suggests that there is no difference between Babble Noise and White Noise in the effect they have on the recognition of the target words. As can be seen in Figure 3.3 in the two right-most panels on the top row, there is indeed no large difference in the amount of target responses depending on the type of noise.

For the LP subset of Fricatives (all LP trials in which the minimal pair had a fricative contrast: N = 719) we included the same two predictors, Noise and Trial Number, with the same random intercepts of Participant and Item as for the model for Plosives. This model contained an additional random slope of Noise for Participant. The model revealed both a significant adverse effect of Noise compared to Quiet (Contrast 1: β = -3.67, SE = 0.56, z = -6.58, p < .001), and a significant adverse effect of White Noise compared to Babble Noise (Contrast 2: β = -0.85, SE = 0.35, z = -2.42, p < .05). These results suggest that unlike Plosive items, Fricative target pairs are more strongly affected by White Noise than Babble noise. This can clearly be seen in Figure 3.3: While in Babble Noise (depicted in the middle panel of the figure) most responses are target responses, in White Noise (right-most panel on the middle row) there are more distractor responses than target responses, showing the difficulty of recognizing fricatives in this type of noise.

Finally, the model for the LP subset of Vowels (all LP trials in which the minimal pair had a vowel contrast: N = 696) also included the same two predictors as before, Noise and Trial Number. It had random intercepts for Participant and Item, with a random slope of Noise for Item. There was a significant effect of Quiet vs Noise (β = -4.58, SE = 1.26, z = -3.62, p < .001) with more target responses in Quiet.

Table 3.6: Model Outcomes for the Subset of Fricatives

	Estimate	*SE*	*Z*-value	*p*-value	
Intercept	1.4	0.39	3.55	$< .001$	***
Noise Contrast 1	-3.67	0.56	-6.58	$<.001$	***
Noise Contrast 2	-0.85	0.35	-2.42	$< .05$	*
Trial Number	0.13	0.15	0.88	.38	

Table 3.7: Model Outcomes for the Subset of Vowels

	Estimate	SE	Z-value	p-value	
Intercept	2.04	0.66	3.08	$< .01$	**
Noise Contrast 1	-4.58	1.26	-3.62	$< .001$	***
Noise Contrast 2	0.76	0.25	3.08	$< .01$	**
Trial Number	-0.01	0.12	-0.1	.92	

Additionally, the second contrast also showed a significant effect ($\beta = 0.76$, $SE = 0.25$, $z = 3.08$, $p < .01$), showing more target responses in White noise compared to Babble noise, an effect in the opposite direction as for Fricatives. Again, this is visible in Figure 3.3, where we see a majority of distractor responses in Babble Noise (middle panel on the bottom row), while in White Noise (right-most panel on the bottom row) there is a majority of target responses, showing this is the easier condition to recognize vowels correctly.

3.3.4 Semantic Fit and Phonetic Distance

To analyze which kind of information participants relied on when making their incorrect responses – top-down predictions or bottom-up auditory processes – we coded the semantic fit and phonetic distance to the target word for all wrong responses (N = 396; both from the HP and LP condition). Empty responses (where participants typed only *?* or -, for example, N = 5) were removed in this analysis. If participants rely more on semantic context to give their answer, we expect that their response fits the sentence context and perhaps has a higher phonetic distance to the target. If, on the other hand, they rely more on the sound signal, we expect that the phonetic distance to the target is smaller, but that the word might not fit the semantic context. We have only analyzed this for the wrong responses, as for the targets and distractors, these values are a given: In HP items, the distractor never fits the sentence, while in

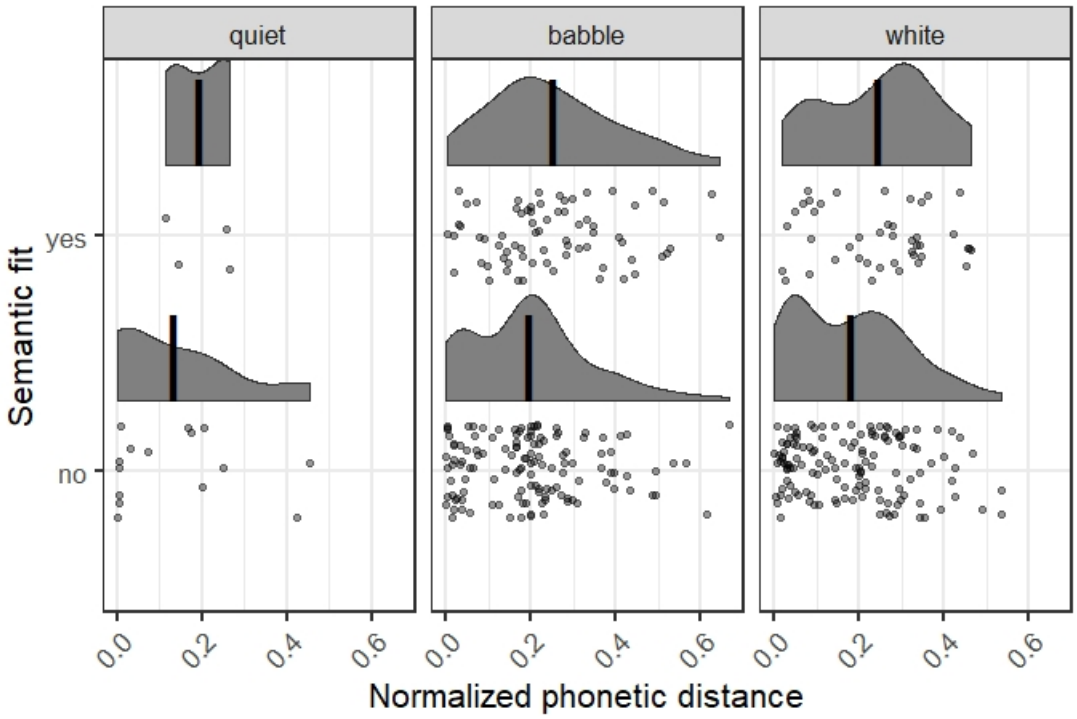

Figure 3.4: This figure shows the wrong responses that semantically fit or did not fit the sentence, plotted with the normalized phonetic distance, in each of the three noise conditions. Lower phonetic distance means more similar to the target item. The vertical black lines show the mean phonetic distance for each condition. Each dot represents a single wrong response, the shaded curves show the density plots for these.

LP the target never fits the sentence semantically, and they have a similar phonetic distance.

These results are presented in Figure 3.4, showing the semantic fit, yes or no, and normalized phonetic distance to the target word for each of the three noise conditions. Lower normalized phonetic distance scores mean that the participant's wrong response sounds more like the target. We see that for all noise conditions, most responses did not fit the sentence semantically: 15 vs 4 for Quiet, 128 vs 71 for Babble Noise, and 135 vs 43 for White Noise. The density peaks of the phonetic distance distributions as well as the mean values (shown by the vertical lines in Figure 3.4) are lying more towards the lower distances for those responses that do not fit the sentence context. This suggests a trade-off between acoustic fit and semantic fit: Participants made their response based on what they heard at a cost of fitting the semantic context.

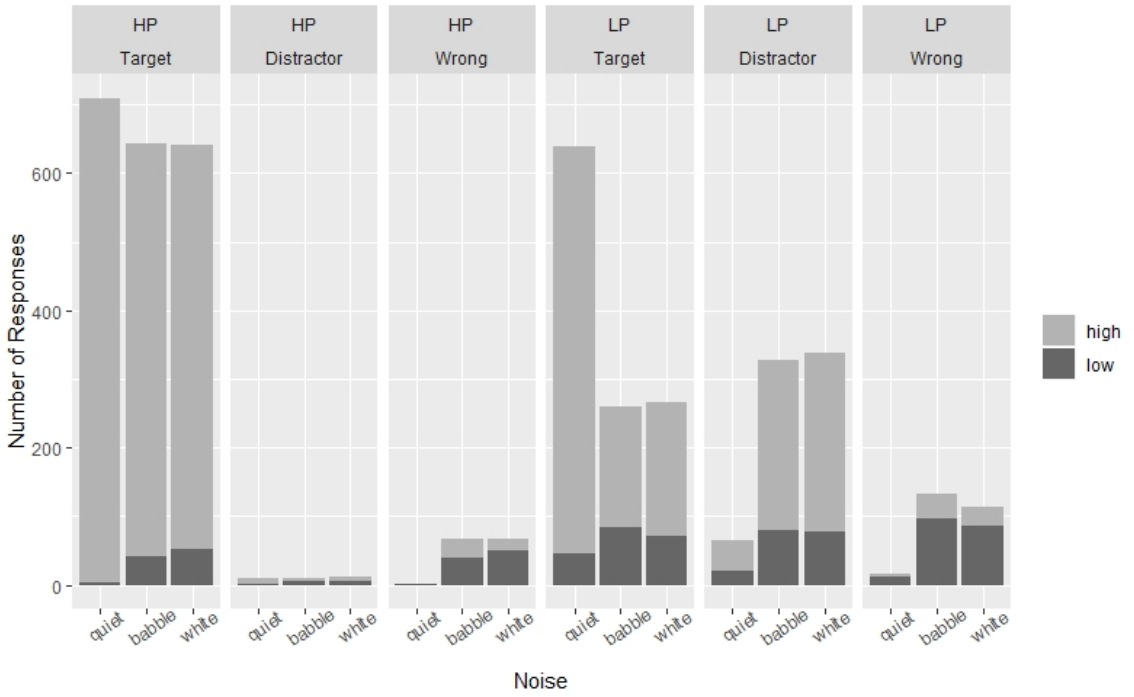

Figure 3.5: For all answer types (Target, Distractor, and Wrong) and predictability levels (High Predictability HP and Low Predictability LP) the responses are presented per noise type (Quiet, Babble, and White), showing high and low confidence ratings.

3.3.5 Confidence Ratings

In the experiment, we asked participants on every trial to rate how confident they were in giving the correct response. For this, they had a four-point scale, ranging from 1 (completely uncertain, guessed) to 4 (completely certain). Previous studies (Failes et al., 2020; Rogers et al., 2012; Sommers et al., 2015) reported false hearing effects, that is, increased confidence in giving the correct response even for incorrect responses when listening in noise. This effect has been found to be larger for older adults than younger adults, although a recent study (Van Os et al., 2021, see also Chapter 4) found no evidence of false hearing for either group of participants. As the present study only tested younger adults, we do not expect to find effects of false hearing. Instead, we expect that confidence ratings will depend on the difficulty of the listening condition, and thus be lower in adverse conditions like background noise, or low predictability items.

Figure 3.5 presents the participants' responses for each of the predictability and response type conditions, split for the three noise levels. We have transformed these responses into a binary variable of low confidence (ratings 1 and 2) and high

confidence (ratings 3 and 4) It is collapsed across the three speech sound types. It shows that the majority of responses in the high predictability condition are correct target responses, made with high confidence. In background noise, the number of wrong responses increases slightly, but these are rated with low confidence. In the low predictability condition, we find still many target responses in quiet, still with predominantly high confidence. In background noise, there is an increase in distractor responses, which are rated with high confidence. Again, wrong responses increase in noise, and are generally rated with low confidence.

In our statistical model, which is presented below, we used the binary confidence rating as in the figure, testing high and low confidence. We used the same statistical methods as described above. The final model included fixed effects of Noise (forward Helmert coded; the first contrast compared Quiet to the mean of both types of noise: weights -1, 0.5, 0.5; the second contrast compared babble Noise and white Noise: weights 0, -0.5, 0.5), Response Type (forward Helmert coded; the first contrast compared target responses to the mean of both other response types: weights -1, 0.5, 0.5; the second contrast compared distractor responses and wrong responses: weights 0, -0.5, 0.5), Predictability (categorical predictor with two levels, mapping the High Predictability condition on the intercept), and Trial Number (continuous predictor of the trial number in each block, scaled to improve convergence). The model also included all two-way interactions between Noise, Response Type, and Predictability. Additionally, the model included random intercepts by Participant and by Item. Inclusion of random slopes led to singular fit of the model.

We found a main effect of Response Type (β = -2.34, SE = 0.19, z = -12.06, $p < .001$), showing that participants were less confident of incorrect (Distractor and Wrong) responses than correct Target responses. The difference between Distractor responses and Wrong responses did not reach significance as a main effect ($p = .10$). A significant main effect of Predictability (β = -0.89, SE = 0.21, z = -4.19, $p < .001$) shows that participants had lower confidence ratings in Low Predictable items. A significant main effect for the first Noise contrast (β = -1.26, SE = 0.34, z = -3.73, $p < .001$) shows that participants were less confident of their responses in background noise compared to quiet. The second Noise contrast (comparing Babble to White Noise) did not reach significance ($p = .06$). We found a significant interaction of Response Type and Predictability ($\beta = 1.40$, $SE = 0.20$, $z = 6.93$, $p < .001$), showing higher confidence ratings for incorrect responses in the Low Predictability condition. This is most likely due to the distractor responses, which did fit the semantic context in this condition, and would therefore be made with higher confidence. In the Low

Table 3.8: Model Outcomes for the Overall Model for Analysis of the Confidence Ratings

	Estimate	*SE*	*Z*-value	*p*-value	
Intercept (Predictability High)	1.52	0.23	6.55	$< .001$	***
Response : Type Contrast 1	-2.34	0.19	-12.06	$< .001$	***
Response : Type Contrast 2	-0.87	0.52	-1.66	0.1	.
Predictability (Low)	-0.89	0.21	-4.19	$< .001$	***
Noise Contrast 1	-1.26	0.34	-3.73	$< .001$	***
Noise Contrast 2	-0.39	0.21	-1.96	0.06	.
Trial Number	-0.08	0.05	-1.52	0.13	
Response Type 1 : Predictability (Low)	1.4	0.2	6.93	$< .001$	***
Response Type 2 : Predictability (Low)	-1.35	0.53	-2.56	$< .05$	*
Noise 1 : Predictability (Low)	0.79	0.34	2.32	$< .05$	*
Noise 2 : Predictability (Low)	0.47	0.25	1.88	0.06	.
Response Type 1 : Noise 1	0.79	0.16	4.88	$< .001$	***
Response Type 2 : Noise 1	-0.61	0.42	-1.44	0.15	
Response Type 1 : Noise 2	-0.15	0.15	-0.95	0.34	
Response Type 2 : Noise 2	-0.28	0.33	-0.83	0.41	

Predictability condition, the confidence ratings for wrong responses were lower than for distractor responses (β = -1.35, SE = 0.53, z = -2.56, $p < .05$), in line with this explanation. The significant interaction of Predictability and Noise suggests that confidence ratings were higher in noise than quiet in the Low Predictability condition (β = 0.79, SE = 0.34, z = 2.32, $p < .05$), again, most likely due to increased confidence in distractor responses in this predictability condition. Finally, there was a significant interaction effect of Response Type and Noise ($\beta = 0.79$, $SE = 0.16$, $z = 4.88$, $p < .001$), showing higher confidence ratings for incorrect (Distractor and Wrong) responses in Background Noise (Babble and White) compared to Quiet. Results for all effects can be found in Table 3.8.

3.4 Discussion

The present study investigated how during speech comprehension multiple sources of information are combined by the listener, filling a gap in the empirical literature that has not combined context predictability, types of background noise, and a systematic manipulation of different types of speech sounds in a single experiment. We examined mishearings occurring when listening to speech in background noise, as a function of the predictability of the context and certain sound characteristics of the target word.

We were particularly interested in the interaction of background noise type (white noise or multi-speaker babble noise) and the sound contrast in the stimuli (minimal pairs of plosives differing in the place of articulation, tense/lax vowel pairs, and pairs of voiceless fricatives and affricates). We expected that, based on previous literature, the plosives would be difficult to recognize correctly in background noise in general, while for fricatives and vowels this would depend on the type of background noise (fricatives have energy in the same frequency bands as white noise (Phatak et al., 2008); while the formant values of vowels are easily lost in fluctuating babble noise (Benkí, 2003; Gordon-Salant, 1985; Pickett, 1957; Weber & Smits, 2003). This was confirmed in our study. In the high predictability condition, both the audio and the presented written sentence context point to the target, while in the low predictability condition the context supported the distractor response. Thus, finding distractor responses in the low predictability subset shows us that participants relied on the semantic context. Of course, as the target and distractor form minimal pairs, the speech signals for both words overlap greatly.

As expected, we found a main effect of noise, with more correct responses in quiet than in either type of background noise. This can be seen as a control condition, replicating various previous experimental findings (Gordon-Salant, 1985; Kalikow et al., 1977; Phatak et al., 2008; Van Os et al., 2021). Differences between babble and white noise occurred in interaction with the type of sound contrast (for detailed discussion, see below). Previous experiments showed conflicting results on whether babble noise or white noise disrupts recognition of speech most (e.g., Gordon-Salant, 1985; Horii et al., 1971; Taitelbaum-Swead & Fostick, 2016). The results from the present study suggest that the effect of the two types of noise is strongly dependent on the characteristics of the stimuli. This might partially explain the conflicting results in previous literature on this topic, besides other factors like the task, noise levels, and the tested population.

We did not find any significant effects of trial number in our data, suggesting there was no learning effect. Previous research (Van Os et al., 2021, see also Chapter 4) did report a learning effect with more correct answers as the blocks of the experiment went on, indicating that participants learned to rely less on the contextual information. Listeners have been found to re-weight acoustic and contextual cues based on their statistical properties (Bushong & Jaeger, 2019). It might be the case that in the present experiment, participants took longer to get used to the different background noise types and did not have time within each block to re-weight the

information from different sources (Van Os et al., 2021, used only a single type of background noise).

We additionally expected to find an interaction of background noise and the items' predictability. This expectation was confirmed by the data. The results provide evidence for the role of the bottom-up acoustic signal, as shown by the ceiling effect in the quiet listening condition also in low predictability items, as well as evidence for the role of the top-down signal, shown by the ceiling effects in the noisy high predictability items. The interaction between predictability and background noise in our data (as well as the interaction between speech sound and background noise type; see below) shows that these two types of evidence are rationally combined, as predicted by the Noisy Channel Model. Participants rely on the information that is most available to them in a particular experimental condition. We also see this trade-off between semantic context and acoustic signal when looking at the subset of wrong responses (those responses that were neither the target nor the distractor word from the minimal pair; N = 369). We coded the semantic fit to the target sentence context as well as the phonetic distance to the target word, and found that when the wrong response does not fit the semantic context, the phonetic distance to the target word is smaller, meaning the words sound more alike. In our experiment, participants, when making a wrong response, based this on the speech signal at a cost of fitting the semantic context. This is interesting, as we presented participants with the written context on the screen. It would have been very easy for them to rely on this information, despite being asked to focus on the speech. It might be the case that over the course of the experiment, participants learned to rely less on the written context, having realized it can be misleading. Of course, this analysis is based on a small subset of the data. Overall, participants tended to rely on the acoustic signal in quiet listening conditions and more on the semantic context when listening conditions were more difficult.

The Noisy Channel Model (Levy, 2008; Levy et al., 2009; Shannon, 1949) predicts that the level of perceived noise should affect how much listeners rely on the sentence context: this is stronger in higher levels of noise. The level of spectral overlap between noise and phoneme should affect how much listeners rely on the sentence context: this is stronger in higher levels of noise. In the present study, we manipulated the amount of overlap between the speech and noise signals by using combinations of different types of background noise and speech sound contrasts. Results followed our predictions. Plosive pairs are the most difficult speech sound (out of the ones tested in this study) to be recognized correctly, and they do not show a difference between

the two types of background noise. We find a difference between babble noise and white noise for the fricatives, whose recognition is impaired in white noise compared to babble noise, and for the vowels, which show the opposite pattern compared to fricatives. A previous study found many errors for fricatives in babble noise (Weber & Smits, 2003). These errors were found to be of varying kinds, namely manner, place, and voicing errors. In the stimuli of the present study, the most likely error was a place error, as the minimal pairs were constructed to differ only in place of articulation, keeping other features the same. Therefore, we would expect fewer errors in the recognition of fricatives in babble noise, as the chance of some errors is reduced. While participants could respond from an open set of candidate words (as it was not a multiple choice task), they most often responded with the distractor rather than a different word.

In their original study testing the predictions of the Noisy Channel Model, Gibson et al. (2013) quantified the edit distance between the plausible and implausible version of their alternations in a change consisting of insertions and deletions of function words. The present study changed the type of background noise to manipulate the distance between the target and distractor. Here the distance between the two depends on the similarity between the acoustic signal of the speech sound and that of the background noise. As such, it is better grounded in the strength of masking of the signal, and less arbitrary than the edit distance measured in terms of insertions and deletions. Using auditory stimuli of a different type than syntactic alterations, we have found support for the Noisy Channel Model's predictions in a more naturalistic setting.

Not only the Noisy Channel Model predicts that multiple sources of information are combined during speech comprehension. This idea goes back to older models of speech comprehension (e.g. Luce & Pisoni, 1998; Oden & Massaro, 1978; Norris & McQueen, 2008), and is not new. However, these models are generally based on empirical data from studies investigating the perception of mono-syllabic words rather than sentences or larger contexts. Additionally, they primarily focus on explaining effects of frequency and small local contexts consisting of surrounding syllables. The Noisy Channel Model, instead, makes specific predictions about listening situations with background noise. This is a common occurrence in every-day language use, but is not explained by other models of speech comprehension.

A theory that stands in contrast to algorithmic computations, such as rational Bayesian models like the Noisy Channel Model, is the Shallow Processing account, also known as "good-enough processing" (Ferreira, 2003; Ferreira et al., 2002; Ferreira

& Patson, 2007). It states that language comprehension relies on heuristic processing as well as algorithms, generating superficial interpretations of sentences that can in fact be inaccurate. Two heuristics proposed by (Ferreira, 2003) are based on plausibility and word order. The first studies investigating this theory based their evidence on thematic role assignment in active and passive sentences, as well as subject- and object cleft sentences and garden-path sentences, and argued that the use of "good-enough representations" is common in language processing in general (Ferreira, 2003; Ferreira et al., 2002). The representations based on shallow processing are often good enough for everyday communication (as opposed to being tested in psycholinguistic experiments), the most common task that listeners perform, leading only occasionally to misunderstandings (Christianson, 2016; Ferreira, 2003). Through shallow processing, comprehension can be fast enough to keep up with dialogue and takes as little effort as possible. Using plausibility-based heuristics in shallow processing is generally quicker than using syntactic algorithms, and it might also be a strategy to conserve available resources, especially for older adults (Ayasse et al., 2021).

The Noisy Channel Model and Shallow Processing are similar in multiple aspects and share some central ideas such as processing guided by context and at times incorrect final interpretations (Christianson, 2016; Kuperberg & Jaeger, 2016; Traxler, 2014), and both explain how misinterpretations during speech comprehension can occur. Previous papers have made suggestions to link (Bayesian) predictive models and the Shallow Processing account (Ferreira & Lowder, 2016). They do this through the notion of information structure, distinguishing given and new information. Following Haviland and Clark (1974), they state that given information has already been introduced and stored away in long-term memory, while new information needs to be integrated with this knowledge. This information can come from various sources: lexical, syntactic, semantic, or pragmatic levels. Ferreira and Lowder (2016) suggest that only the part of the sentence that presents given information is processed shallowly, while the rest of the sentence is processed following (Bayesian) algorithms. As there is a tendency to present given information before new information, comprehenders would assume this is the case, and process the beginning of the sentence shallowly, while processing the latter part deeply (Ferreira & Lowder, 2016). In psycholinguistic experiments, like the current one, usually each item is a separate sentence without context provided by a longer discourse. Thus, there is hardly any information structure on the discourse level when interpreting these separate sentences, which according to this theory would mean that there would be no shallow processing in these experiments. Especially in the present study, the critical part is sentence-final and should be processed following Bayesian algorithms rather

than through shallow processing. It is unclear what predictions are made by Shallow Processing in psycholinguistic experiments where each item consists of an unrelated sentence. Additionally, as mentioned above, Shallow Processing might not line up with the goals of psycholinguistic experiments as these can differ from those in everyday language use. It should be clarified exactly when and where Shallow Processing takes place, so that it can be contrasted to other approaches and directly tested in experiments.

Kuperberg and Jaeger (2016) argued that shallow processing may arise from prediction, such that highly constraining sentence contexts which give rise to strong anticipations should lead to shallow processing. This would fit with our high predictability condition: The prediction in this case would be that misinterpretation of the sentence-final word would be common in the low predictability condition in the present experiment, also in quiet. Here it contrasts with predictions made by the Noisy Channel Model, and, crucially, also the results of the present study. In the low predictability condition in quiet, we find target responses close to ceiling, particularly for fricatives and vowels, showing that participants were relying on the acoustic signal rather than the (misleading) sentence context.

Furthermore, the Shallow Processing account does not make detailed predictions about the different noise and sound contrast conditions in the present experiment. In particular, Shallow Processing does not depend on the clarity of the signal, like the Noisy Channel Model does. Therefore, there are no specific predictions about any possible interaction effects of noise type and speech sound contrast. In order to compare the Noisy Channel Model and the Shallow Processing account in experiments such as the present one, the Shallow Processing account should be extended so that it makes predictions about this interaction and specifies how the clarity of the (acoustic) signal influences processing depth.

Another line of research that investigates how bottom-up and top-down processes interact is that of local coherence effects (e.g. Konieczny et al., 2009; Kukona et al., 2014; Kukona & Tabor, 2011). A sentence like "The coach smiled at the player tossed a Frisbee by the opposing team" contains a locally coherent phrase "The player tossed a Frisbee". Using various methods, studies have shown that these locally coherent phrases influence sentence processing, leading to longer reading times and error detection times (Konieczny, 2005; Konieczny et al., 2009; Tabor et al., 2004). This suggests that participants actively considered the phrase "The player tossed a Frisbee" even though it is incompatible with the earlier parts of the sentence. The work on local coherence shows that the bottom-up cues in language comprehension

get integrated as well, even if they are inconsistent with the top-down predictions of the sentence structure. Thus, top-down predictions do not rule out bottom-up perceptions, people engage in an interpretation that is locally coherent, even though this interpretation is incompatible with the prior context. In the current experiment, this would mean that in the LP condition, participants consider the target response even if they respond with the distractor word in the end (if the word was sufficiently audible). Future research could investigate how the information from the acoustic signal and the semantic context get integrated online in an experiment similar to the one reported here.

In this chapter, we also analyzed the participants' confidence ratings. These results suggest that participants' confidence ratings depend on the difficulty of the listening condition and how much evidence they have to support their response. We see higher confidence ratings in quiet listening conditions, in particular when the response was a correct target one. In the high predictability condition, the confidence ratings are high for the two noise conditions as well. Here, despite the noise, both the provided sentence context and the acoustic signal point to the target, thus giving participants multiple sources to base their response on. This combination of information sources leads to higher confidence. When two sources of information are conflicting, as is the case in the low predictability condition, the number of responses as well as the confidence ratings drop in background noise for target responses, while they rise for distractor responses, which fit the semantic context. Wrong responses, which might fit either the semantic context or the acoustic signal (or neither) are rated with lower confidence. These results are in line with other findings (Van Os et al., 2021, see also Chapter 4).

3.4.1 Limitations

One limitation of the current study is its online design. While it allowed us to collect data in lock-downs during the Covid-19 pandemic, it also meant we were unable to control the audio setting as we would have been able to in a lab study. We could not collect hearing thresholds of participants and had to rely on self reported hearing issues. It is possible that some participants were unaware of existing problems with their hearing. Furthermore, we could not control the equipment participants used to play the audio in the experiment, nor the level they set or the amount of background noise in their surroundings. By using instructions and a post-experimental questionnaire, we tried to get an idea of these factors for each participant. Still, this source of variation due to the online design might have affected our results.

While we tested three different categories of speech sounds (plosives, fricatives, vowels), we did not control for possible variation in recognition within each category. Previous studies have found that coarticulation effects play a role in the recognition of speech sounds, in particular consonants (Alwan et al., 2011; Gordon-Salant, 1985). In our study, we were not able to control the direct phonetic context of the minimal pair contrasts to minimize these effects. Studies have also found that within each type of sound contrast, some sounds might be more robust to noise than others. For example, ʃ (as in *ship*) has been found to be easier to recognize than other voiceless fricatives (Gordon-Salant, 1985; Weber & Smits, 2003). For vowels, the second formant values are most strongly obscured in background noise (Parikh & Loizou, 2005). We used pairs of tense and lax vowels, but for these sounds there are larger differences in the second formant for the back vowels than for most front vowels (Hoole & Mooshammer, 2002). This leads to differences in recognition or adverse effects of noise within our sound contrast categories. Here, the unbalanced contrasts within each category (plosives, fricatives, vowels) might have affected the results, as it is possible that the interference of noise type is stronger for some of the contrast pairs, that might have occurred more or less in our stimuli as a whole. An additional factor here is that our experiment is limited to having a single female speaker. With speakers of the same sex with different speech / voice characteristics or speakers of a different sex, the results might differ. The voices of these speakers might interact slightly differently with the noise types, affecting which sounds are recognized better or worse. The experiment should be replicated with different speakers to test for more robust effects.

Other factors have been found to affect word recognition, such as the lexical status of the word (real word vs non-word), word frequency, length, and neighborhood effects. Due to the way we constructed our stimuli, we were not able to carefully control our items for all of these factors.

3.5 Summary

During speech comprehension, multiple sources of information are available to the listener. Major models of speech comprehension (e.g. FMLP, Oden & Massaro (1978); NAM, Luce & Pisoni (1998); Shortlist B, Norris & McQueen (2008)) already combine multiple sources of information. However, these models are often based on empirical data that is based on mono-syllabic word recognition rather than full sentences or larger contexts. Previous studies that investigated predictability effects in noise

did not carefully control the types of sounds and how they are affected by noise (Boothroyd & Nittrouer, 1988; Dubno et al., 2000; Hutchinson, 1989; Kalikow et al., 1977; Pichora-Fuller et al., 1995; Sommers & Danielson, 1999), while the literature on effects of background noise on speech sounds does not specifically manipulate predictability effects in sentence comprehension (Alwan et al., 2011; Cooke, 2009; Gordon-Salant, 1985; Phatak et al., 2008; Pickett, 1957). Additionally, results on the effect of background noise are inconclusive regarding which type of noise affects comprehension most severely (Danhauer & Leppler, 1979; Gordon-Salant, 1985; Horii et al., 1971; Nittrouer et al., 2003; Taitelbaum-Swead & Fostick, 2016). We addressed this gap by presenting an experiment that investigated how different sources of information are combined while manipulating the predictability of the context and the clarity of the acoustic signal. Our stimuli contain small differences in intelligibility by combining different types of background noise with different speech sound contrasts that are more or less strongly affected by that noise. Our results show that it is important to consider the effect of the type of noise masker. Listeners use all the cues that are available to them during speech recognition, and these cues crucially depend on the masking noise in the background. In this process, listeners are rational and probabilistically combine top-down predictions based on context with bottom-up information from the acoustic signal, leading to a trade-off between the different types of information.

Chapter 4

Exp. 2: Mishearing and False Hearing

In the previous chapter we investigated mishearing in younger adults. We found that this population relied more on sentence context when the listening conditions were more difficult. Here, difficulty is measured by the overlap between the speech signal and background noise signal. However, as described in Chapter 2 there are differences between younger and older adults that warrant a comparison of these two groups. Older adults have more language experience and hence should have better expectations based on context (Pichora-Fuller, 2008; Sheldon et al., 2008), while at the same time they may already be subject to some hearing loss and know to trust the incoming signal less. Given both of these factors, we would predict based on the Noisy Channel Model that older adults show larger effects of the top-down signal on interpretation, and thus be subject to stronger mishearing effects than younger adults.

A phenomenon related to mishearing is that of false hearing. With false hearing, the listener mishears a word, but is very confident that they identified the word correctly. As such, it differs from simple mishearings in the listener's subjective state. Previous studies (Failes et al., 2020; Failes & Sommers, 2022; Rogers et al., 2012; Rogers, 2017; Sommers et al., 2015) have investigated this phenomenon and found that older adults show a stronger false hearing effect than younger adults. Besides mishearing, the present chapter investigates false hearing with carefully controlled stimuli that vary in the clarity of their acoustic signals.

In the following sections we will first briefly summarize the literature that explains how speech comprehension processes change with age and provides details on

false hearing (Section 4.1). This literature has been covered extensively in Chapter 2. The research goals of the experiment and our hypotheses are defined in Section 4.1.1. In Section 4.2 we will describe the methods that were used in the experiment, with its results provided in Section 4.3. Finally, Section 4.4 gives a discussion of the results, and a summary of the chapter is given in Section 4.5. This chapter is based on and is in some places identical with Van Os et al. (2021).

4.1 Introduction

As described in Chapter 2, older adults differ from younger adults in various aspects that are relevant to speech comprehension. First, there are changes in auditory processing with increasing age (Gordon-Salant et al., 2010; Helfer et al., 2020), so that older adults might have hearing loss at the frequencies important for speech comprehension (Gates & Mills, 2005). They might particularly struggle to differentiate between frequencies, distinguish spectral and temporal transition, and localize sound sources (Hnath-Chisholm et al., 2003; Helfer et al., 2020; Schuknecht & Gacek, 1993; Tun et al., 2012). These changes lead older adults to have greater difficulty understanding speech, especially in adverse listening conditions such as the presence of background noise (Li et al., 2004; Pichora-Fuller et al., 1995; Pichora-Fuller et al., 2017; Schneider et al., 2005). Besides these auditory changes, older adults also show decreased attention, working memory, executive functions, and processing speed (Lindenberger & Ghisletta, 2009; Salthouse, 1990; Salthouse, 1996; Tucker-Drob et al., 2019; Tun et al., 2012), negatively impacting speech comprehension. This holds particularly in noisy environments, as studies have found that older adults are more negatively affected by the presence of background noise compared to younger adults (Benichov et al., 2012; Dubno et al., 2000; Hutchinson, 1989; Pichora-Fuller et al., 1995; Sommers & Danielson, 1999).

General language abilities do not tend to degrade with age, which means that older adults can use knowledge-based factors to compensate for reduced auditory and cognitive abilities. For example, they are able to use supportive sentence contexts to make up for the negative effect of adverse listening conditions (Lash et al., 2013; Stine & Wingfield, 1994; Wingfield et al., 1995; Wingfield et al., 2005). This means that older adults generally rely more on the given context than younger adults do, and are particularly adept at this due to their experience with language and every-day challenging listening situations.

In cases where the context is unhelpful or misleading, this tendency to rely on it might lead to mishearings or even false hearing. False hearing is when a listener is very confident of having correctly identified a word, but in reality misperceived it. This has been studied by using a word priming paradigm (Rogers, 2017; Rogers et al., 2012) and by using predictable and unpredictable sentences (Failes et al., 2020; Failes & Sommers, 2022; Sommers et al., 2015). Items were played in background noise, and participants had to identify the target word. Results from these studies showed that older adults performed better than younger adults in the congruent trials (with supportive contexts) but had a higher false alarm rate for incongruent items (with misleading contexts). This means that older adults rely more strongly on the contextual information than younger adults. Older adults were also more confident of these incorrect responses than younger adults, showing the increased false hearing effect for older participants.

Like in Experiment 1 (Chapter 3), in the current experiment, we aimed to test the predictions of the Noisy Channel Model (Levy, 2008; Levy et al., 2009; Shannon, 1949). According to this model, speech comprehension in background noise is a rational process, where the listener utilizes all available sources of information. Thus, bottom-up information is combined with top-down predictions based on context. Previous studies investigating the interpretation of syntactic alternations in both written and auditory form, found that the results followed the predictions made by the Noisy Channel Model (Gibson et al., 2013; Gibson et al., 2016; Gibson et al., 2017; Poppels & Levy, 2016; Ryskin et al., 2018).

4.1.1 Research Goals and Hypotheses

In this second experiment, we aimed to investigate how background noise affects speech comprehension in younger and older adults, in situations where there is a predictive sentence context available that might facilitate or hinder speech recognition. We used a similar design as in the experiment in the previous chapter. In our experiment, both younger and older adults completed a word recognition task, where sentences were either presented in quiet or in background noise, and where the sentence context could be used to either predict the sentence-final target word or mislead the listener. These sentence-final target words were designed to be minimal pairs with respect to pronunciation, so that in the low predictability context, the word sounded very similar to the word that in fact did fit the sentence semantically. Like before, this allowed us to investigate whether listeners are able to rely on small acoustic cues

for word recognition, even in background noise, while keeping sentence contexts equal across conditions.

This experiment investigated how bottom-up auditory processes and top-down predictive processes interact in speech comprehension. We tested both younger and older adults in our experiment, as we expect age differences in the quality of top-down and bottom-up processes. Based on previous literature, we expected older adults to rely more heavily on the provided sentence context to make their responses, and to be more negatively affected by the presence of background noise than younger adults. The main question that the present study aimed to address is the replicability of mishearings for older adults in German. Like previous studies testing English (Failes et al., 2020; Failes & Sommers, 2022; Sommers et al., 2015), we used a paradigm of word recognition in sentences, where the context was predictive or unpredictive of the target word. We added a quiet condition without background noise as a baseline condition, which allowed us to make sure that hearing ability between groups was comparable with respect to our materials. It was also possible that we would observe a general increase in mishearing in older adults compared to younger adults, even in the quiet condition. This would be an interesting finding, as it would show that older adults rely more on context than the acoustic signal even if the acoustic signal was easily accessible. This is comparable to the finding that older adults rely more on domain-general cognitive processes in challenging listening conditions with high intelligibility (Koeritzer et al., 2018). Like previous studies, we collected confidence ratings to investigate false hearing as a second point of interest.

We aimed to test the predictions of the Noisy Channel Model for older adults as well. Like in the previous experiment, we manipulated the edit distance between targets and distractors by using two different types of speech sound contrasts (plosives and vowels) embedded in babble noise at two different levels (0 dB SNR and -5 dB SNR). This differs from previous studies (Gibson et al., 2013; Gibson et al., 2016; Gibson et al., 2017; Poppels & Levy, 2016; Ryskin et al., 2018) which investigated syntactic alternations in English and that did not use acoustic noise. The acoustic properties of our manipulation of speech sounds differ in various ways. First, vowel sounds have a longer and steadier signal compared to the relatively short burst of the plosives. Second, higher frequencies are more informative for plosives than for vowels, in particular for place of articulation (Alwan et al., 2011; Edwards, 1981; Liberman et al., 1954), which was the contrast in our minimal pairs. Based on the Noisy Channel Model, we expected to find that the top-down predictions play a larger role in the case of plosives, as here the signal of the target and distractor are more similar to

each other compared to the vowel condition, and thus will have more flat probability distributions (where both the target and the distractor have a similar probability of leading to the observed acoustic signal) based on the bottom-up processes. Listeners try to overcome this by relying more on the contextual information that is more easily accessible and provides distinguishing information. Furthermore, we expected that this difference between vowels and plosives may also be more pronounced in older adults, as hearing ability in high frequency ranges is known to degrade during aging (Gates & Mills, 2005). Listeners optimally combine bottom-up and top-down probabilities, leading to mishearing in difficult listening conditions where the choice of the most likely word is mostly determined by the top-down prediction, an effect that is stronger for older adults as they compensate for age-related reductions in auditory and cognitive processing, but still rationally combine the acoustic and top-down information that is available to them.

4.2 Method

4.2.1 Participants

A total of 93 native German speakers participated in the present experiment, for which we used the recruitment platform Prolific (www.prolific.co). We excluded seven older participants based on their performance in the quiet condition, because their number of distractor responses exceeded that of the younger adults. In this way, we tried to ensure equal hearing abilities with respect to our stimuli across ages, as we were not able to collect hearing thresholds for our participants (because of the Covid-19 pandemic, in-lab experiments were not possible at the time of conducting this study). The high number of unexpected responses in this relatively easy condition without background noise might also have been due to difficulty playing the audio or doing the task. The mean age of our final group of participants was 40 years (age range = 18–68 years), 43 were male. While all participants were self-reported native speakers of German, their current countries of residence varied: 55 lived in Germany, 12 in the United Kingdom, 4 in Austria, 3 in Ireland and Spain, 2 in the USA, 1 in each of France, Israel, Portugal, Poland, and South Korea. Three did not list their country of residence. Three out of our 87 participants reported to not speak other languages besides German, all three were older adults. From the remaining 84 participants, the languages spoken besides German were most often English (reported by 82 participants), French (reported by 21), and Spanish (reported by 14). In

the post-experimental questionnaire, most participants reported no hearing issues or use of hearing aids. One participant (age 29) reported tinnitus, and one reported reduced hearing in his right ear (age 48, 60% hearing left). In order to check for any effects of education, we computed Spearman's correlation between participants' age and education level. This correlation was small ($\rho = 0.2$, $p = .08$), indicating that the older participants in our study had a slightly higher education than the young participants. All participants gave informed consent, and the study was approved by the Deutsche Gesellschaft für Sprachwissenschaft (DGfS) ethics committee. The experiment lasted approximately twenty minutes and all participants received 3.12 Euro as compensation for their participation. To ensure we could find sufficient older participants, those older than 35 years were allowed to do two experimental lists if they wanted. In this case, they were paid double.

4.2.2 Materials and Task

In this experiment, our stimuli consisted of 120 minimal pairs, around which sentences were constructed where the target word occurred in sentence-final position. These materials were a subset of the set of items constructed in Experiment 1. For more details on the construction, see Section 3.2.2. These 240 sentences were highly predictable (HP; as tested in a cloze task). We additionally constructed 240 low predictability (LP) sentences by swapping the two words of a pair. Example stimuli can be found in Table 4.1, while the full set of items is presented in Appendix A. Approximately half of the minimal pairs (N = 62) differed in the plosive in medial position, while the other half of minimal pairs (N = 58) differed in their medial vowel (the items with a fricative or affricate pair were not used in this experiment). Recordings were made of all HP sentences, and we subsequently constructed the LP sentences by cross-splicing the target words to their other context. This resulted in natural sounding, mostly grammatical sentences. The final 480 recordings were embedded in background noise, a multi-speaker café babble noise (in the current experiment, we did not use white noise). This was done at two different SNR ratios, namely 0 dB SNR and -5 dB SNR. In the 0 SNR condition, the target sound and background noise have an equal intensity and this is thus easier than the -5 SNR condition. This gave a total of 1140 audio files (480 in each noise condition (Quiet, -5 SNR, and 0 SNR), half of which were HP and half of which were LP).

The recorded items were presented to participants, who completed a word recognition task. They listened to each item in turn and then were presented on the screen

Table 4.1: Example Stimuli (Experiment 2)

1A	Am Pool im Hotel gab es nur noch eine freie **Liege**. *At the pool in the hotel there was only one free **lounger** left.*	HP
1B	Nach vier Jahren heiratete Paul seine große **Liebe**. *After four years, Paul married his big **love**.*	HP
1C	Am Pool im Hotel gab es nur noch eine freie **Liebe**. *At the pool in the hotel there was only one free **love** left.*	LP
1D	Nach vier Jahren heiratete Paul seine große **Liege**. *After four years, Paul married his big **lounger**.*	LP

Note. Highly predictable sentences (HP) were made based on minimal pairs (Liebe / Liege, target words in **bold**) in 1A and 1B), then sentence-final target words were swapped to make low predictability items (LP) with the sentence frames of 1A and 1B, resulting in 1C and 1D. English translations have been given in *italics*.

with the sentence context up to the target word. Their task was to type in the word they had heard, and additionally rate their confidence level on a four-point scale.

4.2.3 Design

The experimental items were arranged in a Latin Square design. Twenty-four different lists were constructed, consisting of 60 items each. These lists were constructed in such a way that each noise level and each predictability level occurred the same number of times, and that each item appeared only once per list (same target pair or same predictability sentence). This was done in a crossed design, so that out of the 60 items, 30 were predictable and 30 were unpredictable. Out of each set of 30 items, 10 were presented in quiet, 10 in 0 SNR noise, and the remaining 10 in -5 SNR noise. The items were blocked by noise level, starting with 0 SNR, followed by -5 SNR, and ending with the quiet condition. This blocking was chosen to give participants a chance to maximally adapt to the noise and the task, starting with the relatively easy noise condition before being presented with the relatively hard noise condition. The quiet condition was presented at the end, so as not to give away the goal of the experiment at the start. Each list was preceded by a practice block, consisting of four items. This short practice block made the participants familiar with the task and online testing environment. All noise levels (quiet, 0SNR, and -5 SNR) were presented during the practice block.

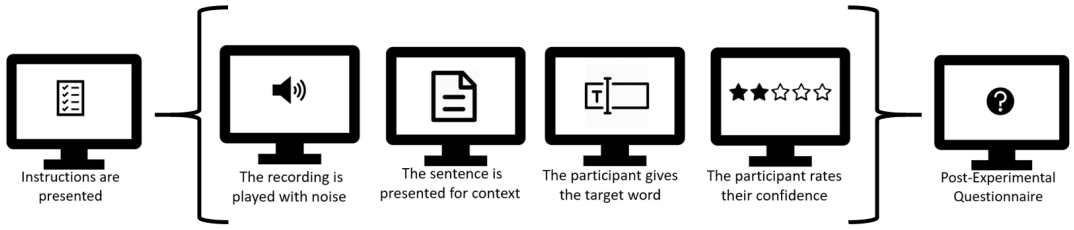

Figure 4.1: This figure shows the different stages of the experiment, with a single trial between brackets. Participants completed four practice trials and sixty experimental trials.

4.2.4 Procedure

The experiment was hosted on Lingoturk, a crowdsourcing client (Pusse et al., 2016). Participants completed the experiment on a computer in a quiet room and using the Chrome webbrowser. They were instructed to use either headphones or speakers. In the experiment, participants had to listen to the sentence and report the final word they had heard. Before the start of the main experiment, the participant saw a series of instructions detailing the task (instructions (in German) are given in Appendix D). Participants were asked to listen carefully and report what they heard. We did not explicitly state that the sentences could be misleading. These screens included a sound check as well, so that the participant had the opportunity to make sure the sound was being played correctly. In the main task of the experiment, the sentence, minus the target word, was presented on the screen in written form. We opted to include the written sentence up until the target word to make sure participants were able to use the context also in noisy conditions. A text box was provided for the participant to type their answer. Additionally, they rated their confidence in having given the correct answer on a scale from 1 (completely uncertain, guessed) to 4 (completely certain). At the start of a trial, the sound played automatically while the screen showed a fixation cross. Next, a screen with the two questions appeared after the recording had finished playing. The next item started playing as soon as the participant clicked to go to the next trial. As mentioned above, the experiment started with a short practice session consisting of four items, which were presented after the participant had seen all instructions. A schematic overview of the experiment is presented in Figure 4.1.

4.2.5 Analyses

After data collection had been completed, all received answers were first classified automatically on whether it was the *target*, the word that was played in the audio, (e.g., in example 1A in Table 4.1 "Liege" / "lounger"), the similar sounding *distractor* (e.g., in 1A "Liebe" / "love"), or was a different word entirely (e.g., in 1A "Platz" / "space", *wrong*). The list of answers that had been classified as wrong was then checked by the first author and a native German-speaking student assistant, to correct misclassifications because of typos. In our statistical analyses, we included the trial number of each block (1-20) as a control variable to check for any learning effects. We analyzed the high predictability and low predictability items separately due to ceiling effects in the high predictability condition. To determine whether participants relied on the sentence context or on the speech signal, we coded the semantic fit of the incorrect responses (fitting or not fitting), as well as the phonetic distance between the incorrect responses and target and distractor items. We made phonetic transcriptions based on the Deutsches Aussprachewörterbuch (German Pronunciation Dictionary; Krech et al., 2009) and calculated the weighted feature edit distance using the Python package *Panphon* (Mortensen et al., 2016). This distance was normalized by dividing it by the longest of the two compared words. The normalized distance fell between 0 and 1.

4.3 Results

In the first part of the results section (Sections 4.3.1, 4.3.2, and 4.3.3), we will report the results on age differences in response accuracy in the high and low predictability conditions investigating mishearing. In the second part (Section 4.3.4), we will analyze confidence ratings, and investigate age differences in the false hearing effect. We used general linear mixed models (GLMM; Quené & Van den Bergh, 2008, for a tutorial see Winter, 2019), implemented in the lme4 package (Bates et al., 2014) in R (R Core Team, 2022) to analyze our data. These models allow both fixed and random effects, letting us control for variation on the participant- and item-level (Baayen et al., 2008; Barr et al., 2013). To improve convergence, all models were run using the bobyqa optimizer and increased iterations to 2·105. Model comparisons were made to guide model selection based on the Akaike Information Criterion (AIC), models with the lowest AIC are reported below.

4.3.1 High Predictability helps Comprehension in Noise

We were interested in whether listeners are able to pick up on small acoustic cues identifying words in minimal pairs, in quiet but also in background noise. For the initial analyses, we used a subset of our data consisting of the participants' target and distractor answers, thus disregarding the wrong responses. We tested the participants' binomial responses (0 = distractor, 1 = target) using a GLMM with a logistic linking function. First, we analyzed the subset of the high predictability items. In this analysis, all confidence ratings were collapsed. The model included fixed effects of Noise (categorical predictor with three levels using dummy coding, mapping the Quiet condition to the intercept), Age (continuous predictor, scaled to improve convergence), ContrastVP (categorical predictor with two levels using dummy coding, mapping Plosive to the intercept), and Trial Number (continuous predictor with trial number within each block, scaled to improve convergence). Additionally, the model included the interaction of ContrastVP and Age (scaled) and the interaction of Trial Number and Age (both scaled). The model included no random effects, since this led to non-converging models or singular fit. The model revealed a significant effect of Noise, where participants more often give the distractor answer noise compared to the quiet condition (β = -1.38, SE = 0.65, z = -2.11, $p < .001$ for 0SNR, and β = -1.61, SE = 0.64, z = -2.48, $p < .001$ for -5SNR). As can be seen in Table 4.2, all other effects were not significant (all p-values $> .35$). The noise effects are relatively small, and overall participants score close to ceiling, where most responses are target responses. These effects can also be seen in the two left-hand panels in Figure 4.2.

4.3.2 Effects of Noise and Phoneme Change on Comprehension

The model for the low predictability subset of the data included the same fixed effects as the high predictability subset, but included by-Participant and by-Item random intercepts and a random slope for Noise for the by-Item random intercept. The model revealed a significant effect of Noise, where the noise conditions had more distractor responses than Quiet (β = -4.87, SE = 0.32, z = -15.22, $p < .001$ for 0SNR, β = -6.32, SE = 0.39, z = -16.09, $p < .001$ for -5SNR). Additionally, the model revealed a significant effect of Trial Number (β = 0.47, SE = 0.08, z = 5.97, $p < .001$), meaning that participants slightly increased the amount of target responses with practice. The interaction of Age and Trial Number was not significant (p = .95), suggesting older adults also showed this learning effect. The model also revealed that items of minimal

Table 4.2: Model Outcomes for High and Low Predictability Items.

	High predictability items subset				
	Estimate	*SE*	*z*-value	*p*-value	
Intercept (quiet, P)	5.77	0.61	9.53	<.001	***
Noise -5SNR	-1.61	0.64	-2.48	<.05	*
Noise 0SNR	-1.38	0.65	-2.11	<.05	*
Age	-0.25	0.27	-0.93	.35	
Trial No	-0.12	0.20	-0.61	.54	
ContrastVP V	0.24	0.40	0.61	.54	
Age:ContrastVP V	0.23	0.40	0.57	.57	
Age:Trial No	0.10	0.20	0.47	0.64	
	Low predictability items subset				
	Estimate	*SE*	*z*-value	*p*-value	
Intercept (quiet, P)	2.16	0.31	7.03	<.001	***
Noise -5SNR	-6.32	0.39	-16.09	<.001	***
Noise 0SNR	-4.87	0.32	-15.22	<.001	***
Age	-0.25	0.21	-1.18	.24	
Trial No	0.47	0.08	5.97	<.001	***
ContrastVP V	1.55	0.34	4.63	<.001	***
Age:ContrastVP V	-0.33	0.14	-2.29	<.05	*
Age:Trial No	-0.004	0.08	-0.06	.95	

Note. This table presents the analyses for the subsets of high and low predictability items. The response variable is the participants' answer type, distractor (0) or target (1).

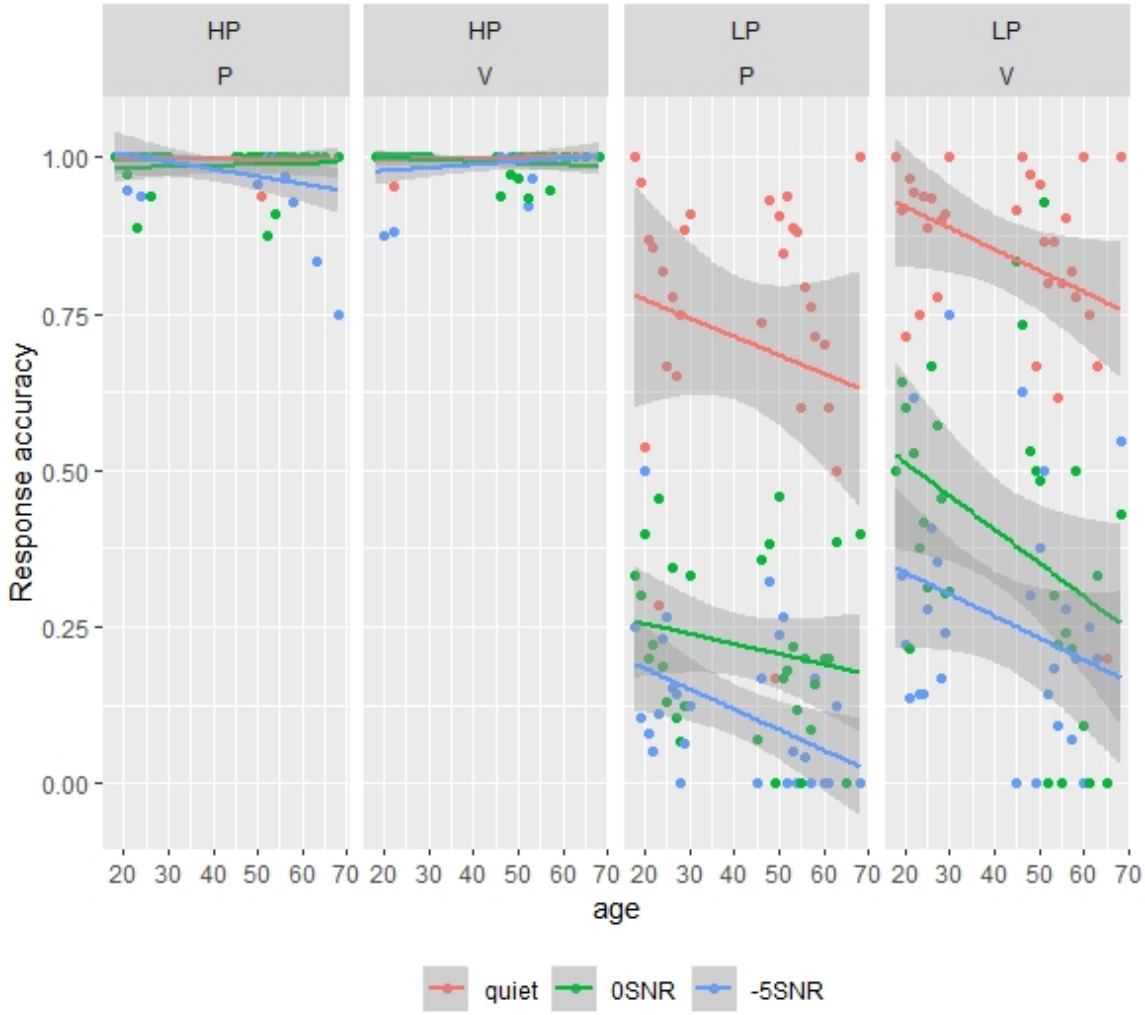

Figure 4.2: This figure shows the participants' answers; split for target and distractor items, with age plotted on the x-axis and answer type on the y-axis. Here 0 denotes the distractor response and 1 the target response. Different line colors show different noise conditions. The different plots show the high (HP) and low predictability (LP) items for stimuli differing in a plosive (P) or vowel (V).

pairs differing in the vowel had more target responses than items of minimal pairs differing in the plosive (β = 1.55, SE = 0.34, z = 4.63, $p < .001$). This was in line with the expectation that words differing in the plosive contrast would be harder to identify correctly than words differing in the vowel. Finally, the interaction of ContrastVP and Age was significant as well (β = -0.33, SE = 0.14, z = -2.29, $p <$ .01), showing that with increasing age, there was a larger decrease in the proportion of target responses for vowel contrasts than plosive contrasts. These effects are presented in Table 4.2 and illustrated in the two right-hand panels of Figure 2. The interaction effect in particular is shown by the steeper downwards slope of the lines in the LP Vowel plot compared to the LP plosive plot.

4.3.3 Semantic Fit and Phonetic Distance

We coded the semantic fit and phonetic distance to the target of the wrong responses, to investigate whether participants relied more on the acoustic signal (low distance) or on the provided context (wrong response fits semantically). This gives more insight in the participants' strategies and allows us to tease apart whether participants relied on top-down (predictions based on context) or bottom-up (acoustic signal) information. Figure 4.3 presents the normalized phonetic distance and semantic fit for the wrong responses in each of the three noise conditions. Lower normalized phonetic distance scores mean that the participant's response sounded more similar to the target word. Responses with a distance score of 1 were empty responses. Figure 4.3 also shows that in a majority of the wrong responses in each of the noise conditions, the participant's response did not fit the sentence semantically (76 vs 12 for Quiet; 177 vs 73 for 0 SNR; 341 vs 208 for -5 SNR). The peaks of the phonetic distance distributions seem to lie more to the right (meaning larger distance to the target) in the semantically fitting responses, suggesting a trade-off between acoustic fit and semantic fit. Participants made their response based on what they heard at a cost of conforming to the semantic context.

4.3.4 Confidence Ratings

We calculated the mean confidence for each of the three response types, namely targets (M = 3.494, SD = 0.806), distractors (M = 2.997, SD = 0.994), and wrong responses (M = 1.756, SD = 0.988), finding similar confidence for targets and distractors overall, and lower confidence for wrong responses. We transformed the participants' confidence responses to a binary variable of low confidence (confidence ratings 1 and 2)

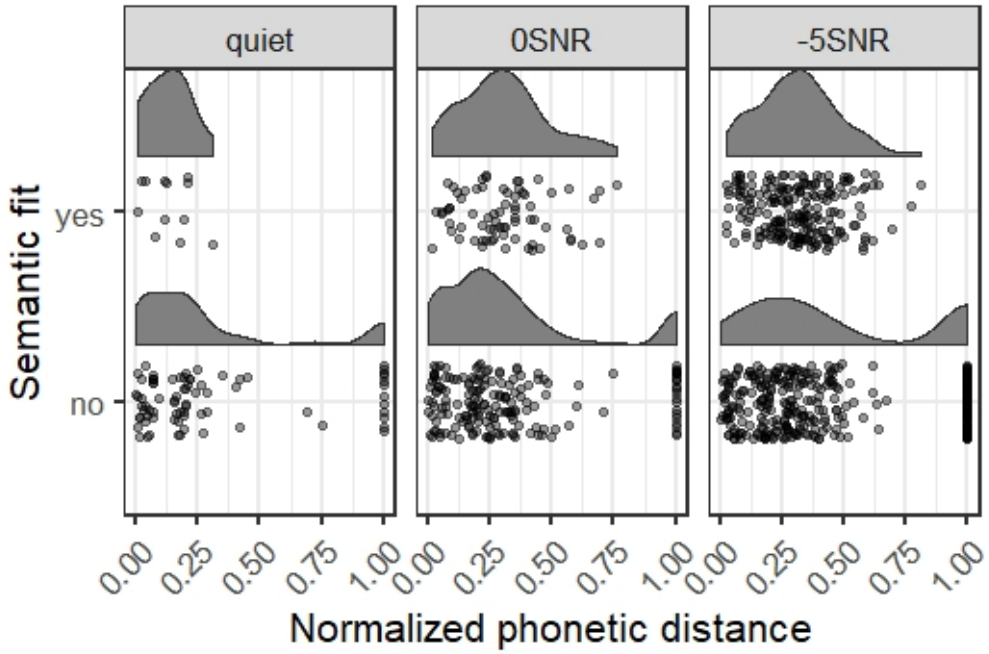

Figure 4.3: This figure shows the wrong responses that semantically fit or did not fit the sentence, plotted with the normalized phonetic distance, in each of the three noise conditions. Lower phonetic distance means more similar to the target item. A distance of 1 means an empty response.

and high confidence (confidence ratings 3 and 4). This binary response variable was tested using a GLMM with logistic linking function. Equivalent results are found with ordinal regression analyses (see Appendix B), but because of better interpretability, we present the binomial regression here. For these analyses, we have taken three subsets of the data: one with the target responses (N = 4161), one with the distractor responses (N = 1438), and one with the wrong responses (N = 881). We expected to find different patterns of confidence ratings for these subsets, because in the wrong responses, participants relied mostly on the sentence context, while in the distractor responses, there was some supporting evidence from the acoustic signal as well. As such, we expected participants to be more certain in general of their distractor items, than of their wrong items, as they realized that the wrong items were not presented to them in the speech signal. These analyses will shed light on how confident participants were in the different response types overall. Subsequently, we will turn to the distractor responses in the three noise conditions, as this was the condition most likely to elicit false hearing. Figure 4.4 presents the participants' confidence ratings from uncertain (1) to certain (4), split for each of the predictability conditions, noise levels, and response types.

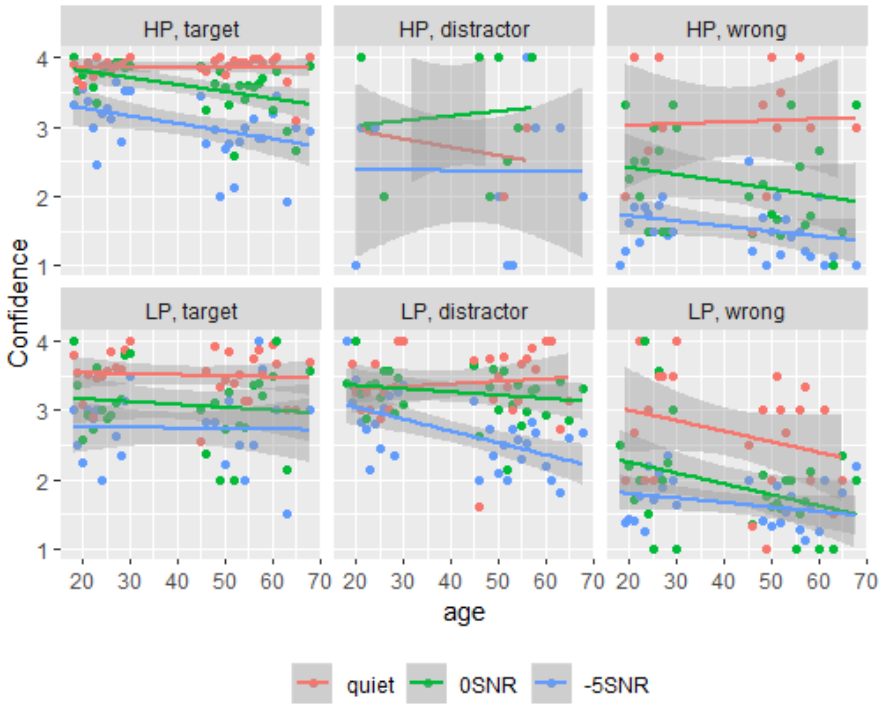

Figure 4.4: This figure shows the participants' confidence ratings; split for the predictability conditions, with HP at the top row and LP at the bottom, as well as the three answer types. Age is plotted on the x-axis and confidence on the y-axis. Here 1 denotes the lowest confidence and 4 the highest confidence. Different line colors show different noise conditions.

Table 4.3: Model Outcomes for the Confidence Rating Analysis (Target Items)

	Estimate	SE	z-value	p-value	
Intercept	5.65	0.48	11.69	<.001	***
Predictability LP	-2.17	0.51	-4.28	<.001	***
Noise -5SNR	-4.10	0.47	-8.78	<.001	***
Noise 0SNR	-1.71	0.46	-3.70	<.001	***
Age	0.15	0.31	0.48	.63	
Trial No	-0.02	0.07	-0.34	.73	
ContrastVP V	0.35	0.20	1.77	.08	.
Predictability LP : Noise -5SNR	0.76	0.55	1.38	.16	
Predictability LP : Noise 0 SNR	-0.77	0.49	-1.56	.12	
Predictability LP : Age	0.05	0.36	0.13	.90	
Noise -5SNR : Age	-0.85	0.34	-2.52	<.05	*
Noise 0SNR : Age	-1.07	0.36	-3.02	<.01	**
Predictability LP : Noise -5SNR : Age	0.42	0.44	0.97	.33	
Predictability LP : Noise 0SNR : Age	0.99	0.41	2.42	<.05	*

Note. This table shows the analysis for the subset of target items. The response variable is the participants' confidence (high or low).

The model for the subset of target responses included fixed effects of Predictability (categorical predictor with two levels using dummy coding, mapping the High Predictability condition on the intercept), Noise, Age, Trial Number, ContrastVP, as well as the three-way interaction of Predictability, Noise, and Age. All were coded and scaled as before. A by-Participant random intercept was included with random slopes for Noise and Predictability, and a by-Item random intercept with a random slope for Predictability. There was a significant effect of Predictability, with lower confidence in LP versus HP (β = -2.17, SE = 0.51, z = -4.28, $p < .001$). The model revealed lower confidence in Noise compared to Quiet (β = -1.71, SE = 0.46, z = -3.70, $p < .001$ for 0SNR, and β = -4.10, SE = 0.47, z = -8.78, $p < .001$ for -5SNR). The interaction of Noise and Age was significant, with lower confidence for older participants in noise (β = -1.07, SE = 0.36, z = -3.02, $p < .01$ for 0SNR, and β = -0.85, SE = 0.34, z = -2.52, $p < .05$ for -5SNR). Finally, the three-way interaction of Predictability, Noise, and Age was significant for the 0SNR condition, with higher confidence ratings with age in LP (β = 0.99, SE = 0.41, z = 2.42, $p < .05$). The other effects were not significant (all p-values $> .08$), all effects can be found in Table 4.3.

The model for the subset of distractor responses included the same fixed effects as the model on the subset of target responses. We removed non-significant interactions and only included the interaction of Noise and Age. A by-Participant random intercept was included, as well as a by-Item random intercept with a random slope of Predictability. Inclusion of other random slopes led to models with a singular fit. We find a significant effect of Noise for the -5SNR condition (β = -1.39, SE = 0.28, z = -4.99, $p < .001$), suggesting lower confidence for loud noise compared to quiet (the 0SNR condition did not differ significantly from quiet: $p = .89$). The model revealed a significant effect of Vowel/Plosive contrast (β = -0.45, SE = 0.20, z = -2.24, $p <$ .05), suggesting that participants were less confident about their answers on items that had a vowel contrast, rather than those with a plosive contrast. Additionally, there was a significant effect of Trial Number, where participants are less confident in later trials (β = -0.19, SE = 0.08, z = -2.52, $p < .05$). The interaction effect of Noise and Age is significant for both noise levels (0SNR: β = -0.64, SE = 0.27, z = -2.33, $p < .05$; -5SNR: β = -0.83, SE = 0.27, z = -3.08, $p < .01$). This suggests that older adults (with increasing age) show lower confidence in background noise compared to younger adults in quiet. The other effects were not significant (all p-values $> .11$). All effects can be seen in Table 4.4.

Table 4.4: Model Outcomes for the Confidence Rating Analysis (Distractor Items)

	Estimate	SE	z-value	p-value	
Intercept	1.73	0.79	2.18	<.05	*
Predictability LP	0.29	0.75	0.39	.70	
Noise -5SNR	-1.39	0.28	-4.99	<.001	***
Noise 0SNR	0.04	2.94	0.14	.89	
Age	0.42	0.27	1.59	.11	
Trial No	-0.19	0.08	-2.52	<.05	*
ContrastVP V	-0.45	0.20	-2.24	< .05	*
Noise -5SNR : Age	-0.83	0.27	-3.08	<.01	**
Noise 0SNR : Age	-0.64	0.27	-2.33	<.05	*

Note. This table shows the analysis for the subset of distractor items. The response variable is the participants' confidence (high or low).

Table 4.5: Model Outcomes for the Confidence Rating Analysis (Wrong Items)

	Estimate	SE	z-value	p-value	
Intercept	0.92	0.36	2.16	<.05	*
Predictability LP	-0.07	0.21	-0.34	.74	
Noise -5SNR	-2.86	0.35	-8.11	<.001	***
Noise 0SNR	-1.77	0.34	-5.16	<.001	***
Age	-0.17	0.13	-1.28	.20	
Trial No	-0.07	0.10	-0.68	.50	
ContrastVP V	-0.30	0.25	-1.20	.23	

Note. This table shows the analysis for the subset of wrong items. The response variable is the participants' confidence (high or low).

The model for the subset of wrong answer items included the same fixed effects as the previous two models, except that this model did not include any interaction effects. A by-Participant random intercept was included, as well as a by-Item random intercept. Inclusion of random slopes led to models with a singular fit. The model revealed a significant effect for both noise conditions. In 0SNR noise, participants were less confident than in quiet (β = -1.77, SE = 0.34, z = -5.16, $p < .001$), an effect that was also found for -5SNR noise (β = -2.86, SE = 0.35, z = -8.11, $p <$ .001). These findings show that generally, confidence ratings reflect the amount of noise that was presented. None of the other effects were significant (all p-values $>$ .20), and all effects are presented in Table 4.4.

Finally, we wanted to investigate directly the false hearing effect in the noise conditions, thus focusing on the confidence ratings in mishearings. We take subsets of the data of all distractor items produced in 0SNR (N = 646), -5SNR (N = 618) and quiet (N = 174). Based on previous findings, we expect to find a false hearing effect in the noise conditions, where participants show high confidence in their incorrect

Table 4.6: Model Outcomes for the False Hearing Analysis

	Quiet subset				
	Estimate	*SE*	*z*-value	*p*-value	
Intercept	0.95	1.77	0.54	.59	
Predictability LP	1.47	1.79	0.82	.41	
Age	0.50	0.35	1.43	.15	
Trial No	-0.36	0.29	-1.23	.22	
ContrastVP V	0.34	0.58	0.59	.55	
	0SNR subset				
	Estimate	*SE*	*z*-value	*p*-value	
Intercept	1.89	0.93	2.03	<.05	*
Predictability LP	0.36	0.92	0.40	.69	
Age	-0.17	0.15	-1.22	.22	
Trial No	-0.36	0.13	-2.85	<.01	**
ContrastVP V	-.91	0.28	-3.29	<.01	**
	-5SNR subset				
	Estimate	*SE*	*z*-value	*p*-value	
Intercept	0.29	0.84	0.35	.73	
Predictability LP	0.22	0.84	0.26	.79	
Age	-0.44	0.14	-2.99	<.01	**
Trial No	<0.001	0.11	0.01	.99	
ContrastVP V	-0.14	0.27	-0.53	.59	

Note. This table shows the analysis for the subset of distractor items in quiet, 0SNR, and -5SNR. The response variable is the participants' confidence (high or low).

responses as these distractor responses were supported by the sentence context. We expect to find an effect of age, so that older participants are more confident of their responses than younger adults. All outcomes from the three GLMMs are presented in Table 4.6.

The model on the subset of 0SNR trials included fixed effects of Predictability, Age, Trial Number, and ContrastVP (all coded and scaled as before). The model also included random intercepts for Subject and Item (random slopes led to non-convergence or singular fit). The model showed significantly lower confidence as the trials went on (β = -0.36, SE = 0.12, z = -2.85, $p < .01$). Additionally, confidence ratings were significantly lower for items with a Vowel contrast compared to items with a Plosive contrast (β = -0.91, SE = 0.28, z = -3.29, $p < .01$). The other effects were not significant (all p-values $> .22$).

The model on the subset of -5SNR trials consisted of the same fixed and random effects as the 0SNR model. We find only a significant effect of Age, where older participants are less confident of their responses than younger adults (β = -0.44, SE = 0.14, z = -2.99, $p < .01$). This is the opposite of what we would expect for false hearing based on previous findings (Failes et al., 2020; Failes & Sommers, 2022; Rogers et al., 2012; Rogers, 2017; Sommers et al., 2015), where older participants are more confident of their responses. None of the other effects were significant (all p-values > .59).

The model on the quiet subset of the data again included the same fixed and random effects as the previous two models. None of the effects were significant (all p-values > .15). These models together show no evidence for false hearing in our data, although mishearings were frequent.

4.4 Discussion

In the present study, we investigated word recognition in background noise in younger and older adults, analyzing to what extent listeners rely on the acoustic speech signal or on top-down predictions made based on the sentence context. In our experiment, participants typed in the last word of the sentence that was played in quiet or embedded in background noise at 0SNR and -5SNR. Additionally, participants rated their confidence in giving the correct answer.

We replicated the results of the experiment described in Chapter 3. The results showed that in quiet listening conditions, listeners of all ages and in both high- and low predictive contexts, mainly make use of the information in the acoustic speech signal. However, they turn more to the sentence context than the acoustic signal as a guide when there is some level of background noise. Regarding the effect of age, we found that this effect is stronger for older adults than for younger adults, and it is more pronounced in higher levels of background noise, in line with our hypotheses. Generally, we find that words with a vowel contrast are easier to recognize than words with a plosive contrast, a benefit that lessens with age, presumably due to floor effects. With regard to the confidence ratings, we generally find lower confidence ratings that reflect more difficult listening conditions and incorrect answers. Words with vowels get lower confidence ratings when the response is incorrect compared to items with a plosive contrast. In none of the conditions in our experiment do we find a false hearing effect where participants rate their incorrect responses with higher confidence, even though mishearings were very common.

4.4.1 Sound Contrast

We carefully controlled the phonetic contrasts of our minimal pairs to investigate how the sound difference of the minimal pair might have an effect on recognition scores. In this experiment, our pairs differed either in a plosive (place of articulation) or in a vowel (tense/lax). We expected that the items differing in the plosive were more difficult to recognize correctly than the items differing in a vowel. Plosives consist of a relatively short sound, especially compared to vowels that have a longer duration and greater amplitude. Thus, plosives are more likely to get lost in the noise, in which case the listener would make use of the provided sentence context and report having heard the distractor item. This expectation was confirmed by our data. Other studies that looked at a wider range of plosives and vowels also found that, especially in more difficult listening situations, vowels led to easier recognition than plosives (Cutler et al., 2004; Fu et al., 1998).

Our results showed an interaction with age: the facilitative effect of a vowel contrast over a plosive contrast decreased as participants were older. The direction of this interaction is unexpected at first glance, as we had hypothesized that older adults would have increased difficulty identifying plosives, as for these sounds the higher frequencies are more informative than for vowels (Alwan et al., 2011; Edwards, 1981). These high frequencies are lost first in age-related hearing loss (Gates & Mills, 2005). We believe however that the observed interaction is the result of a floor effect: older adults have a lot of trouble understanding the plosive correctly in noisy conditions, and almost always mistake the distractor for the target item in this condition. As there is already a substantial number of distractor responses for plosives even in the quiet condition, the decline in noise cannot be as steep as the one observed for vowels, for which comprehension is a lot better in quiet. Another possible explanation for the interaction effect is that the older adults might have had age-induced hearing loss that they were not aware of. This could have led to problems to, among other things, discriminate spectral transitions in noise (Tun et al., 2012). This difference in mishearing between plosives versus vowels suggests that even minor changes in how well the acoustic signal can be perceived affect the probability distribution of the bottom-up information and can lead to a more dominant top-down probability, as predicted by the Noisy Channel Model.

When looking at the confidence ratings, we find an effect of ContrastVP in the subset of distractor responses. This suggests that participants were less confident of their response if the target word was part of a minimal pair containing a vowel contrast, than when the word came from a pair with a plosive contrast. Most distractor

responses were made in the low predictability condition, where the sentence context supported the distractor word, while the acoustic information did not. We also found that in the low predictability condition, words from a pair differing in the vowel generally were easier to identify correctly (participants responding with the target word more often than the distractor). When participants responded incorrectly (with the distractor rather than the target), they were less confident of this, suggesting that they were more aware that they misheard the word than they were for plosive contrasts.

We did not choose our sound contrasts with any models of speech perception in mind. In hindsight, our contrasts might not all be processed in the same way. For example, studies suggest that the coronal place of articulation for consonants is not specified and that it can vary freely for coronal consonants (Friedrich et al., 2006; Lahiri & Reetz, 2010; Roberts et al., 2013). We used the coronal sounds /t/ and /d/ in our consonant minimal pairs, contrasted with other plosives differing in place of articulation. Testing whether these sounds led to more distractor responses due to unspecified coronal place of articulation is outside the scope of this chapter, but would be an interesting question for future research.

4.4.2 Bottom-up and Top-down Processes

This study investigated how bottom-up auditory processes and top-down predictive processes interact in speech comprehension, in particular in noisy conditions and while looking at differences between younger and older adults, following the predictions of the Noisy Channel Model (Levy, 2008; Levy et al., 2009; Shannon, 1949). In the high predictability condition of our experiment, we found an effect of noise, so that there were more distractor responses in the conditions with background noise compared to quiet. This effect was small, and most responses were in fact correct, suggesting a ceiling effect, in particular in quiet. In our paradigm, we presented the sentence context on the screen in written form, which would have led to these ceiling effects. Both the information provided by the speech signal and the information provided by the sentence context pointed to the target word. Participants could thus use information from both sources to recognize the correct word, there was no conflict between them. Especially in the quiet condition, there was no expectation that participants would identify the word incorrectly. The fact that we found this ceiling effect shows that our participants were paying attention to the task. The lack of an age effect in the high predictability condition regarding the number of distractor responses even in noise shows that older adults can make up for difficult listening

conditions by making use of the predictability of the message (Wingfield et al., 2005). As this is arguably the most frequent situation in normal language comprehension – i.e. words fit the context – this is a helpful strategy in everyday listening.

We found different results in the low predictability condition, where the participants' answers depended greatly on the condition the items were presented in. In the low predictability condition, the information provided by the acoustic signal is contradicted by the information given by the sentence context, as both point to different lexical items. On the one hand, the word supported by the context is also partially supported by the speech signal. Because we used minimal pairs, these two words only differed in one single phonetic feature. On the other hand, the word supported by the information from the speech signal is not supported by the sentence context at all. In the quiet condition, participants identified the sentence-final word for the most part correctly. In conditions with background noise, however, participants do rely more on the sentence context to guide word recognition, as shown by the shift to a large proportion of distractor answers. The increased rates of mishearing in noise are observed for both younger and older adults, but the effect is substantially stronger for older adults. This is in line with previous work that has shown that older adults tend to rely more heavily on the sentence context (Dubno et al., 2000; Hutchinson, 1989; Pichora-Fuller et al., 1995; Sommers & Danielson, 1999). Due to the presence of noise, it is more difficult to identify all the sounds in the speech signal, and here listeners turn to the other source of information they have available. This was an expected finding, as in previous studies, also younger adults do rely more on context when listening conditions get harder (Dubno et al., 2000; Hutchinson, 1989; Pichora-Fuller, 2008). We also observed a significant learning effect in our data: as the trials in a block proceed, participants are slightly more likely to get the target item correct. This holds for participants irrespective of age. One possible explanation for this is that they became aware of the manipulation and the fact that the context could be misleading, thus paying more attention to the sound signal than they did before. Listeners have been found to be able to re-weight cues based on their statistical properties (Bushong & Jaeger, 2019). It also shows that older adults are able to adapt to the task, unlike in Rogers et al. (2012). In the present study, they learned over the course of the experiment that context might be misleading and weighing the acoustic information more than the top-down predictions. Adaption suggests that older participants are behaving rationally when showing mishearing.

Analyses of semantic fit and phonetic distance to the target word show that the majority of the wrong responses did not fit the sentence semantically, while dis-

tances were smaller in the semantically incongruent responses. This suggests that participants did try to rely on the acoustic signal rather than the provided context, somewhat against our expectations. It might be the case that they had noticed the sometimes misleading sentence context and relied less on this information. Even though we already find high rates of mishearing in our study, it is likely that this underestimates the amount of mishearing that would occur for these materials in a more naturalistic setting. Participants became aware of the possible semantic mismatches in the presented audio and sentence context, and our analyses show that participants in fact paid considerable attention to the acoustic signal rather than the sentence context.

According to the Noisy Channel Model (Levy, 2008; Levy et al., 2009; Shannon, 1949), information from both sources is combined rationally. However, older adults have been found to rely more on top-down predictive processes than younger adults, which can lead to mishearing in cases when the target is not predicted by the context. A study by Gibson et al. (2013) showed that human language processing relies on rational statistical inference in a noisy channel. Their model predicts that semantic cues should point the interpretation in the direction of plausible meanings even when the observed utterance differs from this meaning, that these non-literal interpretations increase in noisier communicative situations, and decrease when the semantically anomalous meanings are more likely to be communicated. The findings from the present study are in line with the predictions based on the model by Gibson et al.: In more adverse listening conditions, i.e. the conditions with more background noise, listeners rely more on the sentence context to compensate for the difficulties introduced in auditory processing. In these cases, listeners respond that they heard a word that fits the sentence context (plausible meaning), rather than the word that was actually presented to them (implausible meaning). There is contextual information, as well as some sensory information (the shared sounds of the presented word, as these words form a minimal pair) to support the word favored by the sentence context. Finally, following Gibson et al.'s last prediction, over the course of the experiment participants noticed that the sentence context is not always reliable, and showed a learning effect. They came to expect low predictability sentence-final items, which led to less mishearing.

Rationally combining bottom-up and top-down information in speech comprehension is sensible, in particular in cases of a noisy channel, where the bottom-up signal is partially obscured. However, when the top-down predictions form a mismatch with the information being transferred in the signal, a too strong reliance on

top-down processes can lead to problems in communication, in the form of mishearing. These are a side effect of rationally combining bottom-up and top-down information.

4.4.3 False Hearing

We also tested the replicability of the false hearing effect in German, that was reported for English in previous literature (Failes et al., 2020; Failes & Sommers, 2022; Rogers et al., 2012; Rogers, 2017; Sommers et al., 2015). This effect generally has been found to be stronger for older adults than younger adults. Unlike previous studies and against our expectations, we do not find an age effect for false hearing in our study. While there was a substantial amount of mishearing, older participants were not more confident about their responses than younger participants. We also do not find an effect of age on confidence in distractor responses overall. While Rogers et al. (2012) do report a smaller false hearing effect in the condition with loud noise compared to the condition with moderate noise, they do still find a false hearing effect. In the present study, we do not find a significant effect of age at all for the 0 SNR subset, while in -5 SNR the effect is opposite to our expectations: with age, participants become less confident. One possible explanation for this failure to replicate the false hearing effect in noise is the age of the participants: the participants in previous studies were generally older than those in the present study, and thus perhaps more likely to show the false hearing effect due to age-related cognitive declines on top of the effects of mishearing predicted by the noisy channel model. Instead of false hearing, we find that our participants' confidence ratings reflect the difficulty of the listening condition: they tended to be lower in noisy conditions and in low predictability sentences.

4.4.4 Limitations

One of the limitations of this study is that, due to collecting the data via the web, we were not able to collect hearing thresholds of our participants, nor were we able to carefully control the sound levels at which the stimuli were presented. We excluded older participants with a large number of incorrect responses in quiet, so that we make sure that the performance in that condition was equated to younger adults. In hindsight, there is another option for controlling hearing levels among our participants. We could have used an alternative control condition where no context cues are available. These stimuli could have been filler sentences in which participants could only rely on the speech signal to make their response. In this way, auditory

performance could be equated among our groups of younger and older adults. Peelle et al. (2016) showed that for intelligibility ratings, online testing is a feasible method to replace laboratory testing as it gave comparable results to testing in the lab. This suggests that careful control of participants' listening conditions and software used like in lab settings is not necessary to obtain reliable results. Additionally, previous studies have equated overall audibility for older and younger adults using individual speech recognition thresholds, and still found larger false hearing effects for older adults, suggesting it is not directly caused by differences in hearing acuity (Failes et al., 2020; Failes & Sommers, 2022; Rogers et al., 2012; Rogers, 2017; Sommers et al., 2015).

We constructed the items in our low predictability condition by swapping the two words from the minimal pairs we had selected. It should be noted that this led to sentences that, while unpredictable, also were implausible. In fact, in the low predictability condition, the sentences provided a context that was strongly biased for the distractor word. This could have led to larger amounts of mishearing compared to when we would have used sentences that were unpredictable but plausible, in particular for older adults who tend to rely more on context. Due to the strong bias for the distractor and the implausibility of the target word, relying on the context would strongly favor the distractor response. In this way, the sentences can be seen as misleading. Other studies investigating false hearing using sentences varied in whether their low predictability items were plausible or not. Sommers et al. (2015) used unpredictable sentences that were still meaningful (LP: The shepherd watched his sheath), but Failes and Sommers (2022) and Failes et al. (2020) had implausible items. They constructed their unpredictable items by changing one phoneme in the sentence-final target word in the predictable item (HP: She put the toys in the box; LP: She put the toys in the fox). All three of these studies found a larger false hearing effect for older adults, and therefore this effect seems to be independent of the plausibility of the low predictability items. It therefore seems unlikely that our lack of an effect can be explained by having used implausible sentences. The false hearing effect has also been found using a word priming paradigm (Rogers, 2017; Rogers et al., 2012), which suggests that the effect does not depend on the use of a particular paradigm.

Another limitation of the present study is the age of our older adults, which is relatively young. Our oldest participant was 68 years old, and the mean age of the older group was 53. Compare this to the ages of the older participants in Failes et al. (2020), which ranged from 65 to 81, with a mean of 71 years. This might explain the

lack of an age-related false hearing effect in the present study. For our sample, we find that simple mishearing caused by rational processes better explains our results of differences between vowel contrasts and plosive contrasts. However, it could be the case that in an older sample, general cognitive decline plays a larger part as well (Rogers et al., 2012).

The present results are based on a restricted set of minimal pairs, namely pairs of plosives only differing in place of articulation, and tense versus lax vowels, and were tested in multi-speaker babble noise. More research is needed to investigate how these findings generalize to other sound combinations and other types of noise. Future studies could also test at different SNRs, to prevent in particular the floor effects we found in the plosives as noise, as this can shed light on the true nature of the interaction effect of age and sound contrasts in noise. Currently, the Noisy Channel Model does not incorporate meta-cognitive measures like confidence ratings. Confidence could be formulated in terms of the probability distribution between different lexical candidates. If, on the one hand, the probability of one candidate is a lot higher than that of another candidate, high confidence in the response should be reported. On the other hand, if the probabilities of different candidates are more similar, the confidence rating should be lower. The exact modeling of false hearing based on confidence ratings in the Noisy Channel Model can be explored in future research.

4.5 Summary

Previous studies, including the previous chapter of this dissertation, have investigated the mishearing effect, where listeners understand a word different from the one that was spoken. These effects are particularly prevalent in situations where the speech signal is noisy, and the word that is actually understood fits well with the semantic context, indicating that top-down predictability of the word may have overpowered the bottom-up auditory signal. Previously, this effect has been attributed to general deficits in cognitive control, in particular inhibition (Failes et al., 2020; Failes & Sommers, 2022; Rogers, 2017; Rogers et al., 2012; Sommers et al., 2015). In these studies, older participants showed larger effects of both mishearing and even false hearing.

In the present experiment, we argued that the mishearing effect is a natural consequence of rational language processing in noise, and thus does not require to be attributed to deficits in cognitive control. To test this idea, we designed a study

which carefully controls the way in which the target and the distractor words differ from one another. Specifically, we constructed target-distractor pairs which only differed in the articulatory position in a plosive, and another set of target-distractor pairs that differed only in vowel quality. We conducted an online study in German, in which participants listened to sentences in quiet and two levels of background babble noise, and reported the sentence-final word they heard, as well as rated their confidence in this response. Our findings show that participants accurately report the actually spoken word in quiet listening conditions, but that they rely more on sentence context in the presence of background noise, leading to incorrect responses in particular in the low predictability condition. While listeners thus do profit from high predictability in noise (as they do correctly understand the words in this condition), they also suffered the downside of mishearing in the low predictability condition. The mishearing effect was found to be larger in older adults compared to younger adults, replicating previous findings. We explain this within the Noisy Channel Model in terms of increased language experience of older adults, possibly compounded by first experiences of hearing loss.

For our critical phonetic manipulation, we found that stimuli pairs with a vowel contrast were generally easier to identify correctly than pairs with a plosive contrast, although this benefit lessened with age. These different effects for vowels versus plosives suggest that mishearing depends on the quality of the acoustic signal, rather than general deficits in cognitive control or inhibition. We also find a learning effect that suggests that participants of all ages were able to adapt to the task. We think that this finding also underscores the rational account, and is not consistent with an account that relates age differences to a difference in cognitive control. Our findings also add to the literature by replicating the earlier mishearing effects in a different language, German.

Earlier work had however also reported an effect of false hearing, meaning that participants are very confident of their answer even though it is in fact incorrect (Failes et al., 2020; Failes & Sommers, 2022; Rogers, 2017; Rogers et al., 2012; Sommers et al., 2015). In particular, the false hearing effect was found to be increased in older adults. While our experiment was also set up to assess false hearing, we did not find any significant effects of false hearing in the older participants compared to the younger ones. Instead, confidence levels were related to the level of background noise and the difficulty of listening in general.

This experiment, as well as that reported in Chapter 3 investigated how both an (un)predictive context and the presence of background noise affect the speech

comprehension process. We focused on *what* participants heard, and how they used the different sources of information available to them. While this is an interesting and important question in its own right, it raises a follow-up question: How do predictability and background noise affect higher-level processes in communication, beyond the mere recognition of what is being said? In Chapter 5 we will focus on the effects on memory, conducting an experiment with a new set of stimuli. Building on the results of the first two experiments, we investigate how auditorily presented sentences with high or low surprisal affect subsequent memory in a surprise memory test.

Chapter 5

Exp. 3: Surprisal and False Memory

In the previous two chapters we have investigated how the presence of background noise and a predictable or unpredictable context affect language comprehension. In the experiments, we asked participants to type in the word they heard in presented recordings of sentences. However, in every-day life, we do not ask people what they think we just told them, instead we want them to do something with the information we passed on. Generally, we want them to respond to what we said, with an utterance of their own or with an action. Language is primarily used for communication, which involves a wider set of cognitive processes than those used during just the understanding of spoken words. The open question is: How do noise and predictability affect these higher-level processes needed for communication, beyond their effect on speech comprehension? In other words, what are the *consequences* of listening in background noise? We will focus on the process of memory, which is important in many situations, for example in noisy surroundings where instructions need to be given to people who might feel they already know (parts of) the directions, like in traffic or on factory floors. These people might increasingly rely on predictive processes to make up for the adverse listening condition and free up resources for their working memory. In these cases it is good to understand how higher level processes and memory are affected so that proper understanding and recollection can be ensured.

In this chapter we will report on the findings of an experiment that aimed to address these questions. In particular, we tried to find an answer to the question how predictability and noise affect recognition memory. Previous work has found that words that are predicted but not presented linger in memory and affect memory

performance at a later time point (Haeuser & Kray, 2022a; Hubbard et al., 2019), where participants show so called *false memory* effects for these items. We investigated whether these false memories are increased in background noise, as in the two previous chapters we saw that listeners rely more on predictive processes when listening in background noise. For predictability, we now use the measure of surprisal, unlike in Chapter 3 and Chapter 4, where we used cloze probabilities to determine our high and low predictability items. By changing the word order in the sentence, we manipulate whether the target word occurs in the beginning or the end of the sentence, thus also varying its surprisal. When the target word occurs at the end of the sentence, surprisal is lower than when it occurs at the beginning of the sentence, as the preceding sentence context gave clues as to what the word could be. This surprisal manipulation has as a benefit that the content of the sentence between the high and low surprisal items is the same, and all items are plausible. This is unlike the cloze probability manipulation used in the previous experiments, where the low predictability condition by definition differed in meaning from the high predictability condition, and where low predictability items in some cases were implausible.

In the following sections we will summarize the literature relevant to this chapter, which has been extensively covered in Chapter 2. It concerns the effects of surprisal on recognition memory, and the phenomenon of false memory (Section 5.1). In Section 5.1.2 our research questions and hypotheses regarding the experiment in this chapter will be outlined, while Section 5.2 provides details on the method that we used, including the construction of our stimuli. The results are presented in Section 5.3, and discussed in Section 5.4. Finally, a summary of the chapter is given in Section 5.5. This experiment has been preregistered here: https://aspredicted.org/B33_7SY.

5.1 Introduction

To manipulate the difficulty of the listening condition, we vary the noise condition in which an item is presented, as well as the predictability of the target word. The presence of background noise increases the difficulty of the task for the listener. Items occurring with low surprisal (thus high predictability), on the other hand, decrease difficulty, in particular compared to items with high surprisal. These listening conditions should thus lead to varying reliance on predictive processes, which we expect to affect the participants' recognition memory.

Recognition memory is one's ability to (correctly) recognize that something has been experienced before. It consists of two sub-components, namely familiarity and

recollection (Yonelinas et al., 2010). The present chapter focuses on recollection, which is the ability to remember discrete details such as the source of the information or temporal details (Coane et al., 2011; Yonelinas et al., 2010). Recognition memory is often tested after a study phase by presenting participants with previously presented items as well as unseen items, and asking them to judge whether they have seen the item in the study phase. The results are then presented in terms of correct hits (responding with 'old' when the item was old) and false alarms (responding with 'old' when the item was unseen).

The literature shows conflicting results for the effect of predictability on memory. While some studies report better memory for unpredictable words (Haeuser & Kray, 2021; Höltje et al., 2019; Perry & Wingfield, 1994; Riggs et al., 1993; Staresina et al., 2009), other studies found instead that unpredictable words are remembered better (Corley et al., 2007; Federmeier et al., 2007; Rommers & Federmeier, 2018). These opposing findings are possibly due to differences in plausibility in the items used (Haeuser & Kray, 2022b; Kuperberg, 2021; Kuperberg et al., 2020). In the current experiment, we used word order to vary the surprisal of the target word, thus keeping the sentence-level plausibility the same across predictability conditions and eliminating this confound.

We aimed to test for false memories, which are items that participants remember, but that were never presented. The effect depends on the strength of the predictive processes during listening and encoding. Recent studies found false memory effects using read sentences (Haeuser & Kray, 2022a; Hubbard et al., 2019). This paradigm finds these false memories for words that are predicted by a constraining sentence, but not actually presented to participants. In the present experiment, we used a similar paradigm. As shown in Chapters 3 and 4, noisy listening conditions lead to a stronger reliance on predictive processes during speech comprehension. We expected to find effects of the listening condition on the false alarms in our recognition memory test, with more false alarms for non-presented semantically related words ("lures"), in particular in items where participants were more likely to rely on prediction. Finally, we varied the word frequency of the target items, investigating the established Word Frequency Mirror Effect (see below).

5.1.1 Word Frequency

High-frequency words are processed more efficiently than low-frequency words (Brysbaert et al., 2018; Monsell et al., 1989). They are known by more people and processed

faster in several tasks such as word naming and lexical decision. Additionally, word frequency affects memory processes, with different effects depending on the task. When participants are asked to freely recall previously seen words, low-frequency words are harder to recall than high-frequency words. But when participants are asked to discriminate them from non-presented lure items in a recognition task, performance is better for low-frequency words than high-frequency words (Gregg, 1976; Yonelinas, 2002). This recognition effect has been named the Word Frequency Mirror Effect (Glanzer & Adams, 1990). In experiments, the hit rates for old items are higher for low-frequency words, while there is an equally large effect of higher false alarm rates for high-frequency words. In other words, performance is better for low-frequency words compared to high-frequency words, both when they are old (larger hit rates) and new (smaller false alarm rates). This effect stems from several processes: a slower recollective process, a faster familiarity process, and an assessment of change in the relative familiarity of a word (Coane et al., 2011; Mandler, 1980). High-frequency words have a relatively high baseline level of familiarity. Presentation of a high-frequency item in for example an experiment, will not have a large effect on its familiarity level; it is already high. Therefore, high-frequent, unseen distractors are more likely to be incorrectly classified as old. On the other hand, low-frequency items have a low baseline level of familiarity before exposure during an experiment, and this exposure then leads to a larger change in the familiarity level, making it easier to correctly recognize old items and correctly classify unseen as new items (Bridger et al., 2014; Coane et al., 2011). In this work, we further aimed to investigate how word frequency interacts with the noise and predictability conditions, testing participants' recognition memory.

Frequency of occurrence is usually measured in text corpora. In the present experiment, we obtained our frequency values from the CELEX database (Baayen et al., 1995). Common ways of measuring frequency are as frequency per million words, or a logarithmic frequency, which reduces the differences between high frequency words while maintaining a difference between low-frequency words. This is the value that we used.

5.1.2 Research Goals and Hypotheses

In this experiment, we aimed to answer the question how listening to sentences with different predictability levels and background noise levels affects subsequent recognition memory. We were particularly interested in *false memory*, which previous studies have found for predicted but not presented words. The presence of background noise

should make listeners more likely to rely on predictive processes that can lead to false memories. Addressing these questions will provide new insights in how background noise and prediction affect higher-level processes, beyond mere word recognition. This way, the results will speak to how communication is affected by these factors.

In a surprise recognition memory test administered after an exposure phase, participants were presented with three types of words: old items, which were target items from the sentences that were presented to the participants; new items, which were not presented and unrelated to the sentences; and finally items semantically related to the presented target words, which could be predicted from the sentence context, but crucially not presented during the listening phase. We will refer to these items as (semantic) lures. False alarms on these semantic lures would point to false memory effects, where the participant remembers hearing a word that, in fact, was not presented.

Using this surprise recognition memory test, we can probe participants' episodic memory of the sentences they listened to, and compare downstream effects as measured by memory accuracy depending on the surprisal of the target word and the presence of background noise. Previous studies found conflicting results for the effect of predictability on memory (Corley et al., 2007; Federmeier et al., 2007; Haeuser & Kray, 2021; Höltje et al., 2019; Perry & Wingfield, 1994; Riggs et al., 1993; Rommers & Federmeier, 2018; Staresina et al., 2009). These conflicting effects might be due to differences in the plausibility of the target word, where unpredictable words are remembered better when violating plausibility (Haeuser & Kray, 2022b; Kuperberg, 2021; Kuperberg et al., 2020). In our stimuli we varied surprisal based on changes in word order, thus keeping plausibility of the item constant across predictability conditions. We were able to investigate the effect of predictability on memory without the confound of plausibility of the sentence contents. Given that all of our sentences were plausible, we expected better recognition, and thus higher accuracy, for highly predictable target words. This would mean higher hit rates and a lower amount of false alarms in the sentence-final condition.

We expected that these effects are related to integration of the target word when it occurs in sentence-initial position (= high surprisal). In this condition, due to the high surprisal of the target word, it might initially be hard to process. The meaning of the target word might therefore be filled in afterwards based on the rest of the sentence context. This might result in impoverished memory for the target word that was actually presented, but instead only for the general meaning of the word in the sentence. As the semantic lure and target word both would fit the sentence,

we expected lower accuracy due to a lower hit rate and larger false alarm rate in the sentence-initial condition than the sentence-final condition. In the sentence-final condition the preceding sentence context would have helped narrow down the possible candidates for the target word, thus leading to easier processing of the target and better memory of the exact word. To tease apart this effect depending on the meaning of the target word from the overall surprisal effects, we can contrast the false alarm rate of the lures and new items. If the false alarm rate for the lures is higher than that for the unrelated new items, it supports the idea that participants relied on the meaning of the sentence to make their judgement. Alternatively, it might also be the case that the sentence context is not constraining enough to eliminate the false alarm effects on the semantic lures when the target word occurs in sentence-final position, as the lures are related to the target word. This would mean that we find higher false alarm rates for semantic lures than for unrelated new items, regardless of surprisal condition.

To both manipulate the difficulty of the task and increase reliance on predictive processes during listening, we presented our items in either background noise or in quiet. The presence of noise divides the listener's attention during encoding, which is disruptive. This makes the condition with background noise more difficult than the quiet condition. Therefore, in background noise we expected lower accuracy on the memory test compared to quiet. This effect acted as a control of the experimental paradigm: not finding an effect of noise would imply that the experiment was not set up correctly.

Importantly, we expected an interaction of background noise and surprisal. The previous experiments in this dissertation, as well as the literature at large, have shown that in background noise participants increasingly rely on prediction to aid their comprehension. This would primarily be helpful in sentences with low surprisal. Therefore, we might find a larger false alarm rate for the semantic lures in background noise, as participants rely on predictive processes to cope with the noisy listening conditions. As the lures fit the sentence and could have been predicted by participants, it is more difficult to classify them as new, unseen items.

As mentioned above, previous studies have found that low frequency items are recognized correctly more often in memory tests than high frequency words, and that there are more false alarms for high frequency lures than low frequency lures (Word Frequency Mirror Effect; Bridger et al., 2014; Glanzer & Adams, 1990; Yonelinas, 2002). This is thus an effect we expected to find for frequency in general. We might even find a three-way interaction of surprisal, noise, and word frequency, where the

effect of surprisal in background noise might be larger for the low frequency items than the high frequency items. Because low frequency words are harder to process, relying on predictive processes would aid the listener. This would mean that the amount of false alarms increases for lure items related to the low frequency items. Thus, memory accuracy would reflect the item's processing difficulty.

5.2 Method

5.2.1 Participants

We recruited a total of 155 participants via the recruitment platform Prolific (prolific.co). Data from four participants was excluded due to a missing vocabulary task. One participant's data was excluded because they had performed on chance in the memory test, and another participant's data was excluded due to poor performance on the comprehension questions (5 or more incorrect responses out of 20). One participant was excluded for both these reasons (chance performance in the memory test and 5 or more incorrect filler items). The remaining 148 participants completed all parts of the experiment. They were all native speakers of German, aged between 19 and 30 years old ($M = 25$ years). Seventy-six participants were female, seventy-one male, while one participant preferred not to say. One participant reported having a reading/writing disability, and one participant reported having tinnitus. All participants gave informed consent, and the study was approved by the Deutsche Gesellschaft für Sprachwissenschaft (DGfS). They were paid €8 for their participation.

5.2.2 Materials

Our stimuli in the present experiment consisted of 120 pairs of similar-sounding high and low frequency words. We constructed sentences for these items where the target word occurred either at the beginning or the end of the sentence. All items were embedded in -5 dB SNR babble noise.

In the recognition memory test we gave participants a written word and asked them to judge whether they had heard this word in the presented sentences, yes or no, and also rate their confidence on a four-point scale. These memory items could be one of three types: old items, semantic lures, and unrelated new items. We tested the participants' accuracy on this memory task. A detailed description of the stimuli construction can be found in the sections below.

Selection of Items

We started out with a set of 120 target words, which were divided up in similar sounding word pairs (N = 60) and constructed with the help of native German-speaking research assistants. One word of the pair was high frequent compared to the other one, which was low frequent. Unlike in the stimuli for Experiment 1 and 2, we did not have strict criteria for how similar sounding the items had to be and where the differing sounds should occur in the word. We did have some minimal pairs (for example *Jambus* - *Bambus*; iamb - bamboo), but most items were more different from each other (for example *Pergament* - *Parlament*; parchment - parliament, or *Kontradiktion* - *Fiktion*; contradiction - fiction). As the frequency of the words was set relative to the other member of the pair, there was a lot of variation within the set of high frequency words, and there might have been instances where the low frequent word of a pair had a higher frequency than the high frequent word of another pair. We looked up the frequency of the words in the CELEX lexical database (Baayen et al., 1995), taking the logarithmic frequency of the word. The set of sixty low frequency words has a mean log frequency value of 0.11 (SD = 0.28; range 0 - 1.18), the set of sixty high frequency words has a mean log frequency value of 0.76 (SD = 0.78; range 0 - 2.72). It must be noted that a number of words did not occur in the CELEX database, leading to a frequency value of 0, possibly skewing the results and leading to the large standard deviations.

In a pretest, we asked participants whether they were familiar with the word and could give a definition. Participants saw the words presented on the screen one by one and typed their definition in a text box and rated their familiarity. The rating scale ranged from 0 to 6 and had the following descriptions (translated from German):

0. I have never seen or heard the word.
1. I recognize that it is a German word, but I cannot explain it.
2. Barely used; I think that [the participant's given definition] is what it means, but I am not sure.
3. I know the word, but I don't use it.
4. I know the word and use it occasionally.
5. I know the word and I use it.
6. Very well-known word.

We had four experimental lists, each consisting of 38 words. Each list included three non-words, so that we could check participants rated these words as unknown. Approximately half (17 or 18; counterbalanced across lists) of the items were high

frequent words, the other half were low frequent words. Each list was completed by ten participants who would not participate in the main experiment.

We added the question probing for the definition to check whether participants actually had the correct definition of the words, and were not making their rating based on incorrect knowledge. To check this, we randomly selected a subset (N = 98 out of 392) of all low frequency items that had a rating of 3 or higher and coded whether the participants' definition was correct or not. As only 11 out of 98 items had an incorrect definition, and all these items had the lower rating of 3, we assumed we could rely on the participants' rating data in general.

In the rating data, the pretest showed that the high frequency words were indeed better known than the low frequency words, with the non-words not known at all. The average rating for the low frequency items was 2.5 (SD = 1.77), while the average rating for the high frequency items was 4.8 (SD = 1.23). Non-words had a mean rating of 0.3 (SD = 0.7). The smoothed density plots of the ratings are presented in Figure 5.1.

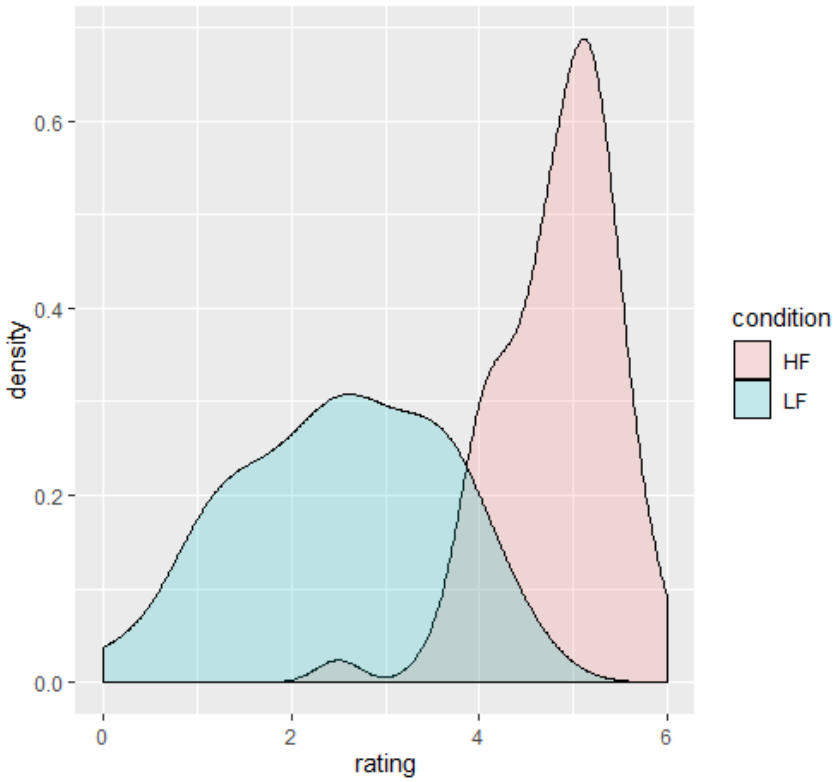

Figure 5.1: This figure shows the participants' familiarity ratings for the high frequency (HF) and low frequency (LF) items. Higher ratings indicate higher familiarity.

Construction of Sentences

For each of the target words we made a sentence context (N = 120) so that the target was predictable. To manipulate surprisal, we changed the word order of the sentence. In one of the conditions, the target word occurred in sentence-initial position (with some articles or other words before it to make a grammatical sentence), while in the other condition the target word occurred in sentence-final position (again, not in all cases the word was absolutely the final word of the sentence to keep the sentence grammatical). This difference in the word order manipulated the surprisal of the target word. The two versions of the sentence were identical except for their word order and any changes to make the new order grammatical.

For each of the sentences, we estimated the target word's surprisal value. This was done using a GPT-2 language model (Greenberg, 2022). Overall, the set of items with the target word in sentence-initial (SI) position had higher surprisal values than the set of items with the target word in sentence-final (SF) position. In a few individual cases, the SF surprisal value was higher than SI surprisal, and several rounds of rewriting and retesting the sentences did not improve this. In the final set, the high surprisal condition, the mean surprisal for the sentences was 12.84 (SD = 3.44; range 5.70 - 29.34), while in the low surprisal condition the mean surprisal value was 7.50 (SD = 3.49; range 0.44 - 15.80).

Besides the experimental sentences, we constructed 41 filler items. These would be used during the experiment as attention check items and also had comprehension questions. We chose not to have questions on experimental items to not point participants' attention to some of these items. In these filler sentences, we tried to vary to word order from only the canonical order to make sure it would blend in with the varied word order in the experimental items. The comprehension questions were simple yes/no questions that participants would not be able to answer without paying attention to the sentence. Approximately half the questions had 'yes' responses, the other half 'no' responses. All items as well as filler items are listed in Appendix C.

Recordings and Preprocessing

Recordings were made of all the sentences, in both order types, as well as of all filler items. These recordings were made in a sound-treated booth by a female native speaker of German. Instructions for the speaker included to speak at a natural, clear speech rate and take care not to make any mispronunciations. Sentences were recorded twice so that the clearest version could be selected later.

Table 5.1: Accuracy on pretest for 0 SNR and -5 SNR

	-5 SNR	0 SNR
Correct	562 (62%)	806 (90%)
Incorrect	338 (38%)	94 (10%)
Total	900	900

All sentences were segmented from the entire recording, taking care to have as little silence before and after as possible. All sound files were normalized to 65 dB, and 300 ms of silence were added before and after the recording. Then noisy versions were created by mixing a café babble noise file (BBC Sound Effects Library, Crowds: Interior, Dinner-Dance, http://bbcsfx.acropolis.org.uk/) at -5 dB SNR, so that the mean intensity of the noise was five decibels more than that of the speech signal. This was calculated based on the average intensity of the entire sentence (as opposed to just the intensity of the target word like in the previous experiments).

We chose this level of noise based on a pretest. We selected thirty items, half of which had the word order where the target occurred in a sentence-initial position, while in the other half the target word occurred at the end of the sentence. Sixty participants (32 male, 19-34 years old, M=25) completed the experiment, in which they were asked to write down the entire sentence they had heard. Half of the participants were presented with the sentences embedded in babble noise at 0 dB SNR, while the other half of participants heard the more difficult -5 dB SNR noise. Items played automatically, and could only be listened to once. We scored accuracy based on whether the target word was part of the participants' transcription. The results are shown in Table 5.1. For the more difficult condition, with noise of -5 SNR, we find that about two-thirds of the responses were correct (586 out of 900), whereas in the easier condition with noise of 0 SNR, about 90% (806 out of 900) responses were correct. We decided that using noise of 0 SNR in the main experiment would be too easy, and possibly lead to no difference between items heard in noise or in quiet. The 60% accuracy reached in -5 SNR seems to be a good balance between challenging, but not too difficult that participants cannot understand the utterance at all. Therefore, we used a noise level of -5 db SNR in the main experiment. None of the participants who completed the pretest took part in the main experiment.

Memory test items

The experiment consisted of two parts. In the first part, participants listened to sentences in both quiet and noisy conditions. In the second part, they were presented

with a surprise memory task, where they saw previously seen target words, related semantic lures, or unrelated unseen new items. These items were presented in written form, because the memory traces of the auditorily presented items might affect the results. In the listening experiment, only target words are presented, which thus have an auditory memory trace, while the other items do not.

For the semantic lures, we selected words that were semantically related to each of the target words. It was important that these words were able to take the target's place in the sentence, and thus be valid lures. We ensured that the lures did not occur in the sentence itself. Controlling this meant we were unable to carefully control the frequency of the lures, and this thus differed from the experimental target items. The semantic lures for the set of low frequency items had a mean log frequency value of 0.48 (SD = 0.64; range 0 - 2.56), which is higher than that of the low frequency target words. The semantic lures for the set of high frequency items had a mean log frequency value of 0.63 (SD = 0.68; range 0 - 2.50), which is a little lower than that of the high frequency target words themselves. For the previously mentioned target words, we had the following lures: *Jambus* - *Versfuß* (iamb - metrical foot), *Bambus* - *Eukalyptus* (bamboo, eucalyptus), *Pergament* - *Schriftrolle* (parchment - scroll), Parlament - Bundestag (parliament - house of representatives), *Kontradiktion* - *Widerspruch* (contradiction - contradiction), *Fiktion* - *Einbildung* (fiction - imagination).

We also constructed unseen, new items for the memory test. For these items, participants should not hesitate to answer 'new' as they had not seen either the exact item, or related words before. We took care to select words that did not occur in any of the experimental items or filler items. The final set of new items consisted of 35 words with a mean log frequency value of 0.59 (SD = 0.96; range 0 - 4.771).

5.2.3 Design

Our total set of 120 experimental items was divided over three study lists of forty items each. This was done to on the one hand prevent items on a single list priming each other or their semantic lures, and on the other hand to make sure a single list would not be too long. We made sure that the subset of forty items did not include both items from the initial high- and low frequency pair. Besides the forty experimental study items, each list included twenty filler items, for a total of sixty items. The memory test list consisted of sixty items as well: twenty old target items, twenty semantic lures, and twenty unrelated new items. Each experimental list started with

a short practice session of four items to make participants familiar with the listening paradigm and the task.

For each of the three unique experimental lists, we constructed different versions in order to counterbalance the different conditions in the experiment. Variables that were counterbalanced were whether an item was presented in background noise or in quiet, whether it was the sentence-initial word order or the sentence-final word order, and for the memory items whether the word had originally been presented in noise or quiet. In a single list, all conditions (noise vs. quiet, SI vs SF) occurred an equal number of times. We additionally counterbalanced the order of the lists, making versions with the order of items reversed. This resulted in 48 different experimental lists.

The items in each list were pseudo-randomized, where we made sure that no more than three items in a row were either quiet or presented in noise, and no more than one filler item consecutively. In the memory test list we similarly ensured no more than three items of a single type were presented after each other. We checked that for each list the CELEX frequency (Baayen et al., 1995) and length in syllables across the old items, the semantic lures, and the new items did not differ significantly. This indeed wasn't the case (all ps $> .06$).

5.2.4 Procedure

The experiment was hosted on Lingoturk, a crowd sourcing client (Pusse et al., 2016). Participants were asked to complete the experiment in a quiet environment on a computer and using the Chrome web browser for best results. They could use either headphones or speakers, and set their volume at a comfortable listening level at the start of the experiment.

The experiment consisted of three different parts; (1) the listening phase, where participants were exposed to the experimental sentence recordings; (2) the memory phase, in which participants completed a surprise memory test, and in between a phase (3) where we tested different individual differences measures. All instructions (in German) for the tasks are provided in Appendix D. In the listening phase, participants would listen to each of the experimental sentences in turn while seeing a fixation cross in the middle of the screen. Then they were asked to rate the intelligibility of the words in the sentence on a five-point scale. This task gave the participants something to do beyond passive listening, which might lead to low attention and high drop-out rates in an online experiment. To make sure that participants did not only listen to

the presence or absence of background noise to make their response to the intelligibility question, but paid attention to the content of the sentence, we included twenty filler items with comprehension questions. These were closed questions, answerable with "yes" or "no".

In the memory phase of the experiment, participants saw written words presented on the screen one by one. For each word, they were asked whether they had heard this exact word in the earlier sentence recordings, yes or no, and additionally to rate their confidence in giving the correct response on a four-point scale, ranging from very unsure to absolutely certain. As described above, the items in the memory phase could be one of three types, *old* words that had been presented before, *semantic lures* similar to the presented target words, or *new* words, unrelated to any of the seen items. The participants' accuracy on this memory task was our main outcome variable of interest.

The experiment ended with a short questionnaire. This questionnaire included questions on demographic information (age, sex, education, current job, possible impairments), linguistic information (type of German, second languages), and on the experiment itself (what the participant thinks it was about, whether they used a strategy, whether they completed the experiment in a quiet environment, and whether they did any other actions simultaneously like texting or watching television). Completing the entire experiment took about 40 to 45 minutes.

Individual Differences Measures

We conducted three different tests to measure individual differences that are of importance in our experiment and which are treated as covariates in our analyses. Their respective relevance will be explained for each test. Two of the three tests were conducted between the listening phase and the memory phase to distract the participants from the heard sentences before the surprise memory tests started.

After the listening phase, we administered a 12-item version[1] of the Raven's Progressive Matrices as a measure of non-verbal IQ (Bilker et al., 2012; Raven, 2000). In this test participants saw a design with a missing part, and had to select the missing part from six to eight patterned response options. The difficulty of the designs differed across the questions. IQ relates to memory performance, and thus collecting this data

[1]This version was based on the 9-item version as designed by Bilker et al. (2012), with three additional items to prevent floor- and ceiling effects. This version has been found to have good reliability (Scholman et al., submitted).

as a covariate may be important to explain variance in results on our recognition memory test.

Subsequently, we asked participants to complete a backwards digit span test. In this task, participants are presented with a string of three to eight digits, which they have to type in in reverse order after having seen all digits. The longer the string of digits, the harder the task is. This is both a measure of working memory and has been used to measure *phonological buffer* (or *phonological loop*; Baddeley, 1968; Baddeley, 1996; Baddeley, 2003; Colle & Welsh, 1976; Olsthoorn et al., 2014; Salame & Baddeley, 1982), which is the loop of generally two seconds in which auditory information is still available to the listener. We expected that this might relate to the memory performance of auditory stimuli.

A third individual differences measure, vocabulary size, was presented after the memory phase, to make sure the items in the vocabulary test would not prime the items in the memory test. As some of our stimuli are quite infrequent words in German, we measured the participants' vocabulary size as a control variable. Larger vocabulary size would mean it is more likely that the participant knows the (low frequent) words they listen to, which might affect recognition performance. It is difficult to find a standardized vocabulary test that is aimed at healthy, monolingual adults, rather than for example children, people with impairments, or second language learners. This is the case for English, but also for German. The vocabulary test we selected, is a subset of the short vocabulary tests designed by the LMU Munich (Engel & Ettinger, 1997; Heubeck, 2001)[2], which has three difficulty levels (easy, medium, and hard). We took the set of middle difficulty and added ten items of both the easy and hard levels to prevent floor- and ceiling effects. This way, our final vocabulary test list consisted of 55 items of varying difficulty.

5.2.5 Analyses

The outcome variable for the Raven IQ test is the number of correctly solved questions, that of the vocabulary test is the number of correct responses, while the digit span score is the largest span which participants remembered correctly. For the three tests we also calculated the accuracy in terms of percentage correct.

We coded the accuracy of the memory test items into a binary variable (correct = 1, incorrect = 0), and then combined this with the outcomes of the individual dif-

[2]Access the original vocabulary test here: https://psytest.psy.med.uni-muenchen.de/kurztests/index.htm

ferences measures and the conditions of the listening phase (like noise and surprisal). These factors were used as predictors in our statistical models.

5.3 Results

In this chapter we present an experiment that aimed to investigate how listening in background noise to sentences with varying predictability levels affects memory performance, and in particular how it affects false memory. False memory has been found for words that are predicted from a sentence context but not actually presented to participants (Haeuser & Kray, 2022a; Hubbard et al., 2019). We expected to find increased false memory in listening conditions where participants were forced to rely more on predictive processes, such as in background noise, see Chapters 3 and 4. In Sections 5.3.1 and 5.3.2, we will report how participants performed on the various sub tasks in the experiment, such as the listening phase and the individual differences measures. In Section 5.3.3 we will analyze the results from the memory task to investigate whether the tested factors affected accuracy on the task and which conditions lead to increased false memory.

5.3.1 Listening Performance

To give participants a task to do during the listening phase, we asked them to rate how well they were able to understand all the words in the sentence, after each sentence that they heard. Participants made their rating on a five-point scale, where 5 indicated that they understood all words, and 1 indicated that they understood none of the words. We expected to find that in background noise, ratings would be lower than in quiet, as the condition with noise leads to more difficult listening conditions, in which words will be missed or identified with increased uncertainty. These findings form a sort of sanity check to control the noise manipulation. Finding no differences would mean that the noisy condition was too easy.

Results showed that in the quiet condition, participants more often used the highest rating. On the other hand, in the noisy condition, ratings were much lower. Thus, this is in line with our expectations, and shows that participants were following instructions and paying attention. All of the participants' ratings are presented in Figure 5.2.

We additionally analysed the participants' ratings by word order, to investigate whether there was a (subjective) difference in how well participants could understand

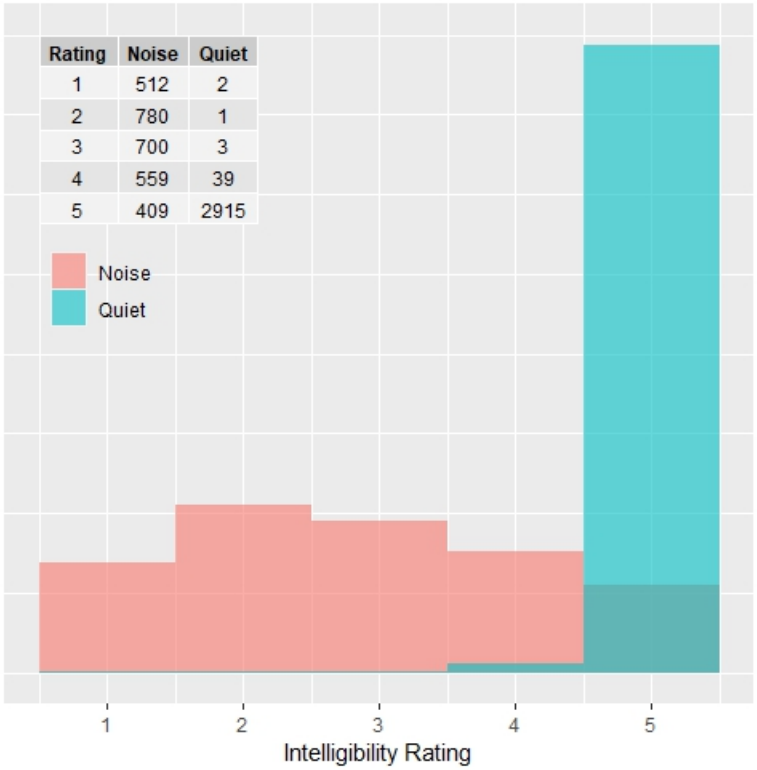

Figure 5.2: This figure shows the participants' ratings on the listening items in Quiet and Noise. Higher ratings mean higher intelligibility.

the different versions of our sentences. As can be seen in Figure 5.3, there are no differences between the two word orders.

5.3.2 Individual Differences Tests

Besides the experimental listening and memory tasks, participants also completed a Raven IQ test, a vocabulary test, and a backwards digit span test. The results from these tests are used as control predictors in the statistical models testing memory performance below. For clarity, the results are described here.

The Raven IQ test consisted of twelve items of varying difficulty. Participants had a mean number of 8.61 (SD = 1.77; range 4-12) questions correct. For the digit span, participants saw a total of 14 sequences of increasing length (ranging from 2 to 8 digits), with each sequence length being presented twice. A common metric to score the digit span is the longest sequence of which participants get both items correct. Our digit span score had a mean of 4.85 (SD = 1.52; range 0-8[3]). The vocabulary

[3]Some participants (N=5) did not respond correctly to both trials with length 2, giving the low digit span score of 0. Further analysis of these participants' individual differences tasks showed that

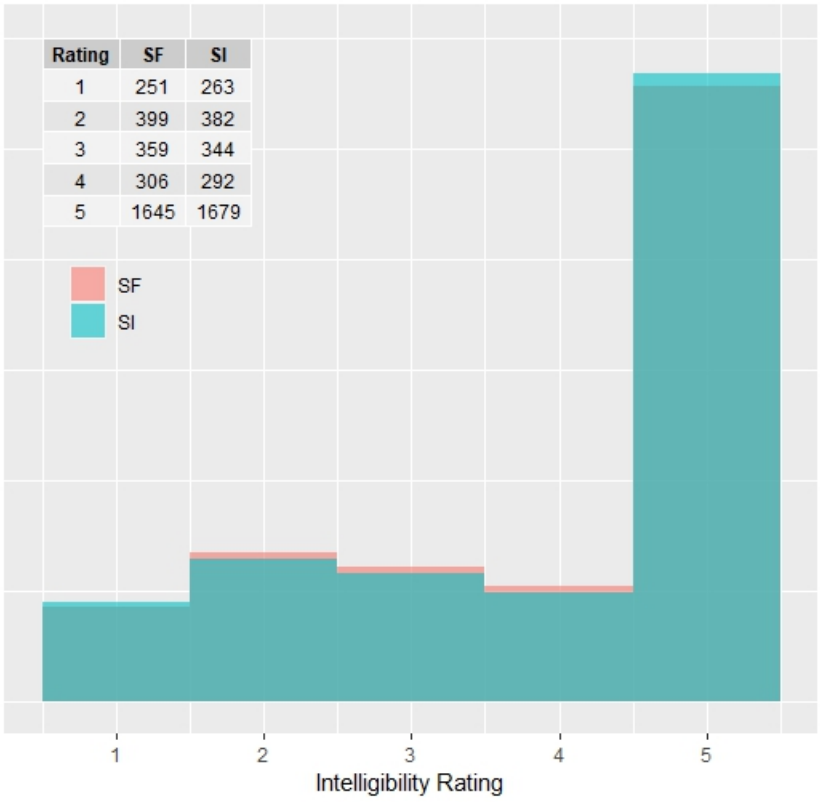

Figure 5.3: This figure shows the participants' ratings on the listening items with sentence-initial (SI) and sentence-final (SF) word order. Higher ratings mean higher intelligibility.

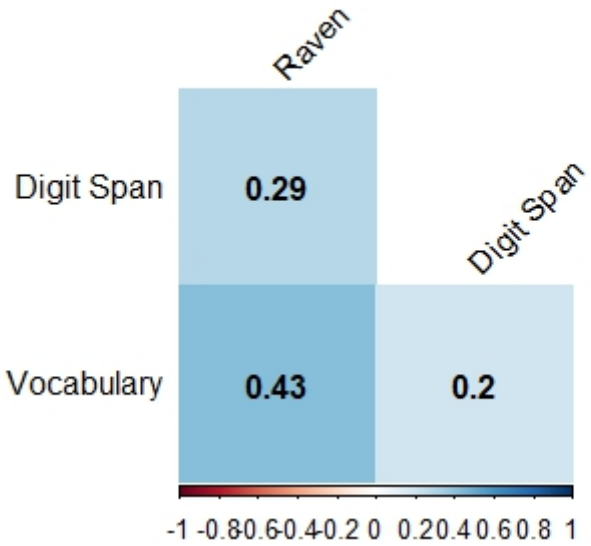

Figure 5.4: This figure shows the correlations between the three individual measures tests (Raven IQ, Vocabulary, and backwards Digit Span test).

test consisted of 55 multiple choice questions. Participants had a mean score of 31.42 correct answers (SD = 7.13; range 13-47).

We correlated the three individual measures with each other. Raven IQ and digit span were weakly correlated ($r(146) = .29$, $p < .001$). Similarly, digit span and the vocabulary test were weakly correlated ($r(146) = .20$, $p < .001$). The correlation between the vocabulary test and the Raven IQ test was a little higher ($r(146) = .43$, $p < .001$). These correlations are presented in Figure 5.4.

5.3.3 Memory performance

In this section we analyze the results from the memory task. Here we will answer the question whether more difficult listening conditions in which listeners rely more on predictive processes (like background noise), lead to increased false memory. This would be indicated by a larger amount of false alarms for our semantic lure items. In our analyses, we use accuracy as the response variable. For our three memory condition types, old items, semantic lures, and new items, we checked whether participants answered the question of having been presented with this word in the heard sentences

two performed relatively badly on all three tasks. We tested our statistical models on a subset of the data with these participants excluded, but the main results did not change. Therefore, we report below the models with these participants included.

correctly ("yes" for the old items, "no" for the semantic lures, and "no" for the new items).

In our statistical analyses we used generalized linear mixed models (GLMMs), implemented in the lme4 package (Bates et al., 2014) in R (R Core Team, 2022). These models allow both fixed and random effects, letting us control for variation on the participant- and item-level (Baayen et al., 2008; Barr et al., 2013). To improve convergence, all models were run using the bobyqa optimizer and increased iterations to 2·105. Model comparisons were made to guide model selection based on the Akaike Information Criterion (AIC), models with the lowest AIC are reported below. We reduced the models' random structure until they converged and did no longer warn of singularity.

The models presented here differ from those in our preregistration. First, it was not possible to compare accuracy performance modulated by surprisal and noise on the old items, the semantic lure items, and the new items directly. The new, unrelated memory items did not have associated surprisal or noise values. To overcome this, we ran two separate analyses: one comparing the new items and semantic lures, disregarding surprisal and noise as predictors; and one comparing the semantic lures and the old items. In this latter analysis, we did add surprisal and noise as predictors. As foreseen in the preregistration, we had to reduce the random effects structure to solve issues of non-convergence and singularity in the models. We did not find all the predicted effects in the preregistered model. We started with the preregistered model, and subsequently reduced the model by removing the interaction terms that did not reach significance. Other (unexpected) effects came out as significant, and we ran exploratory follow up models to investigate these effects further.

Model of Semantic Lures and New Items

To check whether a semantic relation to the experimental items in the listening phase would affect memory performance, we first compared the semantic lures and new items. Neither semantic lures nor new items were actually presented to participants in the listening part of the experiment, but the semantic lures were related to the presented target words. All new items were semantically unrelated. We expected that the semantic relation between the target sentences and the lures would lead to confusions in the memory test, resulting in lower accuracy for semantic lures compared to new items. To test this prediction, we compared the accuracy on the memory test for the semantic lures and new items, both numerically and statistically. For the lures and new items, lower accuracy corresponds to a higher false alarm rate. As

Table 5.2: Accuracy on memory test for semantic lures and new items.

	Semantic Lure	New
Correct	2046 (69%)	2527 (85%)
Incorrect	914 (31%)	433 (15%)
Total	2960	2960

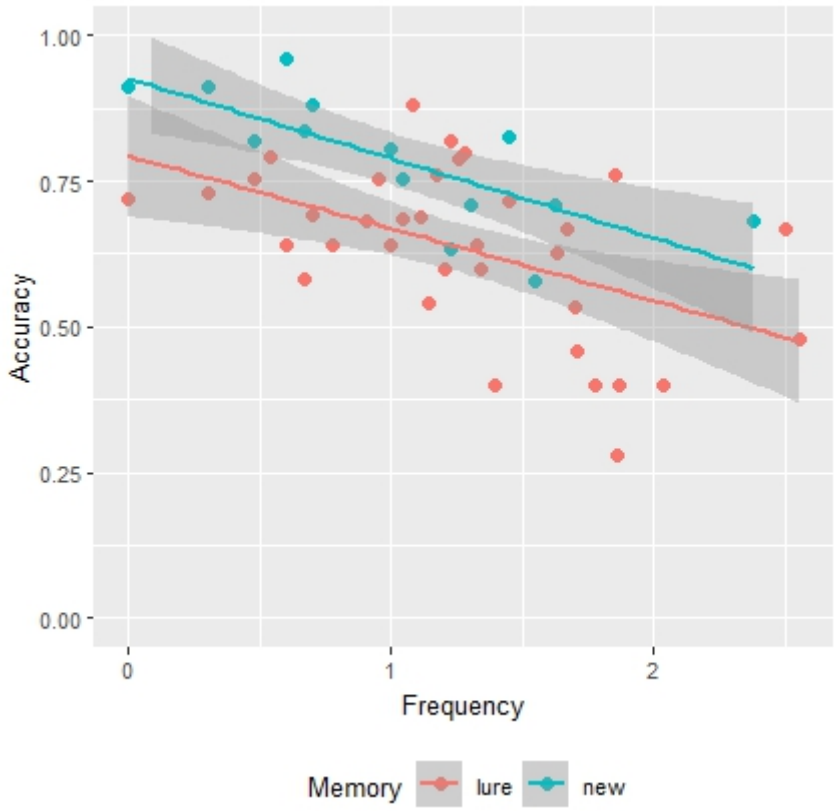

Figure 5.5: This figure shows the accuracy on the memory test by word frequency, split for memory condition (semantic lures and new items)

can be seen in Table 5.3.3, this expectation is supported by the numeric data. We find 69% correct responses for the semantic lures, but 85% correct responses for the new items. Additionally, we expected to find that accuracy was lower for words with higher frequency, as predicted by the established Word Frequency Mirror Effect (Bridger et al., 2014; Glanzer & Adams, 1990; Yonelinas, 2002). Accuracy has been plotted for both memory condition types and word frequency in Figure 5.5, and seems to be in line with our predictions.

We tested this observation in a generalized linear mixed model with logit function (GLMM). Our dependent variable was *Accuracy*, coded as 0 (incorrect) and 1 (correct). The included predictors were *Memory Condition* and *Frequency*. Memory condition coded whether the item was a semantic lure or a new item. These were

Table 5.3: Model Outcomes for Semantic Lure vs New items.

	Estimate	*SE*	*Z*-value	*p*-value	
Intercept	1.99	0.11	17.59	<.001	***
Memory Condition	-0.80	0.10	-8.05	<.001	***
Frequency	-0.76	0.10	-7.50	<.001	***
Memory Condition : Frequency	0.30	0.10	2.97	<.01	**

Note. This table presents the analysis for the Semantic Lures and the New items from the memory test. The response variable is the participants' accuracy, incorrect (0) or correct (1).

sum coded (new = -1, lure = 1). Frequency was the log-transformed word frequency per million as taken from CELEX (Baayen et al., 1995). The model also included the interaction of Memory Condition and Frequency, random intercepts for Item and Participants, and random slopes for Memory Condition for both intercepts.

The model showed a significant effect of Memory Condition (β = -0.80, SE = 0.10, z = -8.05, $p < .001$), showing that accuracy was lower for lures than for new items. The effect of Frequency was also significant (β = -0.76, SE = 0.10, z = -7.50, $p < .001$), so that words with higher frequency showed lower accuracy. These findings were in line with our predictions regarding the effects of semantic interference the Word Frequency Mirror Effect. Finally (and unexpectedly), the interaction effect was significant as well (β = 0.30, SE = 0.10, z = 2.97, $p < .01$), indicating that the adverse effect of higher frequencies was less strong for lures than for new items. The model results are presented in Table 5.3.

Model of Semantic Lures and Old Items

We now turn to the question whether the listening condition and surprisal of the target word affect false memory. Incorrect responses to semantic lures (answering 'yes, this is old') signal false memory: the participant believes they have heard the word when they did not. We expected that this depends on the strength of predictive processing in the listener, and thus that false memory will be more frequent in noisy listening situations. Additionally, we study the effects of word frequency and the three tested individual differences measures. In our analyses, we compared the accuracy on the memory test for the old items and the semantic lures. As the lures were associated with the old items, they had a corresponding noise condition and surprisal value (which were the same as for the associated old item). We used a step-down approach for our models, reducing non-significant interactions, and report here the model with the lowest AIC score in model comparisons.

We expected to find effects of Noise and Frequency, which acted as control effects. Following previous findings in the literature, we expected accuracy to be lower in background noise, as this makes processing harder. We also expected lower accuracy for words with higher frequency, following the Word Frequency Mirror Effect. Regarding Surprisal, we predicted that there would be lower accuracy on the memory test for items with higher surprisal, as these items are harder to process. As participants might remember the meaning of the word (or fill it in based on the subsequent sentence context), there could also be lower accuracy for the semantic lures, as the false alarm rate increases for these words that fit the heard sentences. We expected this effect of Surprisal to interact with Noise, so that the accuracy for high surprisal items in background noise is even lower. Due to the difficulty of processing these items (high surprisal *and* noise), participants might strongly rely on the sentence context rather than the exact word that was presented. This would lead to a higher false alarm rate on the semantic lures. On the other hand, the effect of Noise might be stronger than that of Surprisal: in background noise participants struggle to understand the sentence at all, leading to low accuracy regarding of surprisal condition.

In the model that we preregistered, the dependent variable was accuracy on the memory task, which was a binary variable coded as 0 (incorrect) and 1 (correct). The model included the three-way interaction of Surprisal (numeric variable, scaled), Noise (binary variable, sum-coded: Quiet = -1, Noise = 1), and Frequency (numeric variable, scaled). Other fixed effects were Memory Condition (binary variable, sum-coded: Old = -1, Lure = 1), Trial Number (numeric variable, scaled), and the three individual differences measures: Raven IQ test, Backwards Digit Span, and Vocabulary test (all scaled). The random effects structure had to be reduced from the preregistration, so that the final model consisted of random intercepts for Item and Participant, with random slopes of Noise for both, and additionally a random slope of Trial Number for Item.

The model revealed an expected effect of Noise (β = -0.25, SE = 0.04, z = -5.84, $p < .001$), with lower accuracy in noise than quiet. The effect of Frequency was significant as well (β = -0.23, SE = 0.06, z = -3.69, $p < .001$), showing lower accuracy for words with higher frequency, in line with the Word Frequency Mirror Effect. Accuracy was higher for semantic lures than old items, as shown by the effect of Memory (β = 0.21, SE = 0.07, z = 3.13, $p < .01$). Finally, higher vocabulary scores led to higher accuracy on the memory test (β = 0.08, SE = 0.04, z = 2.00, $p < .05$). Other effects, while marginally significant in some cases, did not reach significance (all ps $> .06$). The model's results are presented in Table 5.4

Table 5.4: Model Outcomes for the Preregistered Model.

	Estimate	*SE*	*Z*-value	*p*-value	
Intercept	0.85	0.07	12.63	<.001	***
Surprisal	0.04	0.04	1.05	.29	
Noise	-0.25	0.04	-5.84	<.001	***
Frequency	-0.23	0.06	-3.69	<.001	***
Raven IQ	0.07	0.04	1.72	.09	.
Digit Span	0.02	0.04	0.60	.55	
Vocabulary Test	0.08	0.04	2.00	<.05	*
Memory	0.21	0.07	3.13	<.01	**
Trial Number	-0.04	0.03	-1.23	.22	
Surprisal : Noise	-0.05	0.04	-1.52	.13	
Surprisal : Frequency	-0.03	0.04	-0.72	.47	
Noise : Frequency	0.08	0.04	1.90	.06	.
Surprisal : Frequency : Noise	0.06	0.04	1.65	.10	.

In our next analyses, we removed the interactions that were not significant. In our exploratory analyses, we found that adding Word Order improved the model fit, and we thus include this in the model. We present here the model with the lowest AIC score and with only significant interaction terms. Like before, our dependent variable is accuracy on the memory task. The model included the following fixed effects: Surprisal, Noise, Frequency, Word Order (binary variable, sum-coded: Sentence-Final = -1, Sentence-Initial = 1), Memory Condition, Trial Number, and the three individual differences measures: Raven IQ test, Backwards Digit Span, and Vocabulary test. All these variables were coded and scaled as described above. The model additionally included interactions of Noise and Surprisal, Noise and Word Order, Noise and Memory, and Memory and Word Order. The random effect structure was the same as in the model above, with random intercepts for Item and Participant, with random slopes of Noise for both, and additionally a random slope of Trial Number for Item.

We found a significant effect of Noise (β = -0.25, SE = 0.04, z = -6.84, $p <$ .001), showing that accuracy on the memory test was lower for the condition with background noise compared to quiet. In line with the Word Frequency Mirror Effect, we found a significant effect of Frequency (β = -0.22, SE = 0.06, z = -3.55, $p <$ .001), so that accuracy is lower for words with higher frequency. There was a significant effect of Vocabulary (β = 0.09, SE = 0.04, z = 2.03, $p <$ .5), showing that participants with higher vocabulary scores perform better on the memory test. A significant effect of Memory Condition (β = 0.13, SE = 0.06, z = 2.20, $p <$ .5) showed that overall, accuracy was higher for semantic lures than for old items. However, we also found two interaction effects with Memory Condition: The interactions with Noise (β = 0.29,

Table 5.5: Model Outcomes for Semantic Lure vs Old items.

	Estimate	*SE*	*Z*-value	*p*-value	
Intercept	0.86	0.07	12.77	<.001	***
Surprisal	-0.00	0.06	-0.02	.98	
Noise	-0.25	0.04	-6.84	<.001	***
Frequency	-0.22	0.06	-3.55	<.001	***
Word Order	0.05	0.05	1.06	.29	
Raven IQ	0.07	0.04	1.69	.09	.
Digit Span	0.02	0.04	0.52	.60	
Vocabulary	0.09	0.04	2.03	<.05	*
Memory	0.13	0.06	2.20	<.05	*
Trial Number	-0.04	0.03	-1.05	.29	
Noise : Surprisal	0.04	-3.52	-3.52	<.001	***
Noise : Word Order	0.10	0.04	2.38	<.05	*
Memory : Noise	0.29	0.03	8.22	<.001	***
Memory : Word Order	-0.14	0.03	-4.38	<.001	***

SE = 0.03, z = 8.22, $p < .001$) and Word Order (β = -0.14, SE = 0.03, z = -4.38, $p < .001$) showed that these predictors have different effects on the different memory conditions. We will have a closer look at these interactions in the next section. The model finally showed significant interactions of both Surprisal and Word Order with Noise; in opposite directions. The interaction of Surprisal and Noise (β = -0.15, SE = 0.04, z = -3.52, $p < .001$) showed lower accuracy for words with higher surprisal, while the interaction of Word Order and Noise (β = 0.10, SE = 0.04, z = 2.38, $p <$ .05) showed higher accuracy for items in Sentence-Initial condition. Other effects (of Surprisal, Raven, Digit Span, and Trial Number) were not significant (ps > .09). All effects are presented in Table 5.5.

Model of Semantic Lures

To investigate these interactions with Memory Condition, we will now turn to subsets of the data: one for the semantic lure items (2960 observations), and one for the old items (2960 observations).

Regarding the semantic lures, we are interested whether the presence of background noise leads to more false alarms, which would signal false memory. We expected these effects to be stronger in background noise compared to the quiet listening condition as here participants would rely more on predictive processes, thus activating the semantic lure items and leading to incorrect responses in the memory test. For the statistical model of lures, neither the interaction of Noise and Surprisal nor

Table 5.6: Model Outcomes for the subset of Semantic Lure items.

	Estimate	*SE*	*Z*-value	*p*-value	
Intercept	1.03	0.10	10.69	<.001	***
Surprisal	0.07	0.08	0.83	0.41	
Word Order	-0.11	0.07	-1.64	0.10	
Noise	0.06	0.04	1.40	0.16	
Frequency	-0.22	0.09	-2.35	<.05	*
Raven IQ	-0.00	0.08	-0.02	0.98	
Digit Span	0.07	0.08	0.97	0.33	
Vocabulary	-0.04	0.08	-0.45	0.65	
Trial Number	0.03	0.04	0.60	0.55	

of Noise and Word Order were significant and were removed from the model. Like before, the dependent variable was Accuracy (a binary variable coded as 0 = incorrect and 1 = correct). For the semantic lures, accuracy directly corresponded to false alarms: lower accuracy means more false alarms. The final model included the following predictors: Surprisal (numerical variable, scaled), Word Order (binary variable, sum-coded: Sentence-Final = -1, Sentence-Initial = 1), Noise (binary variable, sum-coded: Quiet = -1, Noise = 1), Frequency (numeric variable, scaled), Trial Number (numeric variable, scaled), and the three individual differences measures: Raven IQ test, Backwards Digit Span, and Vocabulary test (all scaled numerical variables). The random effects structure included random intercepts for Item and Participant, with a random slope for Frequency by Item, and a random slope of Word Order by Participant. Inclusion of other random slopes led to a singular fit of the model.

The model revealed only a significant effect of Frequency (β = -0.22, SE = 0.09, z = -2.35, $p < .05$), showing that words with higher frequency had lower accuracy in the memory test. This is in line with our predictions and the Word Mirror Frequency Effect. All other effects were not significant (ps > .10). All effects can be found in Table 5.6.

Model of Old Items

We were able to test for the sentence-level effects like the presence of background noise, target word surprisal and word order directly in the old items, as these were the only set of items in the memory test that were actually presented in these different listening conditions. The model for the subset of old items included effects of Surprisal, Frequency, Raven IQ, Digit Span, Vocabulary, and Trial Number (all scaled numerical variables). It also included Word Order (binary variable, sum-coded: Sentence-Final

Table 5.7: Model Outcomes for the subset of Old items

	Estimate	*SE*	*Z*-value	*p*-value	
Intercept	0.83	0.12	6.70	<.001	***
Surprisal	-0.13	0.12	-1.12	0.26	
Word Order	0.28	0.09	3.27	<.01	**
Noise	-0.60	0.05	-12.32	<.001	***
Frequency	-0.21	0.12	-1.78	.08	.
Raven IQ	0.17	0.09	1.96	<.05	*
Digit Span	-0.03	0.08	-0.32	0.75	
Vocabulary	0.25	0.09	2.94	<.01	**
Trial Number	-0.05	0.05	-1.18	.24	
Surprisal : Noise	-0.24	0.06	-3.83	<.001	***
Word Order : Noise	0.20	0.06	3.32	<.001	***

= -1, Sentence-Initial = 1), and Noise (binary variable, sum-coded: Quiet = -1, Noise = 1), as well as the interactions of Noise and Surprisal, and Noise and Word Order. The random effects structure consisted of a random intercept for Item, and a random intercept of Participant with a random slope of Word Order.

The model revealed a significant effect of Noise (β = -0.60, SE = 0.05, z = -12.32, $p < .001$), with lower accuracy in background noise compared to quiet. The effect of Word Order was significant as well (β = 0.28, SE = 0.09, z = 3.27, $p <$ .01), with higher accuracy in items occurring in sentence-initial position compared to sentence-final position. There was a significant effect of Raven IQ test (β = 0.17, SE = 0.09, z = 1.96, $p < .05$), with higher accuracy for participants with higher scores on the Raven test. A similar effect was found for Vocabulary (β = 0.25, SE = 0.09, z = 2.94, $p < .01$), where higher vocabulary scores coincide with higher accuracy on the memory test. There was a significant interaction of Noise and Surprisal (β = -0.24, SE = 0.06, z = -3.83, $p < .001$), showing that accuracy is higher for items with higher surprisal, but this effect is reduced in background noise compared to quiet. The interaction of Noise and Word Order was significant too (β = 0.20, SE = 0.06, z = 3.32, $p < .001$), showing an opposite effect: accuracy is higher for items occurring in sentence-initial position, and this benefit is stronger in background noise compared to quiet. Other effects were not significant (ps > .08), all effects can be found in Table 5.7. The three-way interaction of Noise, Surprisal, and Word Order was not significant and adding the term did not improve the model fit. Figure 5.6 shows the mean accuracy by surprisal in both noise conditions, while Figure 5.7 shows the mean accuracy by word order. The combined effects of Word Order and Surprisal on Accuracy for both noisy and quiet conditions are shown in Figure 5.8.

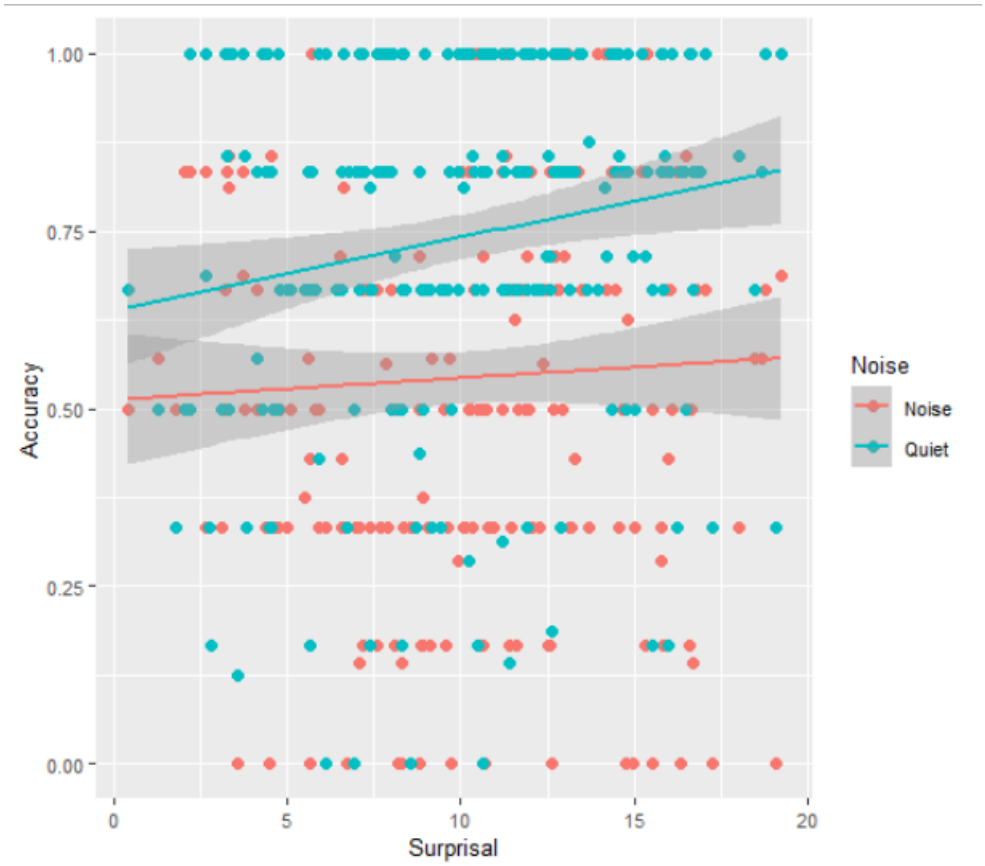

Figure 5.6: This figure shows the accuracy on the memory test for old items by surprisal, split for noise condition.

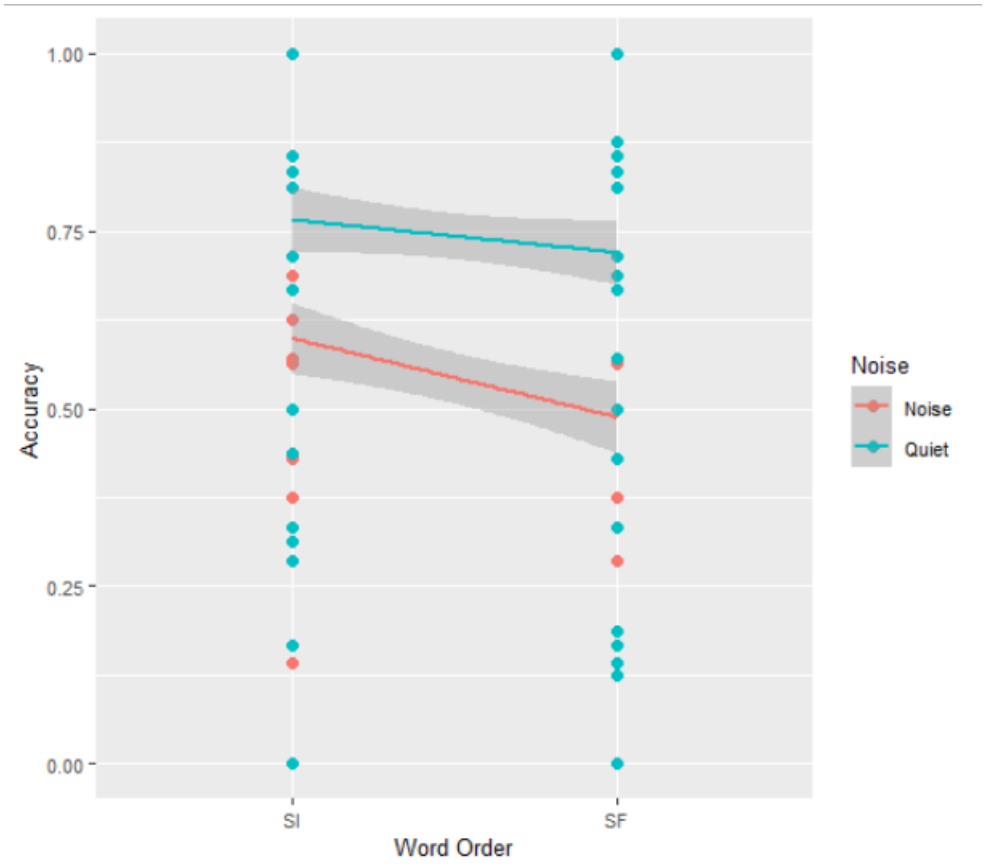

Figure 5.7: This figure shows the accuracy on the memory test for old items by word order, split for noise condition.

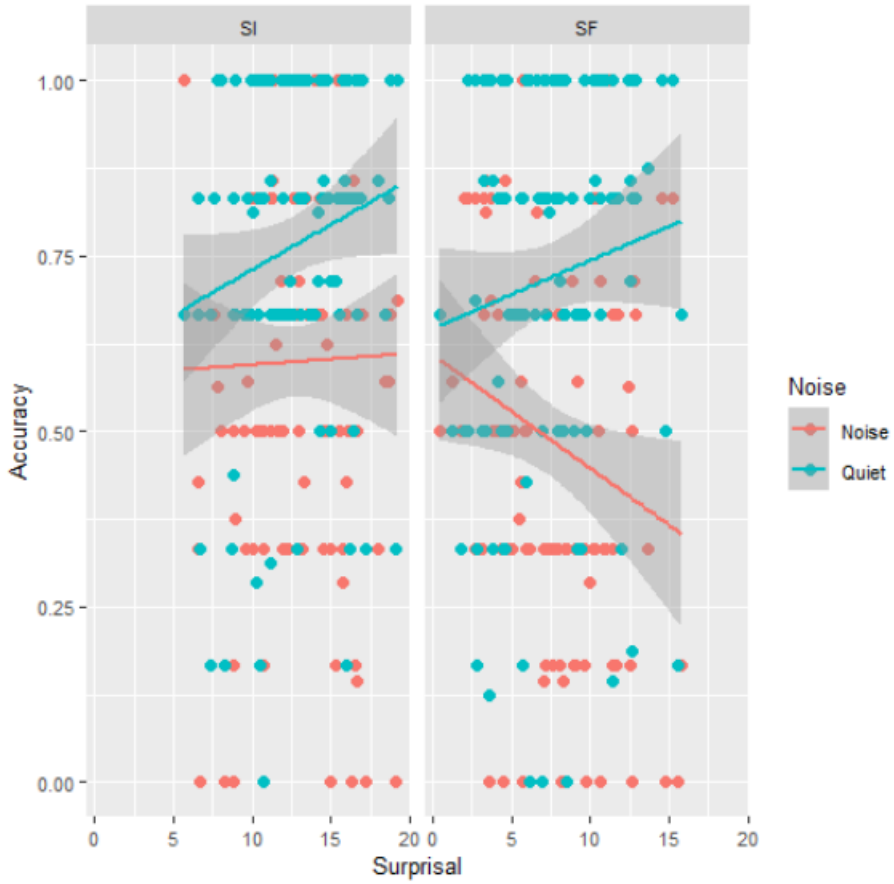

Figure 5.8: This figure shows the accuracy on the memory test for old items by surprisal, split for word order and noise condition.

5.4 Discussion

In this study, we investigated how listening to sentences in background noise affects participants' performance on a surprise memory test, with the aim to test consequences of listening in background noise and different levels of predictability on processing. The results shed light on how higher level processes in communication are affected by the aforementioned factors, beyond word recognition in itself. This can inform how situations with background noise can be managed in which it is important that instructions are understood and remembered correctly. We study false memory effects, which have been found for items that are predicted but not presented (Haeuser & Kray, 2022a; Hubbard et al., 2019). We expected to find these effects especially in conditions where listeners have been found to rely on predictive processes (such as background noise, see Chapters 3 and 4). We varied the frequency of the words as well as the surprisal value of these words in sentences (by changing the word order of the sentence). In the memory test we presented participants with the written form of items heard in the sentences ('old'), semantically related items ('semantic lure') and unrelated, unseen items ('new'). We coded whether participants correctly identified these words as previously heard or not. False alarms for the semantic lures (classifying these items incorrectly as 'old') would point to false memory effects. Participants also completed a Raven IQ test, a backwards digit span test, and a vocabulary test.

Initially, we expected a three-way interaction of Noise, Frequency, and Surprisal. We expected that the (hypothesised) adverse effects of background noise and high surprisal would be further modulated by word frequency. However, we did not find this effect. For our experiment, we based our sample size on a previous study investigating similar memory effects (Nessler et al., 2001), and doubled the number of participants per list to avoid underpowering. However, we might still have been underpowered to find this specific three-way interaction, which the mentioned study did not include. A post-hoc power analysis showed that with the effect size obtained in the current experiment, we would have needed to test close to 700 participants to reach an effect size of 80%.

In our current data, none of the two-way interactions involving Frequency reached significance, either. Across all memory conditions, we did find a main effect of Frequency, which was consistently in the expected direction. These results were in line with previous literature on the Word Frequency Mirror Effect, finding that accuracy is lower on words with higher frequency (Bridger et al., 2014; Glanzer & Adams, 1990; Yonelinas, 2002).

We expected that the semantic lures tested in the memory test would be affected by sentence-level manipulations like Noise and Surprisal (or Word Order). The reasoning here was that in these more difficult listening conditions, participants would rely more on the meaning of the sentence rather than the individual words. This would mean that they might falsely remember the concept of the target word, and make more mistakes in the semantic lure items in the memory test. Similar effects regarding false memories have been found in studies where the predictable word in a highly-constraining sentence was not presented (Haeuser & Kray, 2022a; Hubbard et al., 2019). In our experiment, however, we did not find any evidence for false memory, as none of the tested factors significantly affect the performance accuracy for semantic lures, except for Word Frequency, following the expected direction, in line with the Word Frequency Mirror Effect. Instead, we found higher accuracy for the semantic lure items than the old items, suggesting that participants were able to correctly identify the lures as new more often than to correctly identify the old items as old. Thus, we do not see, as we initially expected, that when participants rely on predictions, they confuse the lures for previously encountered items. A difference between our study and previous studies that did report the prediction-based false memory effect (Haeuser & Kray, 2022a; Hubbard et al., 2019), is in the materials. Compared to ours, theirs consisted of stronger constraining sentences, where the lure word was the most predictable word (highest cloze candidate). Our materials, on the other hand, based the lures on less strong semantic associations. This is more in line with the classic paradigm studying false memory, using word lists of associated items (Deese, 1959; Roediger & McDermott, 1995; Sommers & Lewis, 1999), and it might be the case that more than a single exposure to a set of associated items is necessary to show the false memory effect. Still, we had expected that the presence of background noise would lead to stronger predictions by participants, as also found in the previous two experiments in this dissertation.

It is interesting that none of the sentence-level effects are able to affect the lures. Of course, the semantic lures were never presented in the sentences themselves, but closely related words were. We had expected that predictive processes would lead to some activation of the semantic lures, which we would be able to show in our data as false memories. For example, we had predicted to find a larger amount of false alarms (i.e. lower accuracy) for lures when the related target word was presented in a high surprisal and / or noise condition. In these cases, participants would rely more on prediction to help deal with the increased processing effort. We expected that here they would make more mistakes recalling whether it was exactly the old target item or the related lure that they were presented with. This is not what we found.

Instead, there were no differences between noise, surprisal, and word order conditions in accuracy on the semantic lures.

We did find evidence for a more general interference from the related sentences in that the accuracy on the memory test was lower for the semantic lure items than the new items, which was in line with our hypotheses. Both these types of items were new to the participants, and, all else being equal, should lead to similar levels of performance accuracy. Finding the lower accuracy for the semantic lures compared to the new items suggests that the sentence contexts, regardless of condition, led to priming effects that interfered with correctly classifying the semantic lure items as "new". These priming effects did not depend on the exact conditions in which the sentences were presented to the participants. We additionally found an interaction of memory condition and frequency when comparing the new and semantic lure items. While overall, we find the expected Word Frequency Mirror Effect, this effect was reduced for the lures compared to the new items. It is possible that the priming effect or semantic interference from the sentences with related target words was larger than the Word Frequency Mirror Effect, thus reducing the latter. Within each experimental list, we made sure that the word frequencies and length in syllables did not differ between the different memory condition types. The observed interaction effect of frequency and memory condition, therefore, cannot be attributed to inherently different frequencies in the semantic lures compared to the new items.

Based on our results, we do *not* expect that our participants failed to use predictive processing altogether. We see in the data for the old items that accuracy is affected by surprisal in such a way that the adverse effects of background noise can be overcome by using the sentence as a guide. For items with low surprisal, we see similar accuracy scores regardless of the noise condition, suggesting that the predictability of the sentence helps participants process the sentence in difficult listening conditions. This is in line with the results from the experiments presented in Chapters 3 and 4 and other literature (Boothroyd & Nittrouer, 1988; Dubno et al., 2000; Hutchinson, 1989; Kalikow et al., 1977; Pichora-Fuller et al., 1995; Sommers & Danielson, 1999).

As expected, we find lower accuracy for old items when they occurred in sentences embedded in background noise, compared to those presented in quiet. This could be due to disruptive effects on memory encoding, or because the words were not recognized properly in the noise. If the word was not identified during the listening phase, it will subsequently also not be remembered correctly. When running the same experiment (151 participants) with stimuli embedded in 0 dB SNR background noise, we found no effect of noise at all, suggesting that this was too easy a condition overall

to affect participants' memory. Taken together, these results show that the level of the background noise needs to be chosen with care, so as not to obtain either ceiling or floor effects.

Originally, we expected that our manipulations of word order and surprisal would give similar effects, as they are correlated ($r(5918) = .61$, $p < .005$), and this is the way we constructed our stimuli. However, in the subset of old items, we see that including both factors in the model improves the model fit, and in fact that their effects go in opposite directions. As expected, we find that higher surprisal leads to higher accuracy, but that this effect is more pronounced in quiet compared to background noise. This suggests that the predictability of the target word affected recollection in noise: When the word was harder to predict (high surprisal) *and* there was background noise present, participants struggled to recognize the word correctly in the memory test, possibly because processing was doubly hard in these items.

Conversely, we find that in sentences where the target word occurs in Sentence-Final position, thus corresponding to lower surprisal in general, the accuracy on the memory test is *lower*. This effect is amplified by background noise. We attribute this effect of word order to three points. First, finding higher accuracy in sentence-initial position might be due to a primacy effect, where words at the beginning of the sentence are better remembered than subsequent words. Second, previous studies have found that inaccurate predictions can lead to decreased recognition in latter parts of a sentence (Marrufo-Pérez et al., 2019). This might explain why accuracy reduced more for sentence-final items in background noise than for sentence-final items in quiet, compared to the respective condition in quiet listening conditions. Third, due to how we constructed our stimuli, the level of the background noise might have been relatively higher in the final part of the sentences, thus (inadvertently) affecting participants' accuracy. We will take up this point in the section on our study's limitations.

Our experiment included three additional tests: a German vocabulary test (Engel & Ettinger, 1997; Heubeck, 2001), a set of twelve Raven IQ matrices (Bilker et al., 2012; Raven, 2000), and a backwards digit span test testing working memory (Baddeley, 1968; Baddeley, 1996; Baddeley, 2003; Colle & Welsh, 1976; Olsthoorn et al., 2014; Salame & Baddeley, 1982). We added these three measures to our statistical models and found a significant effect on memory accuracy for the old items of vocabulary score and Raven IQ. We found that participants with larger vocabularies as measured by this test, performed better on the old items in the memory test. Our data contained a large amount of words with low frequency that participants might

not be very familiar with. A better vocabulary score would be an advantage as that increases the chance of knowing the words in our experiment already, making recognition of these words easier. Similarly, a higher score on the Raven IQ test predicted better accuracy on the old items in the memory test. Interestingly, neither of these tests were significant predictors for the new and lure items, even though across the lists these items had equal frequency counts and length in syllables as the old items. The scores of the backwards digit span test did not significantly predict accuracy on the memory test either, against our expectations. As this is a common test for working memory and has also been used to test phonological loop Baddeley, 1968; (Baddeley, 1996; Baddeley, 2003; Colle & Welsh, 1976; Olsthoorn et al., 2014; Salame & Baddeley, 1982), we expected that participants with a higher score on this task would perform better on the memory test as well. One possible explanation for not finding such an effect is that we gave participants a surprise memory task. They were not expecting it, and as such did not actively try to memorise the items in the experiment. It might be the case that this active component was necessary to connect the results of the memory task to the digit span task.

We set out to investigate the consequences of both background noise and predictability on language processing. For this, we used recognition memory as a proxy of processing. This measure was the most straightforward one for our web-based experimental paradigm, but of course reflects subsequent effects of our manipulations rather than immediate effects. There are other measures that give a more immediate picture of how language processing is affected by different conditions. Online measures, such as eye-tracking, pupillometry, or neuro-behavioural measures like ERPs or fMRI could give insight into any changes in processing demands while the stimuli are being presented to the participants. Further demands on the participants' cognitive resources could be made through a dual task setup, for example using a driving simulator. In this way, future research can investigate more directly than we have done here how different levels of predictability and cognitive effort affect language processing, and as such shed more light on false memory effects. Such studies could be grounded in theoretical frameworks that make specific predictions of online processing.

5.4.1 Limitations

As mentioned above, our unexpected word order effect might have been (partially) due to the way we constructed our stimuli. We chose to determine the level of the background noise based on the average intensity of the entire sentence, and subse-

quently computed the intensity of the background noise to reach a signal-to-noise ratio of -5 dB. Because the intensity of naturally spoken sentences tends to drop towards the end of the utterance (Vaissière, 1983), this meant that the background noise might have been relatively of a higher level towards the end of the sentence compared to the beginning. This would mean that these items were more difficult to identify correctly in background noise compared to target words occurring at the start of the sentence, thus adding a confound to the study. An alternative way would have been to determine the intensity level of the background noise on the target word only, rather than the entire sentence. However, this would have led to a higher level of noise on the *beginning* of the sentence, which would make the context harder, or impossible, to understand. Therefore, this method would lead to similar effects of bias in the data as our way of constructing might have done.

Previous studies found that frequency effects mainly depend on the frequency of the root word, rather than surface frequency (Traxler & Gernsbacher, 2011). We retrieved our word frequencies from the CELEX database (Baayen et al., 1995), which are surface frequency counts. Additionally, a CELEX is by now several decades old, the frequency information is a little out-dated. A number of items in our stimuli did not occur in the database, leading to a frequency of 0, even if the word is commonly used nowadays. Despite this, we still find evidence for the expected Word Frequency Mirror Effect in our data, which suggests that our frequency data was fine-grained enough to lead to established effects.

Our memory test consisted of sixty items, where the correct response to 20 was "yes, I have heard this item before", and to the remaining 40 "no, I have not heard this item before". Participants were not warned of this difference between the responses and might have implicitly expected an even split between "yes" and "no" answers. However, participants usually only notice such imbalances when they are more extreme than our 1/3 vs 2/3 split (such as 80% vs 20%, for example). Another effect of our distribution of "yes" and "no" responses might be the following. As people have an inherent bias to respond with "yes" (Budd et al., 1981; Cronbach, 1942), the distribution of correct responses in our memory test might have led participants to respond with "yes" in more cases than warranted, thus reducing accuracy for the semantic lures and new items.

Because we were testing online, we were not able to control the participants' equipment, internet connection, sound settings, comprehension of instructions, or testing environment the way we would when testing in the lab. In our instructions, we gave the participants the chance to solve any possible problems with their audio

equipment and to set their volume at a comfortable level. We included questions in our post-experimental questionnaire to check the participants' surroundings while making the test and possible distractions (such as holding a conversation, watching TV, or texting). We also excluded participants who performed at chance on the memory test and who had too many mistakes in the filler comprehension questions. In this way we tried to minimize the effects of online testing, but it is still a source of variation across participants that might have affected our results.

5.5 Summary

In our previous experiments (Chapters 3 and 4) we studied word recognition in background noise by asking participants to type in what they heard when presented with recorded sentences. In the present experiment, we made a first step away from simple recognition and aimed to shed light on the consequences on higher levels of processing, and thus find implications for real world communication in high-noise situations. We focused on the phenomenon of false memory (Haeuser & Kray, 2022a; Hubbard et al., 2019). As we expected that false memory effects depend on how much the listener relied on predictive processes while being presented with the sentence, we varied the difficulty of the listening condition by adding background noise and changing the surprisal value of the target word. We also investigated how word frequency interacts with these factors. Additionally, we tested if the memory results could be predicted by individual differences tests such as Raven's Progressive Matrices for non-verbal IQ, a backwards digit span as a measure of working memory, and a German vocabulary test. Our dependent variable was accuracy on the surprise recognition memory test at the end of the experiment, where we tested old presented words, semantically related lure items, and unrelated new items.

Our results showed that while there was interference from the sentences, as shown by the reduced accuracy of semantic lures compared to the new items, accuracy on the semantic lures was not affected by any sentence-level factors, such as background noise or surprisal. Thus, we did not find false memories for words predicted by the sentence but not presented to participants. We did find effects of these sentence-level factors for the old items, whereby noise led to lower accuracy. Unexpectedly, word order and surprisal led to opposite effects, which might be explained partially by the way we constructed our stimuli. While previous research found somewhat conflicting results regarding the effect of predictability on memory, our results show better memory for old items when their surprisal is higher, in partic-

ular in quiet listening conditions. Confirming our hypotheses and previous research, we consistently found a main effect of word frequency: words with higher frequency had lower accuracy in the memory test. This is in line with the Word Frequency Mirror Hypothesis. Regarding the individual differences measures, we found that a larger vocabulary score and higher score on the Raven's Progressive Matrices led to higher accuracy on the memory test for the old items.

Further studies should investigate processing effects more directly using online measures in in-lab experiments, such as eye-tracking or neuro-behavioural measures. To increase the cognitive load for participants, a possible experimental design could make use of a dual-task set up, for example in a driving simulator. This would answer open questions regarding immediate effects, rather than effects occurring later in time, such as the memory effects tested here.

Chapter 6

Discussion & Conclusion

In every-day listening situations it is very seldom quiet. In a majority of the time, there is background noise present, which overlaps and obscures the speech signal. In this way, it makes it more difficult to identify what your interlocutor is saying. The present dissertation set out to examine, in broad terms, how speech comprehension is affected by background noise. Specifically, we focused on the benefit that predictability can have when listening in adverse conditions. We investigated the interaction of bottom-up and top-down information streams.

We identified the different research goals that this thesis aimed to address (see also Chapter 1). To start, we aimed to investigate the interplay between context predictability, background noise, and speech sounds on speech recognition. Previous studies investigated up to two of these factors, but never all three together. We were specifically interested in testing the effects of different types of noise, namely babble noise and white noise, as there are conflicting results reported in the literature regarding their masking effects. The different combinations of predictability, noise, and speech sounds lead to fine-grained differences in the intelligibility of the target words, and this allows us to test the predictions of the Noisy Channel Model in so far untested, naturalistic situations. The Noisy Channel Model (Levy, 2008; Levy et al., 2009; Shannon, 1949) is a model of human speech comprehension in noise that poses that listeners combine all sources of information in a rational way. Furthermore, we aimed to test the interaction of top-down and bottom-up processes in older adults, who differ from younger adults as they might have age-induced hearing loss and a tendency to rely more on prediction. These differences allowed us to further test the

predictions of the Noisy Channel Model, as well as to investigate the effect of *false hearing*. This effect of high confidence in incorrectly recognized words has been found to be stronger for older adults than younger adults. Finally, we aimed to test the consequences of speech comprehension in background noise and with more or less predictable sentence contexts. Specifically, we tested the participants' recognition memory and focused on false memory effects, when words that were not presented are remembered as having been encountered before.

To address these research goals, we conducted three experiments, which have been described in the previous chapters. The present chapter will discuss the results of all three experiments and integrate them (Section 6.1). We will additionally discuss the contributions of the dissertation (Section 6.2), as well as their limitations and point out directions for future research (Section 6.3). Finally, we will end with concluding remarks (Section 6.4).

6.1 Main Findings

The first experiment (described in Chapter 3) was a word recognition experiment. We varied the predictability of the target word, the presence or absence of background noise, and speech sounds present in the target word to investigate the interaction of the three factors, while also comparing the effect of different types of noise, namely babble noise and white noise. In the different listening conditions that we thus obtained, we were able to test the predictions made by the Noisy Channel Model. In the high predictability condition, where the sentence context led participants to expect the target word, we found ceiling effects for correct target responses in all conditions, regardless of the noise or speech sound type. In the low predictability condition, where the context pointed to the incorrect distractor word, we found correct target responses in the quiet condition, but not in the noise conditions. Here, the amount of correct responses depended on the speech sound type (plosives were harder to identify correctly than either vowels or fricatives), but crucially also on the interaction of background noise type and speech sound. While there was no difference between the two noise types for plosives, for both fricatives and vowels, one type of noise led to significantly fewer correct responses. For fricatives, white noise is particularly detrimental, while for the recognition of vowels this is babble noise.

Thus, the findings from this first experiment showed that the three factors of interest - predictability, background noise, and speech sounds - interact, leading to fine-grained differences in the clarity of the acoustic signal. According to the Noisy

Channel Model (Levy, 2008; Levy et al., 2009; Shannon, 1949), the observed fine-grained differences in clarity should cause listeners to rely on differing extents on either the acoustic signal or the semantic context. Our results are in line with these predictions, supporting the model. In our results, we observed no overall effect of background noise type, but instead found that its impact depended on the speech signal it obscured.

In the second experiment (described in Chapter 4), we conducted a similar study to the first one. We tested not only younger adults, but also older adults, a population that is of interest given they differ from younger adults in how they combine top-down and bottom-up information. Again, we found ceiling effects in the high predictability condition, where participants correctly responded with the target word, also older adults, regardless of background noise. In the low predictability condition, instead we find that the accuracy depends on the level of the background noise, as well as the age of participants and its interaction with the speech sound. Like in Experiment 1, we found that words with plosive contrasts are harder to identify correctly than words with vowel contrasts. The participants' accuracy scores and mistakes show that older adults rely more on the semantic context than younger adults, like previous research has shown, and also that they are more negatively affected by (louder) noise than younger adults. We additionally investigated the effect of false hearing, for which we collected confidence ratings. We found that participants' confidence corresponded with the difficulty of the listening condition (lower for noisy conditions; lower for incorrect responses), and we did not find any evidence for false hearing, which would have been expressed in the results as high confidence for incorrect answers, in particular for older adults. In this experiment, we found further support for the predictions of the Noisy Channel Model, which hold also for the different trade-offs between top-down and bottom-up information that older adults have.

In the third experiment (described in Chapter 5), we were interested in consequences of listening in background noise on higher-level processes in communication than simply speech recognition, and tested participants' recognition memory. The participants listened to sentences where the predictability of the target word differed depending on the word order in the sentence, and were subsequently asked for a set of words whether they had heard these in the sentences. These words were of one of three types: old items that were presented; unrelated new items that were not presented; and semantic lures that were not presented but semantically related to the target words. We found that this semantic relationship affected recognition perfor-

mance, with more false alarms for the lure items than the new items, even though all of these items were in fact not presented to participants. However, we did not find any effects from the sentence-level (such as predictability or background noise) on the recognition of the lure items, as we had expected. Instead, only word frequency significantly affected the accuracy of the recognition memory test for the lures. Thus, we found no evidence for false memory, for which we would see a higher amount of false alarms for the semantic lures. In the subset of old items, that were presented to participants, we see that the sentence-level factors have significant effects: We find interactions of background noise both with surprisal and word order, going in opposite directions. Memory accuracy is higher for items with higher surprisal, but this effect is reduced in background noise compared to quiet. Regarding word order, accuracy is higher for words in sentence-initial position, and this effect is stronger in background noise.

6.2 Contributions

The findings outlined above help provide new insights in the field of speech comprehension in background noise. Taken together, they make the following contributions.

First, we addressed the question how three factors that have been found to affect speech comprehension, interact. These factors are the background noise, the speech sounds present in the stimuli, and the predictability of the semantic context. Previous studies investigated the factors in isolation, or tested the interaction of pairs of these factors, but their three-way interaction has not been studied in one experiment before. The studies that investigated predictability effects in noise did not carefully control the types of sounds and how they are affected by noise (Boothroyd & Nittrouer, 1988; Dubno et al., 2000; Hutchinson, 1989; Kalikow et al., 1977; Pichora-Fuller et al., 1995; Sommers & Danielson, 1999), while the literature on effects of background noise on speech sounds does not specifically manipulate predictability effects in sentence comprehension (Alwan et al., 2011; Cooke, 2009; Gordon-Salant, 1985; Phatak et al., 2008; Pickett, 1957). Additionally, results on the effect of background noise are inconclusive regarding which type of noise affects comprehension most severely (Danhauer & Leppler, 1979; Gordon-Salant, 1985; Horii et al., 1971; Nittrouer et al., 2003; Taitelbaum-Swead & Fostick, 2016).

We replicate the finding of previous studies that predictable sentences are easier to recognize in background noise, with consistently higher accuracy for items in the high predictability condition than for those in the low predictability condition.

However, we add to this that some sets of items benefit more from the predictive context than others, depending on the speech sounds present in the stimuli. The different speech sounds overlap to varying extents with the different types and levels of background noise, which affects how well they can be recognized. Thus, we show that the three-way interaction of background noise, speech sounds, and predictability is significant, and that it is important to consider the characteristics of the stimuli when studying intelligibility effects in background noise. These effects on the level of the stimuli might explain the so far inconclusive findings in the literature regarding the effects of different types of background noise on speech comprehension. The conflicting findings might in part stem from the differences in the used stimuli, where the combination of the types of background noise and speech sounds present in the items led to differing recognition scores across the studies.

Second, we tested the predictions of the Noisy Channel Model (Levy, 2008; Levy et al., 2009; Shannon, 1949). This is a model of human speech comprehension in background noise. It proposes that language comprehension is a rational process, during which all the different sources of information that are available to the listener are combined. In the model, this is done through Bayes' Rule and by combining prior knowledge in the form of linguistic and world knowledge (which meanings are more plausible; what is the base-rate frequency of certain grammatical constructions), with knowledge about what the most likely corruptions due to the noise might be. Previous studies have tested the predictions that the Noisy Channel Model makes, primarily by investigating the interpretation of written syntactic alternations with plausible and implausible versions (Gibson et al., 2013; Gibson et al., 2016; Gibson et al., 2017; Poppels & Levy, 2016; Ryskin et al., 2018). These studies provide strong evidence in favor of the Noisy Channel Model.

In their original study, Gibson et al. (2013) quantified the edit distance between the plausible and implausible version of their alternations in a change consisting of insertions and deletions of function words (see also Gibson et al., 2016; Gibson et al., 2017). In this dissertation, we changed the type and level of background noise to manipulate the distance between the target and distractor. Here the distance between the two depends on the similarity between the acoustic signal of the speech and that of the background noise. As such, it is better grounded in the strength of masking of the signal, and less arbitrary than the edit distance measured in terms of insertions and deletions. In our experiments, we consistently find supporting evidence for the predictions of the Noisy Channel Model. Thus, we show that these predictions hold also when tested in more naturalistic settings than previously has been done, and with

a different set of stimuli (similar sounding words instead of syntactic alternations). These findings reinforce the belief that human listeners behave rationally.

Third, we investigated how older adults and younger adults differ from each other during speech comprehension in background noise. Previous studies have found that while older adults are generally more negatively affected by the presence of background noise than younger adults (Benichov et al., 2012; Dubno et al., 2000; Hutchinson, 1989; Pichora-Fuller et al., 1995; Sommers & Danielson, 1999), they are able to compensate for this by relying on predictions made from context, as their general language abilities are well-preserved (Lash et al., 2013; Stine & Wingfield, 1994; Wingfield et al., 1995; Wingfield et al., 2005). In fact, older adults have been found to rely on context to a larger extent than younger adults, regardless of listening conditions (Koeritzer et al., 2018; Sheldon et al., 2008; Wingfield et al., 2005). These predictions can come at a cost when they lead the listener to an incorrect word. *False hearing* has been defined as a mistaken high confidence in correctly having recognized a word, when in fact the recognition was incorrect (Failes et al., 2020; Failes & Sommers, 2022; Rogers et al., 2012; Rogers & Wingfield, 2015; Rogers, 2017; Sommers et al., 2015). The effect of false hearing has been found to be stronger for older adults than younger adults.

Like previous studies, we find that older adults indeed rely more on sentence context than younger adults do. Their answers were more in line with the sentence context, even when this was incorrect, and particularly so in louder levels of noise and with larger overlap between speech sounds and noise. However, we were not able to replicate the finding of the false hearing effect. When looking at the confidence ratings, we find that they reflect the difficulty of the listening condition and are lower for incorrect responses. One possible explanation for this can be the differences in experimental set-up: our study was run via the internet and with a group of older adults that were younger than those typically participating in lab studies. It is possible that false hearing effects only increase for these older participants (65+). For our participants, we can conclude that they do show some of the typical effects reported in the literature (strong reliance on predictive processes), but that they behave rationally as well. Their responses in the recognition test are made based on a trade-off of the information that is available to them (acoustic signal vs semantic context), as shown by the small-grained differences depending on the exact listening condition and the interaction of background noise and speech sounds in the stimuli.

Finally, we moved beyond reported word recognition, and investigated consequences of listening in background noise. We focused on the effects of noise and

predictability on recognition memory. Of particular interest were *false memories*, which are items that participants report as having been presented before, but that in fact were new. This effect depends on the strength of the predictive processes during listening and encoding. These effects have been found using lists of semantically related words (Deese, 1959; Roediger & McDermott, 1995; Roediger et al., 2001) and recently also in visually presented sentences, where the sentence context led to the prediction of a particular word that either was or was not the presented continuation (Haeuser & Kray, 2022a; Hubbard et al., 2019).

From our recognition memory experiment, we find no evidence for false memories. Participants do not incorrectly classify the semantic lures as old items when items are presented in noise. This finding is surprising, as we had expected that due to the stronger reliance on predictive processes (as shown by the other experiments in this dissertation), the lure words would be activated and incorrectly "remembered". We did find effects of noise and predictability for the memory of actually presented, old items. Here, the difficulty of the listening condition affected the memory accuracy scores so that there were fewer correct responses in background noise. We find a similar explanation for the confidence ratings, both in the previous word recognition experiments and for the memory task: the more difficult the listening condition, the lower the participants' confidence tended to be. This shows that listeners are aware of the effort that is required during listening and able to judge the accuracy of their responses.

6.3 Limitations and Future Research

The experiments presented in the current dissertation provided results that answer many of the raised research questions and thereby address the set research goals. However, these results also raise new questions that have not been answered yet. Furthermore, the experiments described here were in part affected by limitations on their designs caused by the global Covid-19 pandemic. As such, we can make several suggestions for interesting future work that builds upon the current dissertation. These will be described in this section.

First, we conducted our three experiments online. In part, this was planned: using crowd-sourcing methods has several benefits that outweigh possible downsides. Data collection is fast, much faster than for lab-based experiments. Here participants are typically tested one by one, while online a large group of participants is available, who come from a more varied background than the student population that is

typically recruited. From the start, we planned to run our rating studies and pilot studies (where possible) online to facilitate data collection. The main experiments, in particular those with older adults, we intended to conduct in the lab, which would have given us the ability to invite older participants (65+ years rather than up to 65 years as we were able to recruit online), better control their understanding of instructions and, importantly, test their hearing levels using an audiometer. Covid-19-related lock-downs made it impossible to invite participants to the lab, especially such a vulnerable population of older adults, leaving us with no choice but to recruit them online. The pool of native German-speaking, older adults was very limited on the recruiting platform, especially at the start of the pandemic. In our study, we were unable to replicate previously found effects of false hearing for older adults. It is possible that this lack of an effect on our side was caused by the very different experimental design and tested population compared to these previous studies. Still, our results do show reliable effects of mishearing that increase with age in expected directions, suggesting that our different design was effective also with the tested older adults.

Relatedly, it can be debated whether running studies especially with auditory stimuli online ensures sufficient control of the participant's technical set-up and environment. In a lab study, it is possible to test each participant using the same hardware and software, and make sure the surroundings are quiet and not distracting. In an online set-up, this control is much more limited. A lot of it relies on participants adhering to instructions regarding their surroundings, distractions, and devices. Attention check questions can be used to assess whether participants are focused on the task and whether their results can be used in further analyses. Regardless of these precautions, data obtained via online crowd-sourcing tends to be noisier than that from in-lab experiments (see for example Cooke et al. (2011); Mayo et al. (2012); Slote & Strand (2016); Wolters et al. (2010); but for comparable results between the two populations Peelle et al. (2016)). Recently, Cooke & García Lecumberri (2021) compared results on a speech recognition task of participants tested in the lab and a group of participants drawn from the same population who were tested online. Their results showed that these groups performed comparably when those with low quality technical equipment were removed. This suggests that previous studies' discrepant results between in-lab and online tested groups are caused by on one hand differences in set-up and environment, and on the other in differences in the more varied population that can be reached online (which is in fact one of the advantages of crowd-sourcing experiments). Altogether, it seems that the results from online studies with auditory stimuli can be taken at face value, so long as this increased noise in the data is taken

into account. Further support for the reliability of our online studies comes from the fact that we were able to obtain and replicate the general expected effects in our studies that acted as controls. Consistently, we found adverse effects of background noise, better performance in easier listening conditions both for word recognition as well as our memory measure, and finding the Word Frequency Mirror Effect in the third experiment. This suggests that our experimental designs were effective despite the online set-up.

We were unable to replicate certain expected effects, such as false hearing (higher confidence in incorrectly recognized words) and false memory (recognizing a word as old when it was not presented before), which have been reported in the literature. Regarding the false hearing effect, it is possible that the difference between our lack of effect and previous studies that do report it, is caused by our different experimental design and tested population. As mentioned above, in our online study, we were not able to recruit participants of the same age as in these previous studies, which might explain why we found no effect of false hearing. Future studies should determine, if this is the case, from what age, to what extent, and in which situations, false hearing does occur. The lack of a false memory effect is more puzzling. This effect has been found in crowd-sources studies (Haeuser & Kray, 2022a), and thus the paradigm itself cannot be the reason. One difference is in the stimuli: previous studies finding false memory effects used more constraining sentences, in which the lure word was the candidate word with the highest cloze value. In our study our lures were instead based on semantic associations to the target word. We selected these by checking whether the word had a similar meaning and would fit in the sentence, but did not cloze test these items to ensure they were in fact predicted by participants. Therefore, our semantic lures might not have been sufficiently strongly predicted to lead to false memory. We had expected that the lures would be activated especially in our background noise conditions, which would lead participants to rely on predictive processes (as shown in our other experiments). Classic studies investigating false memory made use of word lists consisting of items with similar semantic associations (Deese, 1959; Roediger & McDermott, 1995; Sommers & Danielson, 1999). It is possible that for the effect to appear in these cases of associations rather than predictions (as tested through cloze testing), multiple exposures are necessary, which was not the case in our experiment. Future studies should determine which situations lead to false memory, and how strong the relation between the sentence context and the presented lure has to be.

In recent years, a lot of research has emerged studying the role of individual differences in language processing (see for example Goodhew & Edwards (2019); Hintz et al. (2021). While traditionally results of a sample are averaged in order to obtain general effects in the population, the development of more powerful analysis methods and crowd-sourcing have led to researchers asking the question how differences between listeners matter. In the present dissertation we made a first step towards testing individual differences in our third experiment by including three tests (a backward digit span test, a vocabulary test, and Raven's Progressive Matrices). Future research could strengthen this focus and include additional measures and increase the sample size of the study to investigate how listeners differ from each other. This is interesting in regard to the recognition memory study, as it is established that there are individual differences in working memory capacity, but also for the prediction effects investigated in the current dissertation. It would be of interest to investigate whether there are differences between listeners in how much they rely on predictive processes in adverse listening conditions, similar to what is known regarding younger and older adults.

One consequence of our online web-based experimental designs is that we were only able to test offline measures of language comprehension. Both our word recognition task, the confidence ratings, and the recognition memory test measure the participants' comprehension after all processes have been completed. However, being able to see what happens *during* these processes would be a valuable extension to the current results, and these results would be able to go beyond the questions addressed in this dissertation. Possible methods that would yield such online measures are eye-tracking, pupillometry, or neurobehavioural measures like ERPs or fMRI. Implementing these methods would require changing the experimental tasks somewhat, for example adding a dual task to increase processing demands on the side of the participant. In this way, future research can investigate more directly than the current dissertation was able to how different levels of predictability and background noise affect speech recognition, both directly and in higher-level processes.

6.4 Conclusion

In this dissertation, we presented three experiments that investigated how human speech comprehension is on the one hand negatively impacted by the presence of background noise, and on the other hand aided by predictive semantic contexts. The results show that small-grained differences in the intelligibility of the stimuli affect

how the listener relies on either the bottom-up signal (acoustic signal) or top-down predictions. Across the three experiments, our findings support the predictions made by the Noisy Channel Model, namely that the reliance on predictive processes is dependent on the effort that is required to process the speech signal and the amount of overlap between speech and background noise. We additionally show that the difficulty of the listening condition affects meta-cognitive judgements, operationalised through confidence ratings: the more difficult the listening condition, the lower listeners' confidence was, both for word recognition and memory. Memory performance itself was similarly affected by the difficulty of the listening condition. We did not find evidence for false hearing or false memory.

In sum, the findings in this dissertation contribute to our understanding of speech recognition in adverse listening conditions, in particular background noise, and how predictive processes can both help and hinder speech perception. The results consistently confirmed the predictions of the Noisy Channel Model, thus providing evidence that human listeners behave rationally, and combine information available from different sources in an optimal way to maximize understanding and to minimize effort. Future studies can shed light on the open questions raised in this dissertation, such as in which experimental situations false hearing and false memory occur exactly, and the implications this has for real-life communication. New insights can be gained in particular through the use of online research methods (eye-tracking, pupillometry, ERP, fMRI, etc.) that open up the possibility for other tasks and experimental set-ups. Additionally, another avenue to advance the field would be through studying individual differences in these language comprehension situations, with a focus on differences in the use of prediction and recognition memory.

List of Figures

List of Tables

Bibliography

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*(4), 419–439.

Alwan, A., Jiang, J., & Chen, W. (2011). Perception of place of articulation for plosives and fricatives in noise. *Speech Communication*, *53*(2), 195–209.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429.

Anderson, J. R., & Bower, G. H. (1974). A propositional theory of recognition memory. *Memory & Cognition*, *2*(3), 406–412.

Arons, B. (1992). A review of the cocktail party effect. *Journal of the American Voice I/O Society*, *12*(7), 35–50.

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), 1–31.

Ayasse, N. D., Hodson, A. J., & Wingfield, A. (2021). The principle of least effort and comprehension of spoken sentences by younger and older adults. *Frontiers in Psychology*, *12*, 1–13.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). The celex lexical database (release 2). Distributed by the Linguistic Data Consortium, University of Pennsylvania.

Baddeley, A. (1968). How does acoustic similarity influence short-term memory? *The Quarterly Journal of Experimental Psychology*, *20*(3), 249–264.

Baddeley, A. (1996). The concept of working memory. In S. E. Gathercole (Ed.), *Models of Short-term Memory* (pp. 1–27). London, UK: Psychology Press.

Baddeley, A. (2003). Working memory: looking back and looking forward. *Nature Reviews Neuroscience*, *4*(10), 829–839.

Balcetis, E., & Dunning, D. (2010). Wishful seeing: More desired objects are seen as closer. *Psychological Science*, *21*(1), 147–152.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:1406.5823.

Benichov, J., Cox, L. C., Tun, P. A., & Wingfield, A. (2012). Word recognition within a linguistic context: Effects of age, hearing acuity, verbal ability and cognitive function. *Ear and Hearing*, *32*(2), 250–256.

Benkí, J. R. (2003). Analysis of english nonsense syllable recognition in noise. *Phonetica*, *60*(2), 129–157.

Bilker, W. B., Hansen, J. A., Brensinger, C. M., Richard, J., Gur, R. E., & Gur, R. C. (2012). Development of abbreviated nine-item forms of the raven's standard progressive matrices test. *Assessment*, *19*(3), 354–369.

Boersma, P., & Weenink, D. (2009). Praat: doing phonetics by computer (version 6.1.05). URL: `http://www.praat.org`.

Boothroyd, A., & Nittrouer, S. (1988). Mathematical treatment of context effects in phoneme and word recognition. *The Journal of the Acoustical Society of America*, *84*(1), 101–114.

Brainerd, C. J., Reyna, V. F., & Ceci, S. J. (2008). Developmental reversals in false memory: a review of data and theory. *Psychological Bulletin*, *134*(3), 343–382.

Bridger, E. K., Bader, R., & Mecklinger, A. (2014). More ways than one: Erps reveal multiple familiarity signals in the word frequency mirror effect. *Neuropsychologia*, *57*, 179–190.

Bronkhorst, A. W. (2000). The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta Acustica united with Acustica*, *86*(1), 117–128.

Brothers, T., & Kuperberg, G. R. (2021). Word predictability effects are linear, not logarithmic: Implications for probabilistic models of sentence comprehension. *Journal of Memory and Language*, *116*, 104174.

Bruner, J. S. (1957). On perceptual readiness. *Psychological Review*, *64*(2), 123–152.

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America*, *109*(3), 1101–1109.

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The word frequency effect in word processing: An updated review. *Current Directions in Psychological Science*, *27*(1), 45–50.

Budd, E. C., Sigelman, C. K., & Sigelman, L. (1981). Exploring the outer limits of response bias. *Sociological Focus*, *14*(4), 297–307.

Bushong, W., & Jaeger, T. F. (2019). Modeling long-distance cue integration in spoken word recognition. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* (pp. 62–70).

Carhart, R., Johnson, C., & Goodman, J. (1975). Perceptual masking of spondees by combinations of talkers. *The Journal of the Acoustical Society of America*, *58*(S1), S35–S35.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, *25*(5), 975–979.

Chingacham, A., Demberg, V., & Klakow, D. (2021). Exploring the potential of lexical paraphrases for mitigating noise-induced comprehension errors. arXiv preprint arXiv:2107.08337.

Christensen, R. H. B. (2015). ordinal-regression models for ordinal data. R package version 2015.6-28.

Christianson, K. (2016). When language comprehension goes wrong for the right reasons: Good-enough, underspecified, or shallow language processing. *Quarterly Journal of Experimental Psychology*, *69*(5), 817–828.

Ciocca, V., & Bregman, A. S. (1987). Perceived continuity of gliding and steady-state tones through interrupting noise. *Perception & Psychophysics*, *42*(5), 476–484.

Clark, H. H. (1996). *Using language*. Cambridge, UK: Cambridge University Press.

Coane, J. H., Balota, D. A., Dolan, P. O., & Jacoby, L. L. (2011). Not all sources of familiarity are created equal: the case of word frequency and repetition in episodic recognition. *Memory & Cognition*, *39*(5), 791–805.

Cohen, G. (1987). Speech comprehension in the elderly: The effects of cognitive changes. *British Journal of Audiology*, *21*(3), 221–226.

Colle, H. A., & Welsh, A. (1976). Acoustic masking in primary memory. *Journal of Verbal Learning and Verbal Behavior*, *15*(1), 17–31.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428.

Cooke, M. (2006). A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America*, *119*(3), 1562–1573.

Cooke, M. (2009). Discovering consistent word confusions in noise. In *Tenth Annual Conference of the International Speech Communication Association*.

Cooke, M., Barker, J., Lecumberri, M. L. G., & Wasilewski, K. (2011). Crowdsourcing for word recognition in noise. In *Twelfth Annual Conference of the International Speech Communication Association*.

Cooke, M., Garcia Lecumberri, M., & Barker, J. (2008). The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception. *The Journal of the Acoustical Society of America*, *123*(1), 414–427.

Cooke, M., & García Lecumberri, M. L. (2021). How reliable are online speech intelligibility studies with known listener cohorts? *The Journal of the Acoustical Society of America*, *150*(2), 1390–1401.

Corley, M., MacGregor, L. J., & Donaldson, D. I. (2007). It's the way that you, er, say it: Hesitations in speech affect language comprehension. *Cognition*, *105*(3), 658–668.

Crocker, M. W. (2005). Rational models of comprehension: Addressing the performance paradox. In A. Cutler (Ed.), *Twenty-first century psycholinguistics: Four cornerstones* (pp. 363–380). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cronbach, L. J. (1942). Studies of acquiescence as a factor in the true-false test. *Journal of Educational Psychology*, *33*(6), 401–415.

Culling, J. F., & Stone, M. A. (2017). Energetic masking and masking release. In J. C. Middlebrooks, J. Z. Simon, A. N. Popper, & R. R. Fay (Eds.), *The auditory system at the cocktail party: Springer Handbook of Auditory Research* (pp. 41–73). New York, NY: Springer.

Cutler, A., Weber, A., Smits, R., & Cooper, N. (2004). Patterns of english phoneme confusions by native and non-native listeners. *The Journal of the Acoustical Society of America*, *116*(6), 3668–3678.

Danhauer, J. L., & Leppler, J. G. (1979). Effects of four noise competitors on the california consonant test. *Journal of Speech and Hearing Disorders*, *44*(3), 354–362.

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology*, *58*(1), 17–22.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*(8), 1117–1121.

Dubno, J. R., Ahlstrom, J. B., & Horwitz, A. R. (2000). Use of context by young and aged adults with normal hearing. *The Journal of the Acoustical Society of America*, *107*(1), 538–546.

Durlach, N. I., Mason, C. R., Shinn-Cunningham, B. G., Arbogast, T. L., Colburn, H. S., & Kidd Jr, G. (2003). Informational masking: Counteracting the effects of stimulus uncertainty by decreasing target-masker similarity. *The Journal of the Acoustical Society of America*, *114*(1), 368–379.

Edwards, T. J. (1981). Multiple features analysis of intervocalic english plosives. *The Journal of the Acoustical Society of America*, *69*(2), 535–547.

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, *20*(6), 641–655.

Engel, R., & Ettinger, U. (1997). Satsergänzungstest 1. Psychiatrische Klinik der LMU München, Abteilung für experimentelle Psychologie und Psychophysiologie.

Failes, E., & Sommers, M. S. (2022). Using eye-tracking to investigate an activation-based account of false hearing in younger and older adults. *Frontiers in Psychology*, *13*, 821044.

Failes, E., Sommers, M. S., & Jacoby, L. L. (2020). Blurring past and present: Using false memory to better understand false hearing in young and older adults. *Memory & Cognition*, *48*(8), 1403–1416.

Federmeier, K. D., & Kutas, M. (2005). Aging in context: age-related changes in context use during language comprehension. *Psychophysiology*, *42*(2), 133–141.

Federmeier, K. D., Kutas, M., & Schul, R. (2010). Age-related and individual differences in the use of prediction during language comprehension. *Brain and Language*, *115*(3), 149–161.

Federmeier, K. D., Van Petten, C., Schwartz, T. J., & Kutas, M. (2003). Sounds, words, sentences: age-related changes across levels of language processing. *Psychology and Aging*, *18*(4), 858.

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.

Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology*, *47*(2), 164–203.

Ferreira, F., Bailey, K. G., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current directions in psychological science*, *11*(1), 11–15.

Ferreira, F., & Lowder, M. W. (2016). Prediction, information structure, and good-enough language processing. *Psychology of Learning and Motivation*, *65*, 217–247.

Ferreira, F., & Patson, N. D. (2007). The 'good enough' approach to language comprehension. *Language and Linguistics Compass*, *1*(1-2), 71–83.

Frank, A. F., & Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. volume 30.

Freyman, R. L., Balakrishnan, U., & Helfer, K. S. (2004). Effect of number of masking talkers and auditory priming on informational masking in speech recognition. *The Journal of the Acoustical Society of America*, *115*(5), 2246–2256.

Friedrich, C. K., Eulitz, C., & Lahiri, A. (2006). Not every pseudoword disrupts word recognition: an erp study. *Behavioral and Brain Functions*, *2*(1), 1–10.

Frisson, S., Rayner, K., & Pickering, M. J. (2005). Effects of contextual predictability and transitional probability on eye movements during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(5), 862–877.

Fu, Q.-J., Shannon, R. V., & Wang, X. (1998). Effects of noise and spectral resolution on vowel and consonant recognition: Acoustic and electric hearing. *The Journal of the Acoustical Society of America*, *104*(6), 3586–3596.

Garcia Lecumberri, M. L., Cooke, M., & Cutler, A. (2010). Non-native speech perception in adverse conditions: A review. *Speech Communication*, *52*(11-12), 864–886.

Gates, G. A., & Mills, J. H. (2005). Presbycusis. *The Lancet*, *366*(9491), 1111–1120.

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, *110*(20), 8051–8056.

Gibson, E., Sandberg, C., Fedorenko, E., Bergen, L., & Kiran, S. (2016). A rational inference approach to aphasic language comprehension. *Aphasiology*, *30*(11), 1341–1360.

Gibson, E., Tan, C., Futrell, R., Mahowald, K., Konieczny, L., Hemforth, B., & Fedorenko, E. (2017). Don't underestimate the benefits of being misunderstood. *Psychological Science*, *28*(6), 703–712.

Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: data and theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(1), 5–16.

Goodhew, S. C., & Edwards, M. (2019). Translating experimental paradigms into individual-differences research: Contributions, challenges, and practical recommendations. *Consciousness and Cognition*, *69*, 14–25.

Gordon-Salant, S. (1985). Phoneme feature perception in noise by normal-hearing and hearing-impaired subjects. *Journal of Speech, Language, and Hearing Research*, *28*(1), 87–95.

Gordon-Salant, S., Frisina, R. D., Fay, R. R., & Popper, A. (2010). *The aging auditory system.* volume 34 of *Springer Handbook of Auditory Research*. New York, NY: Springer Science & Business Media.

Gow, D. W., & Gordon, P. C. (1995). Lexical and prelexical influences on word segmentation: evidence from priming. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(2), 344–359.

Greenberg, C. (2022). *Evaluating Humanness in Language Models*. Ph.D. thesis Saarland University, Saarbrücken, Saarland, Germany.

Gregg, V. (1976). Word frequency, recognition and recall. In J. Brown (Ed.), *Recall and Recognition* (pp. 183–216). Chichester, UK: John Wiley & Sons.

Haeuser, K., & Kray, J. (2021). Effects of prediction error on episodic memory retrieval: evidence from sentence reading and word recognition. *Language, Cognition and Neuroscience*, *38*(4), 558–574.

Haeuser, K., & Kray, J. (2022a). Uninvited and unwanted: False memories for words predicted but not seen. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. volume 44.

Haeuser, K. I., & Kray, J. (2022b). How odd: Diverging effects of predictability and plausibility violations on sentence reading and word memory. *Applied Psycholinguistics*, *43*(5), 1193–1220.

Hagoort, P., & Van Berkum, J. (2007). Beyond the sentence given. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 801–811.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Häuser, K., Demberg, V., & Kray, J. (2018). Surprisal modulates dual-task performance in older adults: Pupillometry shows age-related trade-offs in task performance and time-course of language processing. *Psychology and Aging*, *33*(8), 1168–1180.

Haviland, S. E., & Clark, H. H. (1974). What's new? acquiring new information as a process in comprehension. *Journal of Verbal Learning and Verbal Behavior*, *13*(5), 512–521.

Hay, J. F., & Jacoby, L. L. (1999). Separating habit and recollection in young and older adults: effects of elaborative processing and distinctiveness. *Psychology and aging*, *14*(1), 122–134.

Helfer, K., Bartlett, E., Popper, A., & Fay, R. R. (2020). *Aging and Hearing. Causes and Consequences*. Switzerland: Springer.

Heubeck, V. (2001). *Statistische Analyses eines Wortbedeutungstest*. Master's thesis Universität Konstanz, Konstanz, Baden-Württemberg, Germany.

Hintz, F., Voeten, C. C., Isakoglou, C., McQueen, J. M., & Meyer, A. S. (2021). Individual differences in language ability: Quantifying the relationships between linguistic experience, general cognitive skills and linguistic processing skills. In *the 34th Annual CUNY Conference on Human Sentence Processing (CUNY 2021)*.

Hnath-Chisholm, T., Willott, J., & Lister, J. (2003). The aging auditory system: anatomic and physiologic changes and implications for rehabilitation. *International Journal of Audiology*, *42*, 2S3–2S10.

Höltje, G., Lubahn, B., & Mecklinger, A. (2019). The congruent, the incongruent, and the unexpected: Event-related potentials unveil the processes involved in schematic encoding. *Neuropsychologia*, *131*, 285–293.

Hoole, P., & Mooshammer, C. (2002). Articulatory analysis of the german vowel system. *Silbenschnitt und Tonakzente*, *1*, 129–152.

Horii, Y., House, A. S., & Hughes, G. W. (1971). A masking noise with speech-envelope characteristics for studying intelligibility. *The Journal of the Acoustical Society of America*, *49*(6B), 1849–1856.

Hubbard, R. J., Rommers, J., Jacobs, C. L., & Federmeier, K. D. (2019). Downstream behavioral and electrophysiological consequences of word prediction on recognition memory. *Frontiers in Human Neuroscience*, *13*, 291.

Hutchinson, K. M. (1989). Influence of sentence context on speech perception in young and older adults. *Journal of Gerontology*, *44*(2), P36–P44.

Jacoby, L. L., Rogers, C. S., Bishara, A. J., & Shimizu, Y. (2012). Mistaking the recent past for the present: False seeing by older adults. *Psychology and Aging*, *27*(1), 22–32.

Johnson, K. (2004). Acoustic and auditory phonetics. *Phonetica*, *61*(1), 56–58.

Kahneman, D. (1973). *Attention and effort*. Englewood Cliffs, NJ: Prentice Hall.

Kalikow, D. N., Stevens, K. N., & Elliott, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *The Journal of the Acoustical Society of America*, *61*(5), 1337–1351.

Kidd, G., Mason, C. R., Richards, V. M., Gallun, F. J., & Durlach, N. I. (2008). Informational masking. In W. A. Yost, A. N. Popper, & R. R. Fay (Eds.), *Auditory Perception of Sound Sources* (pp. 143–189). Boston, MA: Springer US.

Kliegl, R., Nuthmann, A., & Engbert, R. (2006). Tracking the mind during reading: the influence of past, present, and future words on fixation durations. *Journal of Experimental Psychology: General*, *135*(1), 12–35.

Koeritzer, M. A., Rogers, C. S., Van Engen, K. J., & Peelle, J. E. (2018). The impact of age, background noise, semantic ambiguity, and hearing loss on recognition memory for spoken sentences. *Journal of Speech, Language, and Hearing Research*, *61*(3), 740–751.

Konieczny, L. (2005). The psychological reality of local coherences in sentence processing. In *Proceedings of the 27th annual conference of the cognitive science society* (pp. 1178–1183). Cognitive Science Society Stresa, Italy.

Konieczny, L., Müller, D., Hachmann, W., Schwarzkopf, S., & Wolfer, S. (2009). Local syntactic coherence interpretation. evidence from a visual world study. In *Proceedings of the 31st annual conference of the Cognitive Science Society* (pp. 1133–1138). Cognitive Science Society Austin.

Krech, E.-M., Stock, E., Hirschfeld, U., & Anders, L.-C. (2009). *Deutsches Aussprachewörterbuch*. Berlin, Germany: De Gruyter.

Kukona, A., Cho, P. W., Magnuson, J. S., & Tabor, W. (2014). Lexical interference effects in sentence processing: Evidence from the visual world paradigm and self-organizing models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(2), 326–347.

Kukona, A., & Tabor, W. (2011). Impulse processing: A dynamical systems model of incremental eye movements in the visual world paradigm. *Cognitive Science*, *35*(6), 1009–1051.

Kuperberg, G. R. (2021). Tea with milk? a hierarchical generative framework of sequential event comprehension. *Topics in Cognitive Science*, *13*(1), 256–298.

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.

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31*(1), 32–59.

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*(5947), 161–163.

Ladefoged, P., & Maddieson, I. (1996). *The Sounds of the World's Languages* volume 1012. Blackwell.

Lahiri, A., & Reetz, H. (2010). Distinctive features: Phonological underspecification in representation and processing. *Journal of Phonetics*, *38*(1), 44–59.

Laney, C., & Loftus, E. F. (2013). Recent advances in false memory research. *South African Journal of Psychology*, *43*(2), 137–146.

Lash, A., Rogers, C. S., Zoller, A., & Wingfield, A. (2013). Expectation and entropy in spoken word recognition: Effects of age and hearing acuity. *Experimental Aging Research*, *39*(3), 235–253.

Levy, R. (2008). A noisy-channel model of human sentence comprehension under uncertain input. In *Proceedings of the 2008 conference on empirical methods in natural language processing* (pp. 234–243).

Levy, R., Bicknell, K., Slattery, T., & Rayner, K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences*, *106*(50), 21086–21090.

Li, L., Daneman, M., Qi, J. G., & Schneider, B. A. (2004). Does the information content of an irrelevant source differentially affect spoken word recognition in younger and older adults? *Journal of Experimental Psychology: Human Perception and Performance*, *30*(6), 1077–1091.

Liberman, A. M., Delattre, P. C., Cooper, F. S., & Gerstman, L. J. (1954). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs: General and Applied*, *68*(8), 1–13.

Lidestam, B., Holgersson, J., & Moradi, S. (2014). Comparison of informational vs. energetic masking effects on speechreading performance. *Frontiers in Psychology*, *5*, 639.

Lindenberger, U., & Ghisletta, P. (2009). Cognitive and sensory declines in old age: gauging the evidence for a common cause. *Psychology and Aging*, *24*(1), 1–16.

Loizou, P. C. (1999). Introduction to cochlear implants. *IEEE Engineering in Medicine and Biology Magazine*, *18*(1), 32–42.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, *19*(1), 1–36.

Maddieson, I. (1984). *Patterns of sounds*. Cambridge University Press.

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review*, *87*(3), 252–271.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York, NY: W.H Freeman and Company.

Marrufo-Pérez, M. I., Eustaquio-Martín, A., & Lopez-Poveda, E. A. (2019). Speech predictability can hinder communication in difficult listening conditions. *Cognition*, *192*, 103992.

Marslen-Wilson, W. (1993). Issues of process and representation in lexical access. In G. T. M. Altmann, & R. Shillcock (Eds.), *Cognitive Models of Speech Processing: The Second Sperlonga Meeting* (pp. 187–210). Lawrence Erlbaum Associates Publishers.

Mattys, S. L., Brooks, J., & Cooke, M. (2009). Recognizing speech under a processing load: Dissociating energetic from informational factors. *Cognitive Psychology*, *59*(3), 203–243.

Mayo, C., Aubanel, V., & Cooke, M. (2012). Effect of prosodic changes on speech intelligibility. In *Thirteenth Annual Conference of the International Speech Communication Association*.

McQueen, J. M. (2005). Speech perception. In K. Lamberts, & R. Goldstone (Eds.), *The Handbook of cognition* (pp. 255–275). London, UK: Sage Publications.

Medina, J. J. (2008). The biology of recognition memory. *Psychiatric Times*, *25*(7), 13–15.

Meister, C., Pimentel, T., Haller, P., Jäger, L., Cotterell, R., & Levy, R. (2021). Revisiting the uniform information density hypothesis. arXiv preprint arXiv:2109.11635.

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some english consonants. *The Journal of the Acoustical Society of America*, *27*(2), 338–352.

Monsell, S., Doyle, M. C., & Haggard, P. N. (1989). Effects of frequency on visual word recognition tasks: Where are they? *Journal of Experimental Psychology: General*, *118*(1), 43–71.

Mortensen, D. R., Littell, P., Bharadwaj, A., Goyal, K., Dyer, C., & Levin, L. (2016). Panphon: A resource for mapping ipa segments to articulatory feature vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 3475–3484).

Nábělek, A. K. (1988). Identification of vowels in quiet, noise, and reverberation: Relationships with age and hearing loss. *The Journal of the Acoustical Society of America*, *84*(2), 476–484.

Nebes, R. D., Boller, F., & Holland, A. (1986). Use of semantic context by patients with alzheimer's disease. *Psychology and Aging*, *1*(3), 261–269.

Nessler, D., Mecklinger, A., & Penney, T. B. (2001). Event related brain potentials and illusory memories: the effects of differential encoding. *Cognitive Brain Research*, *10*(3), 283–301.

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. et al. (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*, *375*(1791), 20180522.

Nittrouer, S., & Boothroyd, A. (1990). Context effects in phoneme and word recognition by young children and older adults. *The Journal of the Acoustical Society of America*, *87*(6), 2705–2715.

Nittrouer, S., Wilhelmsen, M., Shapley, K., Bodily, K., & Creutz, T. (2003). Two reasons not to bring your children to cocktail parties. *The Journal of the Acoustical Society of America*, *113*(4), 2254–2254.

Norris, D., & McQueen, J. M. (2008). Shortlist b: a bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357–395.

Oden, G. C., & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review*, *85*(3), 172–191.

Olsthoorn, N. M., Andringa, S., & Hulstijn, J. H. (2014). Visual and auditory digit-span performance in native and non-native speakers. *International Journal of Bilingualism*, *18*(6), 663–673.

Parikh, G., & Loizou, P. C. (2005). The influence of noise on vowel and consonant cues. *The Journal of the Acoustical Society of America*, *118*(6), 3874–3888.

Peelle, J. E., Zhang, T., Patel, N., Rogers, C. S., & Van Engen, K. J. (2016). Online testing for assessing speech intelligibility. *The Journal of the Acoustical Society of America*, *140*(4), 3214–3214.

Perry, A. R., & Wingfield, A. (1994). Contextual encoding by young and elderly adults as revealed by cued and free recall. *Aging and Cognition*, *1*(2), 120–139.

Phatak, S. A., & Allen, J. B. (2007). Consonant and vowel confusions in speech-weighted noise. *The Journal of the Acoustical Society of America*, *121*(4), 2312–2326.

Phatak, S. A., Lovitt, A., & Allen, J. B. (2008). Consonant confusions in white noise. *The Journal of the Acoustical Society of America*, *124*(2), 1220–1233.

Pichora-Fuller, K. (2008). Use of supportive context by younger and older adult listeners: Balancing bottom-up and top-down information processing. *International Journal of Audiology*, *47*(sup2), S72–S82.

Pichora-Fuller, M. K., Alain, C., & Schneider, B. A. (2017). Older adults at the cocktail party. In J. C. Middlebrooks, J. Z. Simon, A. N. Popper, & R. R. Fay (Eds.), *The auditory system at the cocktail party* (pp. 227–259). Berlin, Germany: Springer.

Pichora-Fuller, M. K., Schneider, B. A., & Daneman, M. (1995). How young and old adults listen to and remember speech in noise. *The Journal of the Acoustical Society of America*, *97*(1), 593–608.

Pickett, J. (1957). Perception of vowels heard in noises of various spectra. *the Journal of the Acoustical Society of America*, *29*(5), 613–620.

Pollack, I. (1975). Auditory informational masking. *The Journal of the Acoustical Society of America*, *57*(S1), S5–S5.

Poppels, T., & Levy, R. (2016). Structure-sensitive noise inference: Comprehenders expect exchange errors. In *Proeedings of the 38th Annual Meeting of the Cognitive Society* (pp. 378–383).

Pusse, F., Sayeed, A., & Demberg, V. (2016). Lingoturk: managing crowdsourced tasks for psycholinguistics. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations* (pp. 57–61).

Quené, H., & Van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, *59*(4), 413–425.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. URL: `https://www.R-project.org/`.

Raven, J. (2000). The raven's progressive matrices: change and stability over culture and time. *Cognitive Psychology*, *41*(1), 1–48.

Rayner, K., Warren, T., Juhasz, B. J., & Liversedge, S. P. (2004). The effect of plausibility on eye movements in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(6), 1290–1301.

Riggs, K. M., Wingfield, A., & Tun, P. A. (1993). Passage difficulty, speech rate, and age differences in memory for spoken text: Speech recall and the complexity hypothesis. *Experimental Aging Research*, *19*(2), 111–128.

Roberts, A. C., Wetterlin, A., & Lahiri, A. (2013). Aligning mispronounced words to meaning: Evidence from erp and reaction time studies. *The Mental Lexicon*, *8*(2), 140–163.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(4), 803–814.

Roediger, H. L., Watson, J. M., McDermott, K. B., Gallo, D. A. et al. (2001). Factors that determine false recall: A multiple regression analysis. *Psychonomic Bulletin and Review*, *8*(3), 385–407.

Rogers, C. S. (2017). Semantic priming, not repetition priming, is to blame for false hearing. *Psychonomic Bulletin & Review*, *24*(4), 1194–1204.

Rogers, C. S., Jacoby, L. L., & Sommers, M. S. (2012). Frequent false hearing by older adults: the role of age differences in metacognition. *Psychology and Aging*, *27*(1), 33–45.

Rogers, C. S., & Wingfield, A. (2015). Stimulus-independent semantic bias misdirects word recognition in older adults. *The Journal of the Acoustical Society of America*, *138*(1), EL26–EL30.

Rommers, J., & Federmeier, K. D. (2018). Lingering expectations: A pseudo-repetition effect for words previously expected but not presented. *NeuroImage*, *183*, 263–272.

Ryskin, R., Futrell, R., Kiran, S., & Gibson, E. (2018). Comprehenders model the nature of noise in the environment. *Cognition*, *181*, 141–150.

Salame, P., & Baddeley, A. (1982). Disruption of short-term memory by unattended speech: Implications for the structure of working memory. *Journal of Verbal Learning and Verbal Behavior*, *21*(2), 150–164.

Salthouse, T. A. (1990). Working memory as a processing resource in cognitive aging. *Developmental Review*, *10*(1), 101–124.

Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review*, *103*(3), 403–428.

Schneider, B. A., Daneman, M., & Murphy, D. R. (2005). Speech comprehension difficulties in older adults: cognitive slowing or age-related changes in hearing? *Psychology and Aging*, *20*(2), 261–271.

Scholman, M., Marchal, M., & Demberg, V. (submitted). *Discourse Processes*.

Schuknecht, H. F., & Gacek, M. R. (1993). Cochlear pathology in presbycusis. *Annals of Otology, Rhinology & Laryngology*, *102*(1_suppl), 1–16.

Shannon, C. E. (1949). Communication in the presence of noise. *Proceedings of the IRE*, *37*(1), 10–21.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, *270*(5234), 303–304.

Sheldon, S., Pichora-Fuller, M. K., & Schneider, B. A. (2008). Priming and sentence context support listening to noise-vocoded speech by younger and older adults. *The Journal of the Acoustical Society of America*, *123*(1), 489–499.

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences*, *12*(5), 182–186.

Simpson, S. A., & Cooke, M. (2005). Consonant identification in n-talker babble is a nonmonotonic function of n. *The Journal of the Acoustical Society of America*, *118*(5), 2775–2778.

Slote, J., & Strand, J. F. (2016). Conducting spoken word recognition research online: Validation and a new timing method. *Behavior Research Methods*, *48*, 553–566.

Slowiaczek, L. M., Nusbaum, H. C., & Pisono, D. B. (1987). Phonological priming in auditory word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(1), 64–75.

Smith, N., & Levy, R. (2011). Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. volume 33.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.

Soli, S. D., & Arabie, P. (1979). Auditory versus phonetic accounts of observed confusions between consonant phonemes. *The Journal of the Acoustical Society of America*, *66*(1), 46–59.

Sommers, M. S., & Danielson, S. M. (1999). Inhibitory processes and spoken word recognition in young and older adults: the interaction of lexical competition and semantic context. *Psychology and Aging*, *14*(3), 458–472.

Sommers, M. S., & Lewis, B. P. (1999). Who really lives next door: Creating false memories with phonological neighbors. *Journal of Memory and Language*, *40*(1), 83–108.

Sommers, M. S., Morton, J., & Rogers, C. (2015). You are not listening to what i said: False hearing in young and older adults. In D. S. Lindsay, C. M. Kelley, A. P. Yonelinas, & H. L. Roediger III (Eds.), *Remembering: Attributions, Processes, and Control in Human Memory (Essays in Honor of Larry Jacoby)* (pp. 293–308). New York, NY: Psychology Press.

Staresina, B. P., Gray, J. C., & Davachi, L. (2009). Event congruency enhances episodic memory encoding through semantic elaboration and relational binding. *Cerebral Cortex*, *19*(5), 1198–1207.

Staub, A., Grant, M., Astheimer, L., & Cohen, A. (2015). The influence of cloze probability and item constraint on cloze task response time. *Journal of Memory and Language*, *82*, 1–17.

Staub, A., Rayner, K., Pollatsek, A., Hyönä, J., & Majewski, H. (2007). The time course of plausibility effects on eye movements in reading: evidence from noun-noun compounds. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(6), 1162–1169.

Steen-Baker, A. A., Ng, S., Payne, B. R., Anderson, C. J., Federmeier, K. D., & Stine-Morrow, E. A. (2017). The effects of context on processing words during sentence reading among adults varying in age and literacy skill. *Psychology and Aging*, *32*(5), 460–472.

Stine, E. A., & Wingfield, A. (1994). Older adults can inhibit high-probability competitors in speech recognition. *Aging and Cogntion*, *1*(2), 152–157.

Tabor, W., Galantucci, B., & Richardson, D. (2004). Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language*, *50*(4), 355–370.

Taitelbaum-Swead, R., & Fostick, L. (2016). The effect of age and type of noise on speech perception under conditions of changing context and noise levels. *Folia Phoniatrica et Logopaedica*, *68*(1), 16–21.

Taylor, W. L. (1953). "cloze procedure": A new tool for measuring readability. *Journalism Quarterly*, *30*(4), 415–433.

Traxler, M., & Gernsbacher, M. A. (2011). *Handbook of psycholinguistics*. Amsterdam, Netherlands: Elsevier.

Traxler, M. J. (2014). Trends in syntactic parsing: Anticipation, bayesian estimation, and good-enough parsing. *Trends in Cognitive Sciences*, *18*(11), 605–611.

Tucker-Drob, E. M., Brandmaier, A. M., & Lindenberger, U. (2019). Coupled cognitive changes in adulthood: A meta-analysis. *Psychological **b**ulletin*, *145*(3), 273–301.

Tun, P. A., Williams, V. A., Small, B. J., & Hafter, E. R. (2012). The effects of aging on auditory processing and cognition. *American Journal of Audiology*, *21*, 344–350.

Vaissière, J. (1983). Language-independent prosodic features. In A. Cutler, & D. R. Ladd (Eds.), *Prosody: Models and Measurements* (pp. 53–66). Heidelberg, Germany: Springer Verlag.

Van Berkum, J. J., 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), 443–467.

Van Engen, K. J., & Bradlow, A. R. (2007). Sentence recognition in native-and foreign-language multi-talker background noise. *The Journal of the Acoustical Society of America*, *121*(1), 519–526.

Van Os, M., Kray, J., & Demberg, V. (2021). Mishearing as a side effect of rational language comprehension in noise. *Frontiers in Psychology*, *12*, 679278.

Van Os, M., Kray, J., & Demberg, V. (2022). Rational speech comprehension: Interaction between predictability, acoustic signal, and noise. *Frontiers in Psychology*, *13*, 914239.

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and erp components. *International Journal of Psychophysiology*, *83*(2), 176–190.

Warren, T., & McConnell, K. (2007). Investigating effects of selectional restriction violations and plausibility violation severity on eye-movements in reading. *Psychonomic Bulletin & Review*, *14*(4), 770–775.

Weber, A., & Smits, R. (2003). Consonant and vowel confusion patterns by american english listeners. In *15th International Congress of Phonetic Sciences [ICPhS 2003]*.

Wingfield, A., Tun, P. A., & McCoy, S. L. (2005). Hearing loss in older adulthood: What it is and how it interacts with cognitive performance. *Current Directions in Psychological Science*, *14*(3), 144–148.

Wingfield, A., Tun, P. A., & Rosen, M. J. (1995). Age differences in veridical and reconstructive recall of syntactically and randomly segmented speech. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *50*(5), P257–P266.

Winter, B. (2019). *Statistics for linguists: An introduction using R*. Evanston, IL: Routledge.

Wolters, M. K., Isaac, K. B., & Renals, S. (2010). Evaluating speech synthesis intelligibility using amazon mechanical turk. In *Proceedings of the 7th ISCA Speech Synthesis Workshop* (pp. 136–141).

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of memory and language*, *46*(3), 441–517.

Yonelinas, A. P., Aly, M., Wang, W.-C., & Koen, J. D. (2010). Recollection and familiarity: Examining controversial assumptions and new directions. *Hippocampus*, *20*(11), 1178–1194.

You, H.-Y. (1979). *An acoustic and perceptual study of English fricatives*. Master's thesis University of Edmonton Edmonton, Canada.

Zwitserlood, P. (1989). The locus of the effects of sentential-semantic context in spoken-word processing. *Cognition*, *32*(1), 25–64.

Appendices

Appendix A

Stimuli Experiments 1 & 2

Table A.1 presents the stimuli that were used in Experiment 1 (Chapter 3) and Experiment 2 (Chapter 4). In Experiment 1 we used the entire set of items, while in Experiment 2 we used a subset: only the items with an item number *without* the prefix *F* were used there. The items are based on minimal pairs, which were the target words of the recognition task, and always occur in sentence-final position. Each item of a pair occurs once in a high predictability version, and once in a low predictability version, which was created by swapping the two sentence-final target words.

In the table we present the predictability condition (HP = high predictability; LP = low predictability) for each item. The item number additionally specifies which items go together. Versions A and B are the high predictable versions, while C and D are the low predictability swapped versions. Versions A and C have the same sentence context, as do B and D. For all high predictability items we give the cloze rating as obtained in our pretests (participant N = 10), and for all items we give the plausibility rating, as well as the speech sound contrast it contains.

Table A.1: Overview of the stimuli used in Experiment 1 and 2

pred.	item#	item	cloze	plaus.	contrast
HP	001A	Die Bauern verteilen zu viel Gülle auf den Äckern.	0.2	4.8	/ɐ/-/ə/
HP	001B	Auch wenn Jennifer sehr sorgfältig putzt, findet ihre Mutter immer noch Staub in den Ecken.	0.4	4.6	/ɐ/-/ə/
LP	001C	Die Bauern verteilen zu viel Gülle auf den Ecken.		2.4	/ɐ/-/ə/
LP	001D	Auch wenn Jennifer sehr sorgfältig putzt, findet ihre Mutter immer noch Staub in den Äckern.		1.7	/ɐ/-/ə/

Table A.1 continued from previous page

pred.	item#	item	cloze	plaus.	contrast
HP	002A	Das schlechte Wetter führte dazu, dass Elias und Karo ihre Pläne für den Feiertag änderten.	0.5	4.9	/ɐ/-/ə/
HP	002B	Bis vor kurzem war es noch so, dass beide Straßen hier endeten.	0.4	4.5	/ɐ/-/ə/
LP	002C	Das schlechte Wetter führte dazu, dass Elias und Karo ihre Pläne für den Feiertag endeten.		2.8	/ɐ/-/ə/
LP	002D	Bis vor kurzem war es noch so, dass beide Straßen hier änderten.		1.2	/ɐ/-/ə/
HP	003A	In der Disco bewegten sich die Freundinnen im Takt des Beats.	0.4	4.9	/i/-/ɪ/
HP	003B	Siegmund kennt sich mit Dateigrößen nicht so gut aus, und fragt nach dem Unterschied zwischen Bytes und Bits.	0.8	4.7	/i/-/ɪ/
LP	003C	In der Disco bewegten sich die Freundinnen im Takt des Bits.		1.4	/i/-/ɪ/
LP	003D	Siegmund kennt sich mit Datengrößen nicht so gut aus, und fragt nach dem Unterschied zwischen Bytes und Beats.		2.1	/i/-/ɪ/
HP	004A	Nach seinem Geburtstag wollte Samuel sich bei allen für die Geschenke bedanken.	1	4.9	/ɛ/-/a/
HP	004B	Bevor man sich dazu entscheidet umzuziehen, gibt es viel zu bedenken.	0.2	5	/ɛ/-/a/
LP	004C	Nach seinem Geburtstag wollte Samuel sich bei allen für die Geschenke bedenken.		2.3	/ɛ/-/a/
LP	004D	Bevor man sich dazu entscheidet umzuziehen, gibt es viel zu bedanken.		1.4	/ɛ/-/a/
HP	005A	Es gibt zunehmend mehr Überflutungen und Dürren, die der Klimawandel bedingt.	0	3.3	/d/-/ɡ/
HP	005B	Würdet ihr diesen Film auch dann herunterladen, wenn ihr damit eine Straftat begingt?	1	2.8	/d/-/ɡ/
LP	005C	Es gibt zunehmend mehr Überflutungen und Dürren, die der Klimawandel begingt.		1.4	/d/-/ɡ/
LP	005D	Würdet ihr diesen Film auch dann herunterladen, wenn ihr damit eine Straftat bedingt.		1.8	/d/-/ɡ/
HP	006A	Sofia wusste sofort, dass die Seide von guter Qualität war, als sie sie im Stoffgeschäft vorsichtig befühlte.	0.1	4.3	/y/-/ʏ/
HP	006B	Daniela naschte ein bisschen von der Schokolade, als sie den Adventskalender für ihre Kinder mit Süßigkeiten befüllte.	0.8	4.6	/y/-/ʏ/
LP	006C	Sofia wusste sofort, dass die Seide von guter Qualität war, als sie sie im Stoffgeschäft vorsichtig befüllte.		1.7	/y/-/ʏ/

Table A.1 continued from previous page

pred.	item#	item	cloze	plaus.	contrast
LP	006D	Daniela naschte ein bisschen von der Schokolade, als sie den Adventskalender für ihre Kinder mit Süßigkeiten befühlte.		2.3	/y/-/ʏ/
HP	007A	Der örtliche Kulturverein hat die alte Tradition des Erntedankumzugs wieder belebt.	0.2	4.8	/p/-/k/
HP	007B	Als Christian den Proviant vorbereitet, möchte seine Mutter wissen, mit was er die Brote belegt.	0.5	4.5	/p/-/k/
LP	007C	Der örtliche Kulturverein hat die alte Tradition des Erntedankumzugs wieder belegt.		2.1	/p/-/k/
LP	007D	Als Christian den Proviant vorbereitet, möchte seine Mutter wissen, mit was er die Brote belebt.		1.1	/p/-/k/
HP	008A	Nach dem Brand vor zwei Jahren, hatte die Versicherung nur ungefähr schätzen können, auf welchen Wert sich die Schäden beliefen.	0.3	4.8	/ɐ/-/ə/
HP	008B	Aufgrund der schlechten Kartoffelernte konnte der Bauer nicht mehr wie gewohnt alle Gemüseläden beliefern.	0.8	4.9	/ɐ/-/ə/
LP	008C	Nach dem Brand vor zwei Jahren, hatte die Versicherung nur ungefähr schätzen können, auf welchen Wert sich die Schäden beliefern.		2.8	/ɐ/-/ə/
LP	008D	Aufgrund der schlechten Kartoffelernte konnte der Bauer nicht mehr wie gewohnt alle Gemüseläden beliefen.		3.6	/ɐ/-/ə/
HP	009A	Für seine erste Fahrt hatte der Kapitän bereits zwei Matrosen gefunden und er suchte noch den Rest der Besatzung.	0.2	4.3	/ɛ/-/a/
HP	009B	Da mehrere Schauspieler erkrankt waren, benötigte der Regisseur eine neue Besetzung.	0.4	4.9	/ɛ/-/a/
LP	009C	Für seine erste Fahrt hatte der Kapitän bereits zwei Matrosen gefunden und er suchte noch den Rest der Besetzung.		3.1	/ɛ/-/a/
LP	009D	Da mehrere Schauspieler erkrankt waren, benötigte der Regisseur eine neue Besatzung.		3.1	/ɛ/-/a/
HP	010A	Da das Land niemandem gehörte, konnten die Einwanderer es direkt besiedeln.	0.5	5	/d/-/g/
HP	010B	Lass uns unsere Freundschaft mit einem Handschlag besiegeln.	0.6	4.7	/d/-/g/
LP	010C	Da das Land niemandem gehörte, konnten die Einwanderer es direkt besiegeln.		1.4	/d/-/g/
LP	010D	Lass uns unsere Freundschaft mit einem Handschlag besiedeln.		2.3	/d/-/g/
HP	011A	Wenn Marie den antiken Holztisch ersteigern möchte, muss sie bereit sein, sehr hoch zu bieten.	0.9	4.3	/i/-/ɪ/

Table A.1 continued from previous page

pred.	item#	item	cloze	plaus.	contrast
HP	011B	Damit Marie sich ein neues Handy kaufen konnte, musste sie ihren Vater um Geld bitten.	0.9	5	/i/-/ɪ/
LP	011C	Wenn Marie den antiken Holztisch ersteigern möchte, muss sie bereit sein, sehr hoch zu bitten.		2.3	/i/-/ɪ/
LP	011D	Damit Marie sich ein neues Handy kaufen konnte, musste sie ihren Vater um Geld bieten.		2	/i/-/ɪ/
HP	012A	Beim Skatspielen bekam Thomas meistens nur Damen und Könige und keine Buben.	0.3	4.3	/b/-/d/
HP	012B	Auf dem Jahrmarkt fahren die Kinder mit den Karussellen und spielen Entenfischen an den Buden.	0.4	4.8	/b/-/d/
LP	012C	Beim Skatspielen bekam Thomas meistens nur Damen und Könige und keine Buden.		1.3	/b/-/d/
LP	012D	Auf dem Jahrmarkt fahren die Kinder mit den Karussellen und spielen Entenfischen an den Buben.		1.3	/b/-/d/
HP	013A	Nachdem Janik seine Freunde länger nicht gesehen hatte, freute er sich auf das Treffen mit seiner Clique.	0.4	5	/p/-/k/
HP	013B	Im letzten Sommerurlaub kletterte Janik am Meer auf eine hohe Klippe.	0.5	4.4	/p/-/k/
LP	013C	Nachdem Janik seine Freunde länger nicht gesehen hatte, freute er sich auf das Treffen mit seiner Klippe.		1.3	/p/-/k/
LP	013D	Im letzten Sommerurlaub kletterte Janik am Meer auf eine hohe Clique.		1.1	/p/-/k/
HP	014A	Da Frau Klein Hunde liebt, steht auf dem Armaturenbrett in ihrem Auto ein kleiner mit dem Kopf wackelnder Dackel.	0.5	4.8	/t/-/k/
HP	014B	Am Probierstand des orientalischen Basars entschied Mona sich nach kurzer Überlegung gegen die Feige und für die Dattel.	0.4	4.5	/t/-/k/
LP	014C	Da Frau Klein Hunde liebt, steht auf dem Armaturenbrett in ihrem Auto ein kleiner mit dem Kopf wackelnder Dattel.		2.7	/t/-/k/
LP	014D	Am Probierstand des orientalischen Basars entschied Mona sich nach kurzer Überlegung gegen die Feige und für die Dackel.		1.5	/t/-/k/
HP	015A	Zu den beliebtesten Jagdhunden gehört trotz kurzer Beine die Rasse der Dackel.	0.8	4.3	/ɛ/-/a/
HP	015B	In der Haushaltswarenabteilung des Kaufhauses suchte Lara nach einem Topf mit Deckel.	1	5	/ɛ/-/a/
LP	015C	Zu den beliebtesten Jagdhunden gehört trotz kurzer Beine die Rasse der Deckel.		1.9	/ɛ/-/a/
LP	015D	In der Haushaltswarenabteilung des Kaufhauses suchte Lara nach einem Topf mit Dackel.		1.7	/ɛ/-/a/

Table A.1 continued from previous page

pred.	item#	item	cloze	plaus.	contrast
HP	016A	Da es in der Hütte kalt war, nahmen sich alle ein paar warme Decken.	0.6	5	/p/-/k/
HP	016B	Da sie beim Schrankaufbau alles falsch gemacht haben, bezeichnete Robin seine Brüder als Deppen.	0	4.5	/p/-/k/
LP	016C	Da es in der Hütte kalt war, nahmen sich alle ein paar warme Deppen.		1.1	/p/-/k/
LP	016D	Da sie beim Schrankaufbau alles falsch gemacht haben, bezeichnete Robin seine Brüder als Decken.		1.2	/p/-/k/
HP	017A	Da der Baum morsch war, ließ der Förster ihn fällen.	0.9	4.9	/ɛ/-/a/
HP	017B	Vor Schreck ließ die kleine Pia ihr Eis fallen.	0.9	5	/ɛ/-/a/
LP	017C	Da der Baum morsch war, ließ der Förster ihn fallen.		3	/ɛ/-/a/
LP	017D	Vor Schreck ließ die kleine Pia ihr Eis fällen.		1.4	/ɛ/-/a/
HP	018A	Herr Pfeifer möchte sein altes Bauernhaus von Schädlingen befreien und findet fast jeden Morgen Mäuse in den aufgestellten Fallen.	1	4.2	/ɛ/-/a/
HP	018B	Die Wilderer machen Pelze aus den weichen Fellen.	0.7	4.9	/ɛ/-/a/
LP	018C	Herr Pfeifer möchte sein altes Bauernhaus von Schädlingen befreien und findet fast jeden Morgen Mäuse in den aufgestellten Fellen.		1.7	/ɛ/-/a/
LP	018D	Die Wilderer machen Pelze aus den weichen Fallen.		1	/ɛ/-/a/
HP	019A	Beinahe wäre Christopher auf den unebenen Pflastersteinen gestolpert, zum Glück aber nur fast.	0.3	4.7	/ɛ/-/a/
HP	019B	Beim Abschied umarmte Christopher seine Freundin fest.	0.4	4.9	/ɛ/-/a/
LP	019C	Beinahe wäre Christopher auf den unebenen Pflastersteinen gestolpert, zum Glück aber nur fest.		1.5	/ɛ/-/a/
LP	019D	Beim Abschied umarmte Christopher seine Freundin fast.		2.1	/ɛ/-/a/
HP	020A	Der Vogel fiel sofort auf mit seinen leuchtend roten Federn.	0.6	4.5	/ɐ/-/ə/
HP	020B	Ein Familienstreit um das Erbe endete früher oft in jahrzehntelangen, blutigen Fehden.	0.2	4.4	/ɐ/-/ə/
LP	020C	Der Vogel fiel sofort auf mit seinen leuchtend roten Fehden.		1.7	/ɐ/-/ə/
LP	020D	Ein Familienstreit um das Erbe endete früher oft in jahrzehntelangen, blutigen Federn.		1.3	/ɐ/-/ə/
HP	021A	Da ihre Lieblingsbluse einen Riss bekommen hat, hofft Melanie, dass ihre Mutter sie ihr wieder flickt.	0.3	4.4	/i/-/ɪ/
HP	021B	Melanie beobachtet den Hubschrauber beim Starten und fragt sich, wohin er wohl fliegt.	0.7	4.7	/i/-/ɪ/
LP	021C	Da ihre Lieblingsbluse einen Riss bekommen hat, hofft Melanie, dass ihre Mutter sie ihr wieder fliegt.		1.3	/i/-/ɪ/

Table A.1 continued from previous page

pred.	item#	item	cloze	plaus.	contrast
LP	021D	Melanie beobachtet den Hubschrauber beim Starten und fragt sich, wohin er wohl flickt.		1.1	/i/-/ɪ/
HP	022A	Als winterliche Deko für die Fenster bastelte Tim eine Flocke.	0.5	4.6	/t/-/k/
HP	022B	Die Engländer gewannen viele Seeschlachten, aufgrund ihrer großen Flotte.	0.6	4.5	/t/-/k/
LP	022C	Als winterliche Deko für die Fenster bastelte Tim eine Flotte.		1.8	/t/-/k/
LP	022D	Die Engländer gewannen viele Seeschlachten, aufgrund ihrer großen Flocke.		1.2	/t/-/k/
HP	023A	Früher hat Aaron viel mehr auf sein Aussehen geachtet.	0.8	5	/ɛ/-/a/
HP	023B	Nach dem Vietnamkrieg wurden chemische und biologische Waffen weltweit geächtet.	0.1	4.3	/ɛ/-/a/
LP	023C	Früher hat Aaron viel mehr auf sein Aussehen geächtet.		1.3	/ɛ/-/a/
LP	023D	Nach dem Vietnamkrieg wurden chemische und biologische Waffen weltweit geachtet.		2.9	/ɛ/-/a/
HP	024A	Letzte Woche haben die Rosen im Garten noch alle geblüht.	0.9	4.5	/b/-/g/
HP	024B	Nachdem die Flammen erloschen waren, haben die Kohlen noch lange geglüht.	1	4.5	/b/-/g/
LP	024C	Letzte Woche haben die Rosen im Garten noch alle geglüht.		1.6	/b/-/g/
LP	024D	Nachdem die Flammen erloschen waren, haben die Kohlen noch lange geblüht.		1.2	/b/-/g/
HP	025A	An der Wand des Urlaubshauses von Anastasia in Indonesien sonnten sich zwei bunte Geckos.	0	4.7	/t/-/k/
HP	025B	Vor den Toren vieler Großstädte leben sehr arme Menschen in Gettos.	0.3	3.4	/t/-/k/
LP	025C	An der Wand des Urlaubshauses von Anastasia in Indonesien sonnten sich zwei bunte Gettos.		1.3	/t/-/k/
LP	025D	Vor den Toren vieler Großstädte leben sehr arme Menschen in Geckos.		1	/t/-/k/
HP	026A	Auf dem Heimweg war Johannes ganz versunken in seine Gedanken.	0.6	4.5	/ɛ/-/a/
HP	026B	Am 100. Geburtstag des Malers gab es in seinem Heimatort eine Feier zu seinem Gedenken.	0.3	4	/ɛ/-/a/
LP	026C	Auf dem Heimweg war Johannes ganz versunken in seine Gedenken.		2.4	/ɛ/-/a/
LP	026D	Am 100. Geburtstag des Malers gab es in seinem Heimatort eine Feier zu seinem Gedanken.		1.4	/ɛ/-/a/
HP	027A	David hat schon lange keine Kopfschmerzen mehr gehabt.	1	5	/p/-/k/

Table A.1 continued from previous page

pred.	item#	item	cloze	plaus.	contrast
HP	027B	Bevor der Koch die Nüsse über das Essen gestreut hat, hat er sie grob gehackt.	0.5	4.8	/p/-/k/
LP	027C	David hat schon lange keine Kopfschmerzen mehr gehackt.		1.2	/p/-/k/
LP	027D	Bevor der Koch die Nüsse über das Essen gestreut hat, hat er sie grob gehabt.		2.4	/p/-/k/
HP	028A	Vor ihrem Einbruch haben die Diebe die Telefonleitung gekappt.	0.7	4.8	/t/-/k/
HP	028B	Bei den Ermittlungen zu dem Einbruch hat die Polizei lange im Dunklen getappt.	0.9	4.5	/t/-/k/
LP	028C	Vor ihrem Einbruch haben die Diebe die Telefonleitung getappt.		1.7	/t/-/k/
LP	028D	Bei den Ermittlungen zu dem Einbruch ist die Polizei lange im Dunklen gekappt.		1.5	/t/-/k/
HP	029A	Aus Nervosität hat Luise während der Klausur auf ihrem Stift gekaut.	0.9	4.8	/t/-/k/
HP	029B	Eislaufen auf dem Weiher war am nächsten Morgen nicht mehr möglich, denn aufgrund der milden Temperaturen hat es über Nacht getaut.	0.8	4.6	/t/-/k/
LP	029C	Aus Nervosität hat Maria während der Klausur auf ihrem Stift getaut.		1.1	/t/-/k/
LP	029D	Eislaufen auf dem Weiher war am nächsten Morgen nicht mehr möglich, denn aufgrund der milden Temperaturen hat es über Nacht gekaut.		1.1	/t/-/k/
HP	030A	Beim Fernsehen hat Petra sich an ihren Freund gekuschelt.	0.7	5	/t/-/k/
HP	030B	Während des Unterrichts haben die Freundinnen die ganze Zeit leise miteinander getuschelt.	0.4	4.9	/t/-/k/
LP	030C	Beim Fernsehen hat Petra sich an ihren Freund getuschelt.		1.1	/t/-/k/
LP	030D	Während des Unterrichts haben die Freundinnen die ganze Zeit leise miteinander gekuschelt.		2.8	/t/-/k/
HP	031A	Da die Holzklötze zu groß für den Kamin waren, hat Herbert sie mit einer Axt in zwei Teile gespaltet.	0.6	4.1	/p/-/t/
HP	031B	Die Organisatoren des Klassentreffens haben den Abend sehr abwechslungsreich gestaltet.	1	5	/p/-/t/
LP	031C	Da die Holzklötze zu groß für den Kamin waren, hat Herbert sie mit einer Axt in zwei Teile gestaltet.		1.7	/p/-/t/
LP	031D	Die Organisatoren des Klassentreffens haben den Abend sehr abwechslungsreich gespaltet.		1.7	/p/-/t/
HP	032A	Während der Klassenarbeiten wird gerne mal gespickt.	0.4	5	/p/-/t/
HP	032B	Früher hat Oma auf die feinen Tischdecken normalerweise Verzierungen gestickt.	0.4	4.2	/p/-/t/

Table A.1 continued from previous page

pred.	item#	item	cloze	plaus.	contrast
LP	032C	Während Klassenarbeiten wird gerne mal gestickt.		1.4	/p/-/t/
LP	032D	Früher hat Oma auf die feinen Tischdecken normalerweise Verzierungen gespickt.		2.4	/p/-/t/
HP	033A	Bevor Sonja das Haus verlassen hat, hatte sie ihr Handy in die Tasche gesteckt.	0.5	4.5	/p/-/k/
HP	033B	Bei der Abschlussfeier hat nach Mitternacht der Bär gesteppt.	0.9	4.2	/p/-/k/
LP	033C	Bevor Sonja das Haus verlassen hat, hatte sie ihr Handy in die Tasche gesteppt.		1	/p/-/k/
LP	033D	Bei der Abschlussfeier hat nach Mitternacht der Bär gesteckt.		1.5	/p/-/k/
HP	034A	Die Uhr hat die ganze Nacht durch getickt.	0.7	4.3	/p/-/k/
HP	034B	Um Jans Aufmerksamkeit zu bekommen, hat Paulina ihm auf die Schulter getippt.	0.4	4.9	/p/-/k/
LP	034C	Die Uhr hat die ganze Nacht durch getippt.		1.4	/p/-/k/
LP	034D	Um Jans Aufmerksamkeit zu bekommen, hat Paulina ihm auf die Schulter getickt.		2.6	/p/-/k/
HP	035A	Da die Metallplatte sonst zu dick gewesen wäre, wurde sie vor der Weiterverarbeitung gewalzt.	0.3	5	/ɛ/-/a/
HP	035B	Da Frederike vor Aufregung nicht schlafen konnte, hat sie sich die ganze Nacht im Bett hin und her gewälzt.	0.7	5	/ɛ/-/a/
LP	035C	Da die Metallplatte sonst zu dick gewesen wäre, wurde sie vor der Weiterverarbeitung gewälzt.		3.5	/ɛ/-/a/
LP	035D	Da Frederike vor Aufregung nicht schlafen konnte, hat sie sich die ganze Nacht im Bett hin und her gewalzt.		2.9	/ɛ/-/a/
HP	036A	Da die neuen Auflagen es Moritz schwer machen, Gewinn zu machen, hat er heftig gegen die Politiker gewettert.	0.2	4.6	/ɐ/-/ə/
HP	036B	Da Moritz sich sehr sicher ist, wer das Pferderennen gewinnen wird, hat er mit seinem Freund um Geld gewettet.	1	4.8	/ɐ/-/ə/
LP	036C	Da die neuen Auflagen es Moritz schwer machen, Gewinn zu machen, hat er heftig gegen die Politiker gewettet.		2.2	/ɐ/-/ə/
LP	036D	Da Moritz sich sehr sicher ist, wer das Pferderennen gewinnen wird, hat er mit seinem Freund um Geld gewettert.		2.3	/ɐ/-/ə/
HP	037A	Den Teppich hat Linda selber aus Wolle gewoben.	0.2	4.6	/b/-/g/
HP	037B	Da Rosa nicht abschätzen konnte, wie viel Reis noch in dem Päckchen war, hat sie es gewogen.	0.7	4.3	/b/-/g/
LP	037C	Den Teppich hat Linda selber aus Wolle gewogen.		1.3	/b/-/g/
LP	037D	Da Rosa nicht abschätzen konnte, wie viel Reis noch in dem Päckchen war, hat sie es gewoben.		1.1	/b/-/g/

Table A.1 continued from previous page

pred.	item#	item	cloze	plaus.	contrast
HP	038A	Alex und Franziska besichtigten am letzten Tag in Paris den Invalidendom mit Napoleons Grab.	0.4	3.6	/p/-/t/
HP	038B	Bei seinem Weg von einem Tal zum nächsten wanderte der Bergsteiger auf einem schmalen Grad.	0.3	3.5	/p/-/t/
LP	038C	Alex und Franziska besichtigten am letzten Tag in Paris den Invalidendom mit Napoleons Grat.		1.7	/p/-/t/
LP	038D	Bei seinem Weg von einem Tal zum nächsten wanderte der Bergsteiger auf einem schmalen Grab.		1.5	/p/-/t/
HP	039A	Zum Auflockern des Bodens und der Beete benutzen Gärtner für gewöhnlich Hacken.	0.2	3.3	/p/-/k/
HP	039B	Da Matthias nicht viel Hunger hatte, sagte er zu seiner Mutter: Gib mir von dem Fleisch nur einen Happen.	0.1	4.6	/p/-/k/
LP	039C	Zum Auflockern des Bodens und der Beete benutzen Gärtner für gewöhnlich Happen.		1.6	/p/-/k/
LP	039D	Da Matthias nicht viel Hunger hatte, sagte er zu seiner Mutter: Gib mir von dem Fleisch nur einen Hacken.		1.4	/p/-/k/
HP	040A	Eine Frau im Supermarkt schimpfte, denn das Kleinkind fuhr ihr an der Kasse mit dem Einkaufswagen ständig in die Hacken.	0.3	4.3	/ɛ/-/a/
HP	040B	Um in den Nachbarsgarten zu gelangen, kletterten die Kinder durch die Hecken.	0.5	4.6	/ɛ/-/a/
LP	040C	Eine Frau im Supermarkt schimpfte, denn das Kleinkind fuhr ihr an der Kasse mit dem Einkaufswagen ständig in die Hecken.		1.3	/ɛ/-/a/
LP	040D	Um in den Nachbarsgarten zu gelangen, kletterten die Kinder durch die Hacken.		1.2	/ɛ/-/a/
HP	041A	Wenn ein Kind etwas zerstört, müssen die Eltern dafür haften.	0.3	5	/ɛ/-/a/
HP	041B	Damit keines der Blätter verloren geht, sollen die Schüler sie aneinander heften.	0.5	4.6	/ɛ/-/a/
LP	041C	Wenn ein Kind etwas zerstört, müssen die Eltern dafür heften.		1.3	/ɛ/-/a/
LP	041D	Damit keines der Blätter verloren geht, sollen die Schüler sie aneinander haften.		2.2	/ɛ/-/a/
HP	042A	Die Müllabfuhr transportiert die Abfälle auf verschiedene Halden.	0.3	4.6	/ɛ/-/a/
HP	042B	Stefan mochte schon immer spannende Filme mit bösen Schurken und mutigen Helden.	1	4.9	/ɛ/-/a/
LP	042C	Die Müllabfuhr transportiert die Abfälle auf verschiedene Helden.		1.2	/ɛ/-/a/
LP	042D	Stefan mochte schon immer spannende Filme mit bösen Schurken und mutigen Halden.		1.2	/ɛ/-/a/

Table A.1 continued from previous page

pred.	item#	item	cloze	plaus.	contrast
HP	043A	Die Schafe hatten die Wiese leer gefressen bis auf den letzten Halm.	0.5	4.5	/ɛ/-/a/
HP	043B	Beim Fahrradfahren trägt Frederik immer einen Helm.	1	5	/ɛ/-/a/
LP	043C	Die Schafe hatten die Wiese leer gefressen bis auf den letzten Helm.		1.4	/ɛ/-/a/
LP	043D	Beim Fahrradfahren trägt Frederik immer einen Halm.		2.1	/ɛ/-/a/
HP	044A	Auf der Tombola wurden viele Kleinigkeiten verlost, neben den drei großen Hauptgewinnen.	0.3	4.3	/ɐ/-/ə/
HP	044B	Die Glücksfee der Lotterie gratulierte den Hauptgewinnern.	0.9	4.4	/ɐ/-/ə/
LP	044C	Auf der Tombola wurden viele Kleinigkeiten verlost, neben den drei großen Hauptgewinnern.		1.7	/ɐ/-/ə/
LP	044D	Die Glücksfee der Lotterie gratulierte den Hauptgewinnen.		2.7	/ɐ/-/ə/
HP	045A	Der Arzt rät davon ab, dass der Patient in den nächsten Wochen Sport treibt oder schwere Gegenstände hebt.	1	5	/p/-/k/
HP	045B	Durch Patricks Verhalten wird Susanne schnell klar, dass Patrick auch heute noch einen Groll gegen sie hegt.	0.7	4.9	/p/-/k/
LP	045C	Der Arzt rät davon ab, dass der Patient in den nächsten Wochen Sport treibt oder schwere Gegenstände hegt.		1.3	/p/-/k/
LP	045D	Durch Patricks Verhalten wird Susanne schnell klar, dass Patrick auch heute noch einen Groll gegen sie hebt.		2.9	/p/-/k/
HP	046A	Es war Jonas unangenehm, dass er nicht mehr wusste, wie seine neuen Kollegen hießen.	0.8	4.9	/i/-/ɪ/
HP	046B	Als sie das offene Meer erreichten, begann die Mannschaft, die Segel zu hissen.	0.3	4.7	/i/-/ɪ/
LP	046C	Es war Jonas unangenehm, dass er nicht mehr wusste, wie seine neuen Kollegen hissen.		1.4	/i/-/ɪ/
LP	046D	Als sie das offene Meer erreichten, begann die Mannschaft, die Segel zu hießen.		2.9	/i/-/ɪ/
HP	047A	An Sandras Geburtstag ließen ihre Freunde sie hochleben.	0.3	4.6	/b/-/g/
HP	047B	Da Sandra sich beim Sport den Knöchel verletzt hatte, sollte sie den Fuß schonen und möglichst oft hochlegen.	0.3	5	/b/-/g/
LP	047C	An Sandras Geburtstag ließen ihre Freunde sie hochlegen.		1.2	/b/-/g/
LP	047D	Da Sandra sich beim Sport den Knöchel verletzt hatte, sollte sie den Fuß schonen und möglichst oft hochleben.		1.1	/b/-/g/
HP	048A	Bei dem Ausflug zum Bauernhof besuchten die Kinder zuerst den Kuhstall und David streichelte über das weiche Fell des neugeborenen Kalbs.	0.6	5	/p/-/k/

Table A.1 continued from previous page

pred.	item#	item	cloze	plaus.	contrast
HP	048B	Das Leitungswasser in Lauras Wohnort enthielt viele Mineralien und immer, wenn sie Wasser kochte, sah sie im Topf die Rückstände des Kalks.	0.4	4.8	/p/-/k/
LP	048C	Bei dem Ausflug zum Bauernhof besuchten die Kinder zuerst den Kuhstall und David streichelte über das weiche Fell des neugeborenen Kalks.		1.6	/p/-/k/
LP	048D	Das Leitungswasser in Annas Wohnort enthielt viele Mineralien und immer, wenn sie Wasser kochte, sah sie im Topf die Rückstände des Kalbs.		1.5	/p/-/k/
HP	049A	Die Kinder trafen sich auf dem Fußballplatz um ein bisschen zusammen zu kicken.	0.4	4.7	/p/-/k/
HP	049B	Nach dem Putzen musste Lena nur noch das dreckige Wasser aus dem Eimer kippen.	0.5	4.9	/p/-/k/
LP	049C	Die Kinder trafen sich auf dem Fußballplatz um ein bisschen zusammen zu kippen.		1.1	/p/-/k/
LP	049D	Nach dem Putzen musste Lena nur noch das dreckige Wasser aus dem Eimer kicken.		1.3	/p/-/k/
HP	050A	Nachdem Jonas die ganze Zeit mit dem Stuhl hin und her geschaukelt hatte, brachte er den Stuhl schließlich zum Kippen.	0.7	4.4	/p/-/t/
HP	050B	Der Fensterrahmen hatte einen Spalt bekommen, daher suchte Benjamin etwas zum Kitten.	0.1	3.5	/p/-/t/
LP	050C	Nachdem Jonas die ganze Zeit mit dem Stuhl hin und her geschaukelt hatte, brachte er den Stuhl schließlich zum Kitten.		1.5	/p/-/t/
LP	050D	Der Fensterrahmen hatte einen Spalt bekommen, daher suchte Benjamin etwas zum Kippen.		1.9	/p/-/t/
HP	051A	Da Erik sehr viel gelernt hatte, war die Klassenarbeit für ihn ein Klacks.	0.3	4.9	/p/-/k/
HP	051B	Da das Pferd nicht weiter gehen wollte, gab Erik ihm einen leichten Klaps.	0.5	4.6	/p/-/k/
LP	051C	Da Erik sehr viel gelernt hatte, war die Klassenarbeit für ihn ein Klaps.		1.9	/p/-/k/
LP	051D	Da das Pferd nicht weiter gehen wollte, gab Erik ihm einen leichten Klacks.		1.6	/p/-/k/
HP	052A	Als der Arzt den Ärmel hochschob, um den Arm des Patienten zu untersuchen, sah er, dass unter dem Stoff eine große Wunde klaffte.	0.2	4.7	/ɛ/-/a/
HP	052B	Paulina erschreckte sich, als der kleine Hund plötzlich laut und schrill kläffte.	0	4.8	/ɛ/-/a/
LP	052C	Als der Arzt den Ärmel hochschob, um den Arm des Patienten zu untersuchen, sah er, dass unter dem Stoff eine große Wunde kläffte.		2.4	/ɛ/-/a/

Table A.1 continued from previous page

pred.	item#	item	cloze	plaus.	contrast
LP	052D	Paulina erschreckte sich, als der kleine Hund plötzlich laut und schrill klaffte.		2.3	/ɛ/-/a/
HP	053A	Gerade hat der Neustart nicht funktioniert, aber beim nächsten Mal sollte es klappen.	0.6	5	/ɐ/-/ə/
HP	053B	Mir ist so kalt, dass meine Zähne klappern.	0.8	4.6	/ɐ/-/ə/
LP	053C	Gerade hat der Neustart nicht funktioniert, aber beim nächsten Mal sollte es klappern.		2.5	/ɐ/-/ə/
LP	053D	Mir ist so kalt, dass meine Zähne klappen.		1.7	/ɐ/-/ə/
HP	054A	Wenn du das Buch zuschlägst, sei vorsichtig, dass du dabei nicht die Ecken der Seiten knickst.	0.8	4.9	/p/-/k/
HP	054B	Da du die bessere Kamera hast, ist es besser, wenn du das knipst.	0	3.9	/p/-/k/
LP	054C	Wenn du das Buch zuschlägst, sei vorsichtig, dass du dabei nicht die Ecken der Seiten knipst.		1.4	/p/-/k/
LP	054D	Da du die bessere Kamera hast, ist es besser, wenn du das knickst.		1.2	/p/-/k/
HP	055A	Die Tapferkeitsmedaille verlieh man nur den Kühnsten.	0	4.6	/y/-/ʏ/
HP	055B	Malerei, Bildhauerei und Fotografie zählt man zu den Bildenden Künsten.	0.9	4.6	/y/-/ʏ/
LP	055C	Die Tapferkeitsmedaille verlieh man nur den Künsten.		1.9	/y/-/ʏ/
LP	055D	Malerei, Bildhauerei und Fotografie zählt man zu den Bildenden Kühnsten.		1.6	/y/-/ʏ/
HP	056A	Nach einem langen Marsch den Berg hinauf gelangten die Spaziergänger schließlich zur schneebedeckten Kuppe.	0.1	4.1	/p/-/t/
HP	056B	Im Kloster tragen die Mönche alle eine Kutte.	0.5	4.6	/p/-/t/
LP	056C	Nach einem langen Marsch den Berg hinauf gelangten die Spaziergänger schließlich zur schneebedeckten Kutte.		1.4	/p/-/t/
LP	056D	Im Kloster tragen die Mönche alle eine Kuppe.		1.4	/p/-/t/
HP	057A	Ben sieht zu, wie sich die Ziegen an der Milch laben.	0.1	3.5	/b/-/d/
HP	057B	Bevor Ben mit seinem Handy wieder telefonieren kann, muss er es laden.	0.8	4.9	/b/-/d/
LP	057C	Ben sieht zu, wie sich die Ziegen an der Milch laden.		1.2	/b/-/d/
LP	057D	Bevor Ben mit seinem Handy wieder telefonieren kann, muss er es laben.		1.4	/b/-/d/
HP	058A	Zum Tafelwischen benutzt die Lehrerin immer nasse Schwämme oder feuchte Lappen.	1	4.9	/p/-/t/
HP	058B	Um den Garten von dem Nachbarsgrundstück abzugrenzen, baut Bruno einen Holzzaun aus weißen Latten.	0.6	5	/p/-/t/
LP	058C	Zum Tafelwischen benutzt die Lehrerin immer nasse Schwämme oder feuchte Latten.		1.1	/p/-/t/

Table A.1 continued from previous page

pred.	item#	item	cloze	plaus.	contrast
LP	058D	Um den Garten von dem Nachbarsgrundstück abzugrenzen, baut Bruno einen Holzzaun aus weißen Lappen.		1.4	/p/-/t/
HP	059A	Beim Fußball traf Peter zur Enttäuschung aller beim Elfmeter mal wieder an die Latte.	0.8	3.8	/ɛ/-/a/
HP	059B	Peters neuer Nachbar kommt aus dem Baltikum und ist ein Lette.	0.2	4.4	/ɛ/-/a/
LP	059C	Beim Fußball traf Peter zur Enttäuschung aller beim Elfmeter mal wieder an die Lette.		1.7	/ɛ/-/a/
LP	059D	Peters neuer Nachbar kommt aus dem Baltikum und ist ein Latte.		2.2	/ɛ/-/a/
HP	060A	Die beiden Verliebten trafen sich heimlich im Garten in der Laube.	0.3	4.5	/b/-/g/
HP	060B	Tina schwenkt die Brezeln vor dem Backen durch einen Topf mit Lauge.	0.5	3.9	/b/-/g/
LP	060C	Die beiden Verliebten trafen sich heimlich im Garten in der Lauge.		1.7	/b/-/g/
LP	060D	Tina schwenkt die Brezeln vor dem Backen durch einen Topf mit Laube.		2	/b/-/g/
HP	061A	Michael fragte mich erstaunt, ob ich wirklich in einer Villa lebe.	0.4	4.9	/b/-/g/
HP	061B	Ich schüttele immer zuerst meine Kissen aus, bevor ich mich ins Bett lege.	0.8	4.7	/b/-/g/
LP	061C	Michael fragte mich erstaunt, ob ich wirklich in einer Villa lege.		1.2	/b/-/g/
LP	061D	Ich schüttele immer zuerst meine Kissen aus, bevor ich mich ins Bett lebe.		1.5	/b/-/g/
HP	062A	Nach vier Jahren heiratete Paul seine große Liebe.	1	5	/b/-/g/
HP	062B	Am Pool im Hotel gab es nur noch eine freie Liege.	0.7	4.7	/b/-/g/
LP	062C	Nach vier Jahren heiratete Paul seine große Liege.		1.2	/b/-/g/
LP	062D	Am Pool im Hotel gab es nur noch eine freie Liebe.		1.1	/b/-/g/
HP	063A	Die Kinder haben sich bei der Familienfeier so gut benommen, dass alle Gäste sie loben.	0.7	4.3	/b/-/g/
HP	063B	Peter und Klaus gaben beim Verhör an, dass sie unschuldig seien und von nichts wüssten, aber die Richterin vermutete, dass sie logen.	0.3	4	/b/-/g/
LP	063C	Die Kinder haben sich bei der Familienfeier so gut benommen, dass alle Gäste sie logen.		1.2	/b/-/g/
LP	063D	Peter und Klaus gaben beim Verhör an, dass sie unschuldig seien und von nichts wüssten, aber die Richterin vermutete, dass sie loben.		1.2	/b/-/g/
HP	064A	Bei der TÜV Untersuchung fand der Gutachter keine gravierenden Mängel.	0.5	5	/ɛ/-/a/

Table A.1 continued from previous page

pred.	item#	item	cloze	plaus.	contrast
HP	064B	Da sie ständig seinen Unterricht störten, nahm der Lehrer die frechen Schüler in die Mangel.	0.6	4.5	/ɛ/-/a/
LP	064C	Bei der TÜV Untersuchung fand der Gutachter keine gravierenden Mangel.		3	/ɛ/-/a/
LP	064D	Da sie ständig seinen Unterricht störten, nahm der Lehrer die frechen Schüler in die Mängel.		2.5	/ɛ/-/a/
HP	065A	Damit seine Unterlagen nicht zerknittern, verstaut Torsten sie immer in einer Mappe.	0.5	4.9	/p/-/t/
HP	065B	Da Torsten Zahlen gerne mochte, war sein Lieblingsfach in der Schule Mathe.	0.4	4.8	/p/-/t/
LP	065C	Damit seine Unterlagen nicht zerknittern, verstaut Torsten sie immer in einer Mathe.		1.1	/p/-/t/
LP	065D	Da Torsten Zahlen gerne mochte, war sein Lieblingsfach in der Schule Mappe.		1.5	/p/-/t/
HP	066A	Um nicht erkannt zu werden, tragen Diebe häufig Masken.	0.5	4.9	/t/-/k/
HP	066B	Als Hannes das Schiff betrat, staunte er über die hohen Masten.	0.5	4.9	/t/-/k/
LP	066C	Um nicht erkannt zu werden, tragen Diebe häufig Masten.		1.6	/t/-/k/
LP	066D	Als Hannes das Schiff betrat, staunte er über die hohen Masken.		1.8	/t/-/k/
HP	067A	Bevor Verena das Marzipan formen konnte, vermischte sie die Teigzutaten zu einer klebrigen Masse.	1	4.8	/ɛ/-/a/
HP	067B	Sonntagsmorgens gehen Herr und Frau Bach immer im Dom in die heilige Messe.	0.6	3.9	/ɛ/-/a/
LP	067C	Bevor Verena das Marzipan formen konnte, vermischte sie die Teigzutaten zu einer klebrigen Messe.		1.6	/ɛ/-/a/
LP	067D	Sonntagsmorgens gehen Herr und Frau Bach immer im Dom in die heilige Masse.		1.1	/ɛ/-/a/
HP	068A	Der Forscher inspizierte etwas Hausstaub unter dem Mikroskop und sah die Milbe.	0.7	4.4	/b/-/d/
HP	068B	Selbst wenn sie Fehler machten, behandelte der Chef seine Angestellten mit Nachsicht und Milde.	0	5	/b/-/d/
LP	068C	Der Forscher inspizierte etwas Hausstaub unter dem Mikroskop und sah die Milde.		1.4	/b/-/d/
LP	068D	Selbst wenn sie Fehler machten, behandelte der Chef seine Angestellten mit Nachsicht und Milbe.		1.1	/b/-/d/
HP	069A	Beim Gedanken an den Abschied verzog Klara die Miene.	0.7	4.1	/i/-/ɪ/
HP	069B	Im Mittelalter bezeichnete man die Liebe auch als Minne.	0.2	4.4	/i/-/ɪ/

Table A.1 continued from previous page

pred.	item#	item	cloze	plaus.	contrast
LP	069C	Beim Gedanken an den Abschied verzog Klara die Minne.		1.6	/i/-/ɪ/
LP	069D	Im Mittelalter bezeichnete man die Liebe auch als Mine.		2.3	/i/-/ɪ/
HP	070A	Ich finde es kindisch, wenn du in der Öffentlichkeit wie eine Kuh muhst.	0.5	4.7	/u/-/ʊ/
HP	070B	Die Lehrerin hat gesagt, dass du dein Handy vor der Klausur abgeben musst.	0.8	4.8	/u/-/ʊ/
LP	070C	Ich finde es kindisch, wenn du in der Öffentlichkeit wie eine Kuh musst.		1.6	/u/-/ʊ/
LP	070D	Die Lehrerin hat gesagt, dass du dein Handy vor der Klausur abgeben muhst.		1.2	/u/-/ʊ/
HP	071A	Da Markus jede Menge Äpfel hatte, kochte er daraus ein paar Gläser voll Mus.	0.4	4.3	/u/-/ʊ/
HP	071B	Bei einem Trip nach Paris ist die Besichtigung des Eiffelturms ein Muss.	0.7	4.9	/u/-/ʊ/
LP	071C	Da Markus jede Menge Äpfel hatte, kochte er daraus ein paar Gläser voll Muss.		1.6	/u/-/ʊ/
LP	071D	Bei einem Trip nach Paris ist die Besichtigung des Eiffelturms ein Mus.		1.5	/u/-/ʊ/
HP	072A	Nele hatte von ihrer Großmutter gelernt, selbst Kleider zu nähen.	0.9	5	/ɐ/-/ə/
HP	072B	Um das Pferd nicht zu erschrecken, mussten die Reiter sich ihm vorsichtig nähern.	0.8	4.9	/ɐ/-/ə/
LP	072C	Nele hatte von ihrer Großmutter gelernt, selbst Kleider zu nähern.		2.4	/ɐ/-/ə/
LP	072D	Um das Pferd nicht zu erschrecken, mussten die Reiter sich ihm vorsichtig nähen.		1.1	/ɐ/-/ə/
HP	073A	Als Simon einen neuen Vorschlag gemacht hatte, sah er seinen Bruder zustimmend nicken.	0.5	5	/p/-/k/
HP	073B	Da der Tee noch sehr heiß war, konnte Simon nur daran nippen.	0.9	4.9	/p/-/k/
LP	073C	Als Simon einen neuen Vorschlag gemacht hatte, sah er seinen Bruder zustimmend nippen.		1.3	/p/-/k/
LP	073D	Da der Tee noch sehr heiß war, konnte Simon nur daran nicken.		1.5	/p/-/k/
HP	074A	Im Zimmer war es so staubig, dass viele der Besucher niesten.	0.4	4.6	/i/-/ɪ/
HP	074B	Im Garten gab es viele Plätze, an denen Vögel nisten.	0.3	4.5	/i/-/ɪ/
LP	074C	Im Zimmer war es so staubig, dass viele der Besucher nisten.		1.9	/i/-/ɪ/
LP	074D	Im Garten gibt es viele Plätze, an denen Vögel niesten.		2.1	/i/-/ɪ/
HP	075A	Die Rosen im Garten gefielen Kathi so gut, dass sie ein paar für einen Strauß pflückte.	0.5	4.5	/y/-/ʏ/

Table A.1 continued from previous page

pred.	item#	item	cloze	plaus.	contrast
HP	075B	Der Bauer fuhr den Traktor aus der Scheune, bevor er sein Feld pflügte.	0.5	4.8	/y/-/ʏ/
LP	075C	Die Rosen im Garten gefielen Kathi so gut, dass sie ein paar für einen Strauß pflügte.		1.9	/y/-/ʏ/
LP	075D	Der Bauer fuhr den Traktor aus der Scheune, bevor er sein Feld pflückte.		2.1	/y/-/ʏ/
HP	076A	Es ist unverschämt, dass Johanna die Rosinen aus dem Kuchen pickt.	0.3	4.7	/i/-/ɪ/
HP	076B	An der Rose aus dem Garten ist ein Dorn, der ein bisschen pikt.	0.2	4.1	/i/-/ɪ/
LP	076C	Es ist unverschämt, dass Johanna die Rosinen aus dem Kuchen pikt.		3.7	/i/-/ɪ/
LP	076D	An der Rose aus dem Garten ist ein Dorn, der ein bisschen pickt.		2.4	/i/-/ɪ/
HP	077A	Nachdem der kleine Vogel aus dem Nest gefallen war, hat er mitleiderregend gepiepst.	0.4	4.9	/p/-/k/
HP	077B	Als Jule geimpft wurde, hat die Nadel nur leicht gepikst.	0.7	4.8	/p/-/k/
LP	077C	Nachdem der kleine Vogel aus dem Nest gefallen war, hat er mitleiderregend gepikst.		1.7	/p/-/k/
LP	077D	Als Jule geimpft wurde, hat die Nadel nur leicht gepiepst.		2	/p/-/k/
HP	078A	Die Stadt Krakau liegt in Polen.	0.8	5	/o/-/ɔ/
HP	078B	Im Frühling reagieren viele Menschen allergisch auf Pollen.	1	5	/o/-/ɔ/
LP	078C	Die Stadt Krakau liegt in Pollen.		1.7	/o/-/ɔ/
LP	078D	Im Frühling reagieren viele Menschen allergisch auf Polen.		1.6	/o/-/ɔ/
HP	079A	Dem Arzt erschien der Bruch zu kompliziert zum Schienen, weshalb er zu einer Operation riet.	0.4	4.4	/i/-/ɪ/
HP	079B	Als Andrea auf dem Land wohnte, besaß sie ein Pferd, mit dem sie gerne zum Wald ritt.	0.8	4.6	/i/-/ɪ/
LP	079C	Dem Arzt erschien der Bruch zu kompliziert zum Schienen, weshalb er zu einer Operation ritt.		1.2	/i/-/ɪ/
LP	079D	Als Andrea auf dem Land wohnte, besaß sie ein Pferd, mit dem sie gerne zum Wald riet.		1.8	/i/-/ɪ/
HP	080A	Dresden ist die Landeshauptstadt von Sachsen.	0.8	5	/ɛ/-/a/
HP	080B	Beim Mensch ärger dich nicht Spielen hat Mareike meistens Glück und würfelt viele Sechsen.	0.6	4.5	/ɛ/-/a/
LP	080C	Dresden ist die Landeshauptstadt von Sechsen.		1.5	/ɛ/-/a/
LP	080D	Beim Mensch ärgere dich nicht Spielen hat Mareike meistens Glück und würfelt viele Sachsen.		1.1	/ɛ/-/a/
HP	081A	Weil Susi am Geburtstag ihrer Schwester nicht da sein konnte, wollte sie ihr ein Päckchen schicken.	0.8	4.8	/p/-/k/

Table A.1 continued from previous page

pred.	item#	item	cloze	plaus.	contrast
HP	081B	Da es in der Nacht geschneit hatte, musste Herr Meier am nächsten Morgen Schnee schippen.	0.5	4.3	/p/-/k/
LP	081C	Weil Susi am Geburtstag ihrer Schwester nicht da sein konnte, wollte sie ihr ein Päckchen schippen.		1.7	/p/-/k/
LP	081D	Da es in der Nacht geschneit hatte, musste Herr Meier am nächsten Morgen Schnee schicken.		1.2	/p/-/k/
HP	082A	Da es bergauf geht und der Kinderwagen recht schwer ist, ist Inken froh, dass ich ihn jetzt schiebe.	0.8	4.2	/b/-/d/
HP	082B	Es wäre tragisch, wenn unsere Großmutter in nächster Zeit von uns schiede.	0.1	4.2	/b/-/d/
LP	082C	Da es bergauf geht und der Kinderwagen recht schwer ist, ist Inken froh, dass ich ihn jetzt schiede.		1.8	/b/-/d/
LP	082D	Es wäre tragisch, wenn unsere Großmutter in nächster Zeit von uns schiebe.		1.1	/b/-/d/
HP	083A	Um im Schwimmbad nicht auszurutschen, benutzt Bettina in der Dusche ihre Schlappen.	0.1	4.5	/ɛ/-/a/
HP	083B	Die Hochzeitskleider, die die zukünftige Braut anprobierte, hatten unterschiedlich lange Schleppen.	0.2	4.6	/ɛ/-/a/
LP	083C	Um im Schwimmbad nicht auszurutschen, benutzt Bettina in der Dusche ihre Schleppen.		2.4	/ɛ/-/a/
LP	083D	Die Hochzeitskleider, die die zukünftige Braut anprobierte, hatten unterschiedlich lange Schlappen.		2.4	/ɛ/-/a/
HP	084A	Luisa wunderte sich, dass die Kinder so spät noch nicht schliefen.	0.4	4.9	/i/-/ɪ/
HP	084B	Luisa sah den Handwerkern dabei zu, wie sie die Tischplatte mit Schmirgelpapier glatt schliffen.	0.3	4.5	/i/-/ɪ/
LP	084C	Luisa wunderte sich, dass die Kinder so spät noch nicht schliffen.		1.1	/i/-/ɪ/
LP	084D	Luisa sah den Handwerkern dabei zu, wie sie die Tischplatte mit Schmirgelpapier glatt schliefen.		2.2	/i/-/ɪ/
HP	085A	Es geht leichter, wenn ihr eure Tabletten mit Wasser schluckt.	0.5	5	/u/-/ʊ/
HP	085B	Ich weiß noch, dass ihr eure Eltern regelmäßig im Schach schlugt.	0.2	2.8	/u/-/ʊ/
LP	085C	Es geht leichter, wenn ihr eure Tabletten mit Wasser schlugt.		2.6	/u/-/ʊ/
LP	085D	Ich weiß noch, dass ihr eure Eltern regelmäßig im Schach schluckt.		1.5	/u/-/ʊ/
HP	086A	Ein Schwert muss man im heißen Feuer schmieden.	1	4.7	/d/-/g/
HP	086B	Die Katze mochte es, sich mit dem Kopf an Lauras Wange zu schmiegen.	0.5	4.3	/d/-/g/
LP	086C	Ein Schwert muss man im heißen Feuer schmiegen.		1.8	/d/-/g/

Table A.1 continued from previous page

pred.	item#	item	cloze	plaus.	contrast
LP	086D	Die Katze mochte es, sich mit dem Kopf an Lauras Wange zu schmieden.		1.8	/d/-/g/
HP	087A	Wie die Gärtnerin der kleinen Emily zeigte, wachsen Erbsen in Schoten.	0.4	4.5	/o/-/ɔ/
HP	087B	Kerstin liebt die Dudelsackmusik der Schotten.	0.7	4.9	/o/-/ɔ/
LP	087C	Wie die Gärtnerin der kleinen Emily zeigte, wachsen Erbsen in Schotten.		1.8	/o/-/ɔ/
LP	087D	Kerstin liebt die Dudelsackmusik der Schoten.		2.6	/o/-/ɔ/
HP	088A	Das Tragen von knielangen karierten Röcken ist charakteristisch für Schotten.	0.4	5	/p/-/t/
HP	088B	Gestern traf ich meine Tante in der Innenstadt beim Shoppen.	0.3	4.6	/p/-/t/
LP	088C	Das Tragen von knielangen karierten Röcken ist charakteristisch für Shoppen.		1.8	/p/-/t/
LP	088D	Gestern traf ich meine Tante in der Innenstadt beim Schotten.		2.4	/p/-/t/
HP	089A	Wenn Jäger mehrere Projektile mit einem Schuss verschießen möchten, dann schießen sie mit Schrot.	0.4	4.6	/o/-/ɔ/
HP	089B	Nach dem Unfall war das Auto nur noch Schrott.	0.9	4.9	/o/-/ɔ/
LP	089C	Wenn Jäger mehrere Projektile mit einem Schuss verschießen möchten, dann schießen sie mit Schrott.		2.5	/o/-/ɔ/
LP	089D	Nach dem Unfall war das Auto nur noch Schrot.		1.7	/o/-/ɔ/
HP	090A	Mithilfe eines Tricks brachte der Zauberer den Hut zum Schweben.	0.5	4.9	/b/-/d/
HP	090B	Das Möbelhaus IKEA stammt ursprünglich aus Schweden.	1	4.9	/b/-/d/
LP	090C	Mithilfe eines Tricks, brachte der Zauberer den Hut zum Schweden.		1.4	/b/-/d/
LP	090D	Das Möbelhaus IKEA stammt ursprünglich aus Schweben.		1.1	/b/-/d/
HP	091A	Wenn man Klöße kocht, sollte das Wasser am besten nicht kochen, sondern nur sieden.	0.3	4	/d/-/g/
HP	091B	In den Perserkriegen gelang es den Griechen über die Perser zu siegen.	0.9	4.7	/d/-/g/
LP	091C	Wenn man Klöße kocht, sollte das Wasser am besten nicht kochen, sondern nur siegen.		1.8	/d/-/g/
LP	091D	In den Perserkriegen gelang es den Griechen über die Perser zu sieden.		1.1	/d/-/g/
HP	092A	Als ihr Mathelehrer Anna nach dem Abitur das Du anbot, fand sie es ungewohnt, ihn nicht mehr zu siezen.	0.9	4.8	/i/-/ɪ/
HP	092B	Da der Hörsaal überfüllt war, mussten einige der Studenten während der Vorlesung auf dem Boden sitzen.	1	4.6	/i/-/ɪ/

Table A.1 continued from previous page

pred.	item#	item	cloze	plaus.	contrast
LP	092C	Als ihr Mathelehrer Anna nach dem Abitur das Du anbot, fand sie es ungewohnt, ihn nicht mehr zu sitzen.		1.3	/i/-/ɪ/
LP	092D	Da der Hörsaal überfüllt war, mussten einige der Studenten während der Vorlesung auf dem Boden siezen.		1.4	/i/-/ɪ/
HP	093A	Zur Taufe des Babys versammelte sich die gesamte Sippe.	0	4.2	/p/-/t/
HP	093B	Sich zur Begrüßung auf die Wangen zu küssen ist nicht in jedem Land eine Sitte.	0.4	4.9	/p/-/t/
LP	093C	Zur Taufe des Babys versammelte sich die gesamte Sitte.		2.1	/p/-/t/
LP	093D	Sich zur Begrüßung auf die Wangen zu küssen ist nicht in jedem Land eine Sippe.		1.3	/p/-/t/
HP	094A	In den Haaren trug Mia eine hübsche Spange.	0.3	4.9	/p/-/t/
HP	094B	Beim Hochsprung benutzen die Turner eine Stange.	0.7	4.6	/p/-/t/
LP	094C	In den Haaren trug Mia eine hübsche Stange.		1.7	/p/-/t/
LP	094D	Beim Hochsprung benutzen die Turner eine Spange.		1.9	/p/-/t/
HP	095A	Dankbar empfing der Verein die Spende.	0.8	4.8	/p/-/t/
HP	095B	Auf dem Weihnachtsmarkt bewunderten die Besucher die vielfältigen Auslagen der Stände.	0.3	4.8	/p/-/t/
LP	095C	Dankbar empfing der Verein die Stände.		2.3	/p/-/t/
LP	095D	Auf dem Weihnachtsmarkt bewunderten die Besucher die vielfältigen Auslagen der Spende.		1.7	/p/-/t/
HP	096A	Marlene erzählt den Besuchern des Theaters, dass sie auf der Bühne am liebsten die Julia spielt.	0.5	4.8	/p/-/t/
HP	096B	Da Merlin sich immer gerne in den Vordergrund drängt, befürchtet Korinna, dass er ihr beim Auftritt die Show stiehlt.	0.7	4.9	/p/-/t/
LP	096C	Marlene erzählt den Besuchern des Theaters, dass sie auf der Bühne am liebsten die Julia stiehlt.		1.6	/p/-/t/
LP	096D	Da Merlin sich immer gerne in den Vordergrund drängt, befürchtet Korinna, dass er ihr beim Auftritt die Show spielt.		1.8	/p/-/t/
HP	097A	Da das Fleisch sehr fest war, ließ es sich nicht gut auf die Gabel spießen.	0.5	4.5	/p/-/t/
HP	097B	Die Archäologen freuten sich, als sie bei ihren Ausgrabungen auf alte Gräber stießen.	0.3	5	/p/-/t/
LP	097C	Da das Fleisch sehr fest war, ließ es sich nicht gut auf die Gabel stießen.		1.8	/p/-/t/
LP	097D	Die Archäologen freuten sich, als sie bei ihren Ausgrabungen auf alte Gräber spießen.		1.3	/p/-/t/
HP	098A	Im Zoo erklärte Simone ihrem Sohn, dass man das Lama nicht ärgern darf, da es sonst spuckt.	0.7	4.7	/u/-/ʊ/
HP	098B	Wenn bei Gewitter und Wind das Gebälk des Hauses knarzt, könnte man meinen, dass es hier spukt.	1	4.6	/u/-/ʊ/

Table A.1 continued from previous page

pred.	item#	item	cloze	plaus.	contrast
LP	098C	Im Zoo erklärte Simone ihrem Sohn, dass man das Lama nicht ärgern darf, da es sonst spukt.		3.3	/u/-/ʊ/
LP	098D	Wenn bei Gewitter und Wind das Gebälk des Hauses knarzt, könnte man meinen, dass es hier spuckt.		2.7	/u/-/ʊ/
HP	099A	Nach dem Frühstück stellte Jana das dreckige Geschirr in die Spüle.	0.5	4.9	/p/-/t/
HP	099B	Nachdem Jana und Niklas für das Fest den Tisch ins Wohnzimmer getragen hatten, holten sie noch die Stühle.	0.7	3.9	/p/-/t/
LP	099C	Nach dem Frühstück stellte Jana das dreckige Geschirr in die Stühle.		1	/p/-/t/
LP	099D	Nachdem Jana und Niklas den Tisch ins Nebenzimmer getragen hatten, holten sie noch die Spüle.		1.4	/p/-/t/
HP	100A	Nach dem Ausritt bringt Anna ihr Pferd zurück in seine Box in der Stallung.	0.2	3.5	/ɛ/-/a/
HP	100B	Erst eine Woche nach den Anschuldigungen bezog die Regierung zu den Vorwürfen Stellung.	0.9	5	/ɛ/-/a/
LP	100C	Nach dem Ausritt bringt Elena ihr Pferd zurück in seine Box in der Stellung.		2.6	/ɛ/-/a/
LP	100D	Erst eine Woche nach den Anschuldigungen bezog die Regierung zu den Vorwürfen Stallung.		1.3	/ɛ/-/a/
HP	101A	Der Lehrer fragt die Schüler, welche Bäume ursprünglich nicht aus Europa stammen.	0.5	4.9	/ɛ/-/a/
HP	101B	Um Muskeln aufzubauen, geht Lars jede Woche Gewichte stemmen.	0.4	4.9	/ɛ/-/a/
LP	101C	Der Lehrer fragt die Schüler, welche Bäume ursprünglich nicht aus Europa stemmen.		1.5	/ɛ/-/a/
LP	101D	Um Muskeln aufzubauen, geht Lars jede Woche Gewichte stammen.		1.9	/ɛ/-/a/
HP	102A	Der Petersdom ist eine heilige Stätte.	0.5	4.9	/p/-/t/
HP	102B	Der bevorzugte Lebensraum der Präriehunde ist die Steppe.	0.2	5	/p/-/t/
LP	102C	Der Petersdom ist eine heilige Steppe.		1.3	/p/-/t/
LP	102D	Der bevorzugte Lebensraum der Präriehunde ist die Stätte.		1.3	/p/-/t/
HP	103A	Julian hatte schon immer einmal auf dem Dachboden übernachten wollen, doch seine Mutter sagte, es sei zu staubig.	0.4	4.6	/b/-/d/
HP	103B	Laut biologischer Klassifizierung sind Bergkiefern strauchig, und Bananenpflanzen sind staudig.	0.1	3.8	/b/-/d/
LP	103C	Julian hatte schon immer einmal auf dem Dachboden übernachten wollen, doch seine Mutter sagte, es sei zu staudig.		1.3	/b/-/d/

Table A.1 continued from previous page

pred.	item#	item	cloze	plaus.	contrast
LP	103D	Laut biologischer Klassifizierung sind Bergkiefern strauchig, und Bananenpflanzen sind staubig.		1.3	/b/-/d/
HP	104A	Es kann sein, dass die Diebin im Geschäft etwas stiehlt.	0.3	4.4	/i/-/ɪ/
HP	104B	Die junge Mutter sagt, dass sie gerade das Kind stillt.	0.3	4.8	/i/-/ɪ/
LP	104C	Es kann sein, dass die Diebin im Geschäft etwas stillt.		1.5	/i/-/ɪ/
LP	104D	Die junge Mutter sagt, dass sie gerade das Kind stiehlt.		1.6	/i/-/ɪ/
HP	105A	Nachdem Pia ihrer Familie die Neuigkeiten verkündet hatte, herrschte einige Augenblicke lang überraschte Stille.	1	4.2	/i/-/ɪ/
HP	105B	Die Weißweingläser haben alle ziemlich lange Stiele.	0.5	4.8	/i/-/ɪ/
LP	105C	Nachdem Pia ihrer Familie die Neuigkeiten verkündet hatte, herrschte einige Augenblicke lang überraschte Stiele.		1.5	/i/-/ɪ/
LP	105D	Die Weißweingläser haben alle ziemlich lange Stille.		1.7	/i/-/ɪ/
HP	106A	Bei der Schlussszene des Horrorfilms ist Nora vor Schreck der Atem gestockt.	0.4	4.4	/p/-/k/
HP	106B	Die Polizei hat den Raser an einer Straßensperre gestoppt.	0.2	5	/p/-/k/
LP	106C	Bei der Schlussszene des Horrorfilms ist Nora vor Schreck der Atem gestoppt.		3.8	/p/-/k/
LP	106D	Die Polizei hat den Raser an einer Straßensperre gestockt.		1.4	/p/-/k/
HP	107A	Als Sophia an dem Bienenstock vorbeikam hörte sie die Bienen summen.	1	4.3	/u/-/ʊ/
HP	107B	Sophia durfte sich dem Löwen nicht weiter nähern, um das Foto zu machen, doch zum Glück konnte sie mit ihrem Handy zoomen.	0.6	4.7	/u/-/ʊ/
LP	107C	Als Sophia an dem Bienenstock vorbei kam hörte sie die Bienen zoomen.		1.7	/u/-/ʊ/
LP	107D	Sophia durfte sich dem Löwen nicht weiter nähern, um das Foto zu machen, doch zum Glück konnte sie mit ihrem Handy summen.		1.3	/u/-/ʊ/
HP	108A	Da es im Zimmer dunkel war, musste Silvia sich an der Wand entlang tasten.	0.9	5	/ɛ/-/a/
HP	108B	Bevor eine Firma ein Produkt auf den Markt bringen kann, muss sie es ausgiebig testen.	0.9	5	/ɛ/-/a/
LP	108C	Da es im Zimmer dunkel war, musste Silvia sich an der Wand entlang testen.		1.9	/ɛ/-/a/
LP	108D	Bevor eine Firma ein Produkt auf den Markt bringen kann, muss sie es ausgiebig tasten.		2	/ɛ/-/a/
HP	109A	Es gab viel, was Karina an Kindern mochte, abgesehen von deren lautem Geschrei und wildem Toben.	0.5	4.8	/b/-/d/

Table A.1 continued from previous page

pred.	item#	item	cloze	plaus.	contrast
HP	109B	Laura überlegte sich eine Geschichte über das letzte Leben einer Katze nach ihren sechs Toden.	0.3	3.7	/b/-/d/
LP	109C	Es gab viel, was Karina an Kindern mochte, abgesehen von deren lautem Geschrei und wildem Toden.		1.5	/b/-/d/
LP	109D	Laura überlegte sich eine Geschichte über das letzte Leben einer Katze nach ihren sechs Toben.		1.1	/b/-/d/
HP	110A	Die Verkäuferin packte die Einkäufe des Kunden in Tüten.	0.5	4.9	/p/-/t/
HP	110B	Wenn Katja abends alleine unterwegs ist, meidet sie die Begegnung mit seltsamen Typen.	0	4.3	/p/-/t/
LP	110C	Die Verkäuferin packte die Einkäufe des Kunden in Typen.		1.3	/p/-/t/
LP	110D	Wenn Katja abends alleine unterwegs ist, meidet sie die Begegnung mit seltsamen Tüten.		1.6	/p/-/t/
HP	111A	Das Hotelzimmer war in keinem guten Zustand, aber davon ließen sich Petras Eltern den Urlaub nicht vermiesen.	0.3	4.9	/i/-/ɪ/
HP	111B	Ella wusste schon jetzt: sie würde ihren Freund während des Auslandssemesters vermissen.	0.6	4.8	/i/-/ɪ/
LP	111C	Das Hotelzimmer war in keinem guten Zustand, aber davon ließen sich Petras Eltern den Urlaub nicht vermissen.		1.6	/i/-/ɪ/
LP	111D	Ella wusste schon jetzt: sie würde ihren Freund während des Auslandssemesters vermiesen.		1.5	/i/-/ɪ/
HP	112A	Nachdem sie verspätet am Flughafen ankamen, sind Renate und Karla zum Terminal gerannt, damit sie den Flieger nicht verpassten.	0.3	4.5	/ɛ/-/a/
HP	112B	Renate engagiert sich für Fahrverbote in Städten, da sie nicht möchte, dass die Autos die Luft mit Abgasen verpesten.	0.5	4.8	/ɛ/-/a/
LP	112C	Nachdem sie verspätet am Flughafen ankamen, sind Renate und Karla zum Terminal gerannt, damit sie den Flieger nicht verpesten.		1.2	/ɛ/-/a/
LP	112D	Renate engagiert sich für Fahrverbote in Städten, da sie nicht möchte, dass die Autos die Luft mit Abgasen verpassten.		1.5	/ɛ/-/a/
HP	113A	Die Besucher haben sich vor dem Podium versammelt.	0.6	5	/ɛ/-/a/
HP	113B	Miriam hat die letzte Prüfung ziemlich versemmelt.	0.1	4.7	/ɛ/-/a/
LP	113C	Die Besucher haben sich vor dem Podium versemmelt.		1.2	/ɛ/-/a/
LP	113D	Miriam hat die letzte Prüfung ziemlich versammelt.		1.3	/ɛ/-/a/
HP	114A	Beim Lagerfeuer fragte Herr Kraus irritiert, ob die Kinder die Blätter absichtlich versengten.	0	3.6	/t/-/k/

Table A.1 continued from previous page

pred.	item#	item	cloze	plaus.	contrast
HP	114B	Die Kinder machten sich einen Spaß daraus, Papierschiffchen im See zu versenken.	0.9	4.4	/t/-/k/
LP	114C	Beim Lagerfeuer fragte Herr Kraus irritiert, ob die Kinder die Blätter absichtlich versenken.		3	/t/-/k/
LP	114D	Die Kinder machten sich einen Spaß draus, Papierschiffchen im See zu versengten.		1.5	/t/-/k/
HP	115A	Lara wechselte schnell das Thema, da sie bemerkte, dass ihre Freunde sich bei der Unterhaltung vor Unbehagen geradezu wanden.	0.1	4.4	/ɛ/-/a/
HP	115B	Als Joachim falsch abbog, schlug das Navi vor, bitte bei der nächsten Gelegenheit zu wenden.	0.9	4.7	/ɛ/-/a/
LP	115C	Lara wechselte schnell das Thema, da sie bemerkte, dass ihre Freunde sich bei der Unterhaltung vor Unbehagen geradezu wenden.		3.1	/ɛ/-/a/
LP	115D	Als Joachim falsch abbog, schlug das Navi vor, bitte bei der nächsten Gelegenheit zu wanden.		1.6	/ɛ/-/a/
HP	116A	Bevor ihre Mutter nach Hause kam und schimpfen konnte, haben die Kinder ihren Dreck wieder weggemacht.	0.3	4.7	/t/-/k/
HP	116B	Auf den letzten hundert Metern hat der Läufer seinen Zeitverlust wieder wettgemacht.	0.2	4.2	/t/-/k/
LP	116C	Bevor ihre Mutter nach Hause kam und schimpfen konnte, haben die Kinder ihren Dreck wieder wettgemacht.		2.1	/t/-/k/
LP	116D	Auf den letzten hundert Metern hat der Läufer seinen Zeitverlust wieder weggemacht.		3.1	/t/-/k/
HP	117A	Der Stich war mit Sicherheit von einer Wespe.	0.5	5	/p/-/t/
HP	117B	Da die Hochzeitsfeier hauptsächlich draußen stattfinden sollte, zog Carlo unter seine Anzugjacke noch eine dünne Weste.	0.3	4.4	/p/-/t/
LP	117C	Der Stich war mit Sicherheit von einer Weste.		1.6	/p/-/t/
LP	117D	Da die Hochzeitsfeier hauptsächlich draußen stattfinden sollte, zog Carlo unter seine Anzugjacke noch eine dünne Wespe.		1.1	/p/-/t/
HP	118A	Im Erste-Hilfe-Kurs lernten die Schüler wie man jemanden, der nicht mehr atmet, wiederbelebt.	0.5	4.7	/p/-/k/
HP	118B	Da ihre Note im letzten Seminar nicht so gut war, hat die Studentin das Seminar für das kommende Semester wiederbelegt.	0.1	4.3	/p/-/k/
LP	118C	Im Erste-Hilfe-Kurs lernten die Schüler wie man jemanden, der nicht mehr atmet, wiederbelegt.		2.1	/p/-/k/
LP	118D	Da ihre Note im letzten Seminar nicht so gut war, hat die Studentin das Seminar für das kommende Semester wiederbelebt.		1.2	/p/-/k/

Table A.1 continued from previous page

pred.	item#	item	cloze	plaus.	contrast
HP	119A	Im Kindergarten bastelt Mira aus buntem Karton einen Stern mit fünf Zacken.	0.5	4.8	/ɛ/-/a/
HP	119B	Da der Hund den ganzen Tag durch die Wiese gelaufen war, durchsuchte Aaron abends sein Fell nach Zecken.	0.9	4.9	/ɛ/-/a/
LP	119C	Im Kindergarten bastelt Mira aus buntem Karton einen Stern mit fünf Zecken.		1.1	/ɛ/-/a/
LP	119D	Da der Hund den ganzen Tag durch die Wiese gelaufen war, durchsuchte Aaron abends sein Fell nach Zacken.		1.3	/ɛ/-/a/
HP	120A	Oliver erzählt seinen Freunden, dass er von Zuhause auszieht, um mit seiner Freundin zusammenzuleben.	0.2	4.9	/b/-/g/
HP	120B	Da das Geschenk sehr teuer war, beschlossen die Gäste, ihr Geld zusammenzulegen.	0.4	4.9	/b/-/g/
LP	120C	Oliver erzählt seinen Freunden, dass er von Zuhause auszieht, um mit seiner Freundin zusammenzulegen.		1.4	/b/-/g/
LP	120D	Da das Geschenk sehr teuer war, beschlossen die Gäste, ihr Geld zusammenzuleben.		1.2	/b/-/g/
HP	F001A	Andreas will unbedingt nach Italien auswandern und ist von dieser Idee wie besessen	0.8	4.9	/s/-/ts/
HP	F001B	Weil ihre Angestellte neue Arbeit gefunden hat, musste sie den Posten neu besetzen	0.7	4.4	/s/-/ts/
LP	F001C	Andreas will unbedingt nach Italien auswandern und ist von dieser Idee wie besetzen.		1.8	/s/-/ts/
LP	F001D	Weil ihre Angestellte neue Arbeit gefunden hat, musste sie den Posten neu besessen.		1.4	/s/-/ts/
HP	F002A	Maria sieht zu, wie die Biene die Blume bestäubt	0.7	4.8	/ʃ/-/∅/
HP	F002B	Um dem Patienten seinen Weisheitszahn ziehen zu können, wird der Unterkiefer vom Zahnarzt mit einer Spritze betäubt	0.9	5	/ʃ/-/∅/
LP	F002C	Maria sieht zu, wie die Biene die Blume betäubt.		2.4	/ʃ/-/∅/
LP	F002D	Um dem Patienten seinen Weisheitszahn ziehen zu können, wird der Unterkiefer vom Zahnarzt mit einer Spritze bestäubt.		1.4	/ʃ/-/∅/
HP	F003A	Der tägliche Futterbedarf eines Hundes schlägt finanziell schon relativ hoch zu Buche	0.5	3.9	/s/-/x/
HP	F003B	Während der christlichen Fastenzeit verzichten die Gläubigen auf viele Verführungen und tun für ihre Sünden Bußе	0.9	3.6	/s/-/x/
LP	F003C	Der tägliche Futterbedarf eines Hundes schlägt finanziell schon relativ hoch zu Bußе.		4.3	/s/-/x/
LP	F003D	Während der christlichen Fastenzeit verzichten die Gläubigen auf viele Verführungen und tun für ihre Sünden Buche.		1.4	/s/-/x/

Table A.1 continued from previous page

pred.	item#	item	cloze	plaus.	contrast
HP	F004A	In ihren Gesichtern haben Babys süße rote, dicke, kleine Bäckchen	0.5	4.5	/ts/-/∅/
HP	F004B	Für Babys gibt es in Schwimmbädern ein eigenes Becken	0.9	4.8	/ts/-/∅/
LP	F004C	In ihren Gesichtern haben Babys süße rote, dicke, kleine Becken.		1.3	/ts/-/∅/
LP	F004D	Für Babys gibt es in Schwimmbädern ein eigenes Bäckchen.		1.3	/ts/-/∅/
HP	F005A	Dem Häftling gelang es am helllichten Tag aus dem Gefängnis zu fliehen	0.7	4.8	/s/-/∅/
HP	F005B	Die Familie Peters will das komplette Haus sanieren und kauft dafür Fliesen, mit denen sie den Badezimmerboden fliesen	0.4	4.9	/s/-/∅/
LP	F005C	Dem Häftling gelang es am helllichten Tag aus dem Gefängnis zu fliesen.		1.7	/s/-/∅/
LP	F005D	Die Familie Peters will das komplette Haus sanieren und kauft dafür Fliesen, mit denen sie den Badezimmerboden fliehen.		1.2	/s/-/∅/
HP	F006A	Mathematik war eines seiner liebsten Fächer	0.9	5	/ʃ/-/ts/
HP	F006B	Im Weinkeller stehen bereits große Fässer	0.7	4.7	/ʃ/-/ts/
LP	F006C	Mathematik war eines seiner liebsten Fässer.		1.6	/ʃ/-/ts/
LP	F006D	Im Weinkeller stehen bereits große Fächer.		3.2	/ʃ/-/ts/
HP	F007A	Die Milch im heißen Kaffee war schon sehr alt, also hat sie geflockt	0.4	4.1	/f/-/∅/
HP	F007B	Die Polizisten haben die Einbrecher in eine Falle gelockt	0.5	4.8	/f/-/∅/
LP	F007C	Die Milch im heißen Kaffee war schon sehr alt, also hat sie gelockt.		1.8	/f/-/∅/
LP	F007D	Die Polizisten haben die Einbrecher in eine Falle geflockt.		1.4	/f/-/∅/
HP	F008A	Da Laura aufgeregt war, ob ihr Date nun kommen würde, hat sie sich die ganze Zeit aus Nervosität mit ihren Händen in ihrem Gesicht herum gefummelt	0.4	4.5	/f/-/ʃ/
HP	F008B	Beim Mau-Mau Spiel hat das Kind heimlich geschummelt	0.5	4.8	/f/-/ʃ/
LP	F008C	Da Laura aufgeregt war, ob ihr Date nun kommen würde, hat sie sich die ganze Zeit aus Nervosität mit ihren Händen in ihrem Gesicht herum geschummelt.		1.6	/f/-/ʃ/
LP	F008D	Beim Mau-Mau Spiel hat das Kind heimlich gefummelt.		1.9	/f/-/ʃ/
HP	F009A	Seitdem er vom Baum gefallen ist und sich das Bein gebrochen hat, hat Ulrich immer ein bisschen auf dem rechten Bein gehinkt	0.6	4.7	/h/-/ts/
HP	F009B	Katharina gewinnt immer, denn ihre Glückswürfel sind gezinkt	0.7	3.9	/h/-/ts/

Table A.1 continued from previous page

pred.	item#	item	cloze	plaus.	contrast
LP	F009C	Seitdem er vom Baum gefallen ist und sich das Bein gebrochen hat, hat Ulrich immer ein bisschen auf dem rechten Bein gezinkt.		1.8	/h/-/ts/
LP	F009D	Katharina gewinnt immer, denn ihre Glückswürfel sind gehinkt.		1.5	/h/-/ts/
HP	F010A	Die Gestalt war durch das Licht im Flur hinter ihr in Schatten gehüllt	0.4	4.1	/f/-/h/
HP	F010B	Dieser Krapfen vom Bäcker ist mit Erdbeermarmelade gefüllt	0.8	4.9	/f/-/h/
LP	F010C	Die Gestalt war durch das Licht im Flur hinter ihr in Schatten gefüllt.		1.7	/f/-/h/
LP	F010D	Dieser Krapfen vom Bäcker ist mit Erdbeermarmelade gehüllt.		1.4	/f/-/h/
HP	F011A	Hier wird nicht gekleckert, sondern geklotzt	0.7	4	/pf/-/ts/
HP	F011B	Anstatt zu klingeln hat Maria an der Tür geklopft	0.9	4.9	/pf/-/ts/
LP	F011C	Hier wird nicht gekleckert, sondern geklopft.		2.5	/pf/-/ts/
LP	F011D	Anstatt zu klingeln hat Maria an der Tür geklotzt.		1.7	/pf/-/ts/
HP	F012A	Martin hat mit der Taschenlampe unter das Bett geleuchtet	0.8	5	/x/-/∅/
HP	F012B	Um 12 Uhr mittags haben die Kirchenglocken geläutet	0.9	4.7	/x/-/∅/
LP	F012C	Martin hat mit der Taschenlampe unter das Bett geläutet.		1.4	/x/-/∅/
LP	F012D	Um 12 Uhr mittags haben die Kirchenglocken geleuchtet.		1.8	/x/-/∅/
HP	F013A	Wenn er Lust auf Kuchen hat, sagt Opa, es hat ihn schon den ganzen Tag nach Kuchen gelüstet	0.4	4.4	/f/-/s/
HP	F013B	Aufgrund von Corona hat die Lehrerin das Klassenzimmer alle 20 Minuten gelüftet	0.9	4.8	/f/-/s/
LP	F013C	Wenn er Lust auf Kuchen hat, sagt Opa, es hat ihn schon den ganzen Tag nach Kuchen gelüftet.		1.3	/f/-/s/
LP	F013D	Aufgrund von Corona hat die Lehrerin das Klassenzimmer alle 20 Minuten gelüstet.		1	/f/-/s/
HP	F014A	Während des 2. Weltkrieges haben Josefs Großeltern in einer kleinen Baracke gehaust	0.2	4.5	/s/-/x/
HP	F014B	Damit die Gläser seiner Brille beschlagen waren, hat Luis mit seinem warmen Atem auf seine Brille gehaucht	0.7	4.7	/s/-/x/
LP	F014C	Während des 2. Weltkrieges haben Josefs Großeltern in einer kleinen Baracke gehaucht		1.4	/s/-/x/
LP	F014D	Damit die Gläser seiner Brille beschlagen waren, hat Luis mit seinem warmen Atem auf seine Brille gehaust		1.9	/s/-/x/
HP	F015A	Am Ende des Arbeitstages hat der Chef immer das Geld in der Kasse gezählt	0.7	4.6	/ʃ/-/ts/

Table A.1 continued from previous page

pred.	item#	item	cloze	plaus.	contrast
HP	F015B	Als Anna noch als Küchenhilfe arbeitete, hat sie immer die Kartoffeln geschält	0.9	4.8	/ʃ/-/ts/
LP	F015C	Am Ende des Arbeitstages hat der Chef immer das Geld in der Kasse geschält.		1.6	/ʃ/-/ts/
LP	F015D	Als Anna noch als Küchenhilfe arbeitete, hat sie immer die Kartoffeln gezählt.		2.8	/ʃ/-/ts/
HP	F016A	Um die Erdbeeren zu pflanzen, lockern sie den Boden mit einer Schaufel und einer Hacke	0.5	4.4	/s/-/∅/
HP	F016B	Der Unterschenkel von Schwein, Kalb oder Lamm ist eine Spezialität und heißt auch Haxe	0.6	4.5	/s/-/∅/
LP	F016C	Um die Erdbeeren zu pflanzen, lockern sie den Boden mit einer Schaufel und einer Haxe.		2.3	/s/-/∅/
LP	F016D	Der Unterschenkel von Schwein, Kalb oder Lamm ist eine Spezialität und heißt auch Hacke.		2.3	/s/-/∅/
HP	F017A	Da ihr Sohn noch nicht strafmündig war, mussten die Eltern für ihn haften	0.6	4.9	/f/-/s/
HP	F017B	Schulfächer wie Mathe, sind Fächer, die viele Schüler schon immer hassten	0.7	4.6	/f/-/s/
LP	F017C	Da ihr Sohn noch nicht strafmündig war, mussten die Eltern für ihn hassten.		1.4	/f/-/s/
LP	F017D	Schulfächer wie Mathe, sind Fächer, die viele Schüler schon immer haften.		1.4	/f/-/s/
HP	F018A	Susi und Peter mochten noch nie gerne Gemüse und Rosenkohl war das, was sie am meisten hassten	0.9	3.8	/s/-/∅/
HP	F018B	Christina erinnert sich gerne zurück an die Zeit, als sie und ihre Geschwister noch Spaß hatten	1	4.7	/s/-/∅/
LP	F018C	Susi und Peter mochten noch nie gerne Gemüse und Rosenkohl war das, was sie am meisten hatten.		1.9	/s/-/∅/
LP	F018D	Christina erinnert sich gerne zurück an die Zeit, als sie und ihre Geschwister noch Spaß hassten.		1.9	/s/-/∅/
HP	F019A	Die Blätter gehen nicht verloren, wenn du sie in einen Ordner heftest	0.6	4.6	/f/-/∅/
HP	F019B	Du wärst nicht durch die Prüfung gefallen, wenn du vorher genug dafür gelernt hättest	1	4.6	/f/-/∅/
LP	F019C	Die Blätter gehen nicht verloren, wenn du sie in einen Ordner hättest.		2.2	/f/-/∅/
LP	F019D	Du wärst nicht durch die Prüfung gefallen, wenn du vorher genug dafür gelernt heftest.		1.4	/f/-/∅/
HP	F020A	Im Frühjahr schneidet Harald im Garten immer die große Hecke	0.8	5	/s/-/∅/
HP	F020B	Bibi Blocksberg ist eine Hexe	1	4.9	/s/-/∅/
LP	F020C	Im Frühjahr schneidet Harald im Garten immer die große Hexe.		1.4	/s/-/∅/

Table A.1 continued from previous page

pred.	item#	item	cloze	plaus.	contrast
LP	F020D	Bibi Blocksberg ist eine Hecke.		1.2	/s/-/∅/
HP	F021A	Der Bauer war so wütend, dass die Nachbarskinder immer seine frische Milch klauten, dass er beim nächsten Mal den Hund auf sie hetzte	0.5	4.4	/s/-/∅/
HP	F021B	Wilma hätte einen Regenschirm mitgenommen, wenn sie von dem Schauer gewusst hätte	1	4.8	/s/-/∅/
LP	F021C	Der Bauer war so wütend, dass die Nachbarskinder immer seine frische Milch klauten, dass er beim nächsten Mal den Hund auf sie hätte.		1.6	/s/-/∅/
LP	F021D	Wilma hätte einen Regenschirm mitgenommen, wenn sie von dem Schauer gewusst hetzte.		1.4	/s/-/∅/
HP	F022A	Mara kauft für ihr Handy eine Hülle	0.7	4.9	/s/-/∅/
HP	F022B	Erbsen, Bohnen und Linsen gehören alle zur gleichen Gattung und werden umschlossen von einer Hülse	0.8	4.5	/s/-/∅/
LP	F022C	Mara kauft für ihr Handy eine Hülse.		2.3	/s/-/∅/
LP	F022D	Erbsen, Bohnen und Linsen gehören alle zur gleichen Gattung und werden umschlossen von einer Hülle.		4	/s/-/∅/
HP	F023A	Ausgegeben wird die Suppe in der Kantine von den Mitarbeiterinnen mit großen Kellen	0.6	4.8	/x/-/∅/
HP	F023B	Früher tranken Könige nicht aus Bechern, sondern aus goldenen Kelchen	0.7	5	/x/-/∅/
LP	F023C	Ausgegeben wird die Suppe in der Kantine von den Mitarbeiterinnen mit großen Kelchen.		3	/x/-/∅/
LP	F023D	Früher tranken Könige nicht aus Bechern, sondern aus goldenen Kellen.		2.6	/x/-/∅/
HP	F024A	Für das neue Bett brauchen wir noch Decken und Kissen	0.9	5	/ʃ/-/ts/
HP	F024B	Die Rehmutter säugt ihre beiden kleinen Kitzen	0.9	4	/ʃ/-/ts/
LP	F024C	Für das neue Bett brauchen wir noch Decken und Kitzen.		1.7	/ʃ/-/ts/
LP	F024D	Die Rehmutter säugt ihre beiden kleinen Kissen.		1.1	/ʃ/-/ts/
HP	F025A	Helmut erzählt seiner Nachbarin abfällig: "Weil Sabine verrückt geworden ist, muss sie jetzt in die Klapse	0.4	4.8	/s/-/∅/
HP	F025B	Mara sagt zu ihrem Bruder: "Sei endlich ruhig und halt die Klappe	0.9	4.6	/s/-/∅/
LP	F025C	Helmut erzählt seiner Nachbarin abfällig: "Weil Sabine verrückt geworden ist, muss sie jetzt in die Klappe."		1.6	/s/-/∅/
LP	F025D	Mara sagt zu ihrem Bruder: "Sei endlich ruhig und halt die Klapse."		1.2	/s/-/∅/
HP	F026A	Conni erzählt ihrem Freund, dass ihre Freunde bei der letzten Party wegen des vielen Alkohols in den Garten kotzten	0.6	5	/ʃ/-/ts/
HP	F026B	Im Schuhgeschäft fragt Lea den Verkäufer wie viel die Schuhe kosten	0.8	4.9	/ʃ/-/ts/

Table A.1 continued from previous page

pred.	item#	item	cloze	plaus.	contrast
LP	F026C	Conni erzählt ihrem Freund, dass ihre Freunde bei der letzten Party wegen des vielen Alkohols in den Garten kosten.		1.5	/ʃ/-/ts/
LP	F026D	Im Schuhgeschäft fragt Lea den Verkäufer wie viel die Schuhe kotzten.		1	/ʃ/-/ts/
HP	F027A	Mönche tragen ein langes Gewand mit Kapuze und man nennt es Kutte	0.8	4.8	/s/-/∅/
HP	F027B	Als es noch keine Autos gab, fuhren reiche Leute mit einer Kutsche	1	4.8	/s/-/∅/
LP	F027C	Mönche tragen ein langes Gewand mit Kapuze und man nennt es Kutsche.		1.9	/s/-/∅/
LP	F027D	Als es noch keine Autos gab, fuhren reiche Leute in einer Kutte.		1.4	/s/-/∅/
HP	F028A	Um auf die Piste zu gelangen, nehmen Skifahrer und Snowboarder den Lift	0.6	5	/f/-/x/
HP	F028B	Am Ende des Tunnels sah Fatima ein Licht	0.9	4.9	/f/-/x/
LP	F028C	Um auf die Piste zu gelangen, nehmen Skifahrer und Snowboarder den Licht.		1.7	/f/-/x/
LP	F028D	Am Ende des Tunnels sah Fatima ein Lift.		2.2	/f/-/x/
HP	F029A	Die Angestellten sind überglücklich, dass ihr Vorgesetzter endlich entlassen wurde, weil sie alle unter seinen Wutausbrüchen litten	0.7	5	/f/-/∅/
HP	F029B	Frau Geiger will bloß nicht alt werden und Falten bekommen, deswegen lässt sie sich jegliche Fältchen liften	0.2	4.6	/f/-/∅/
LP	F029C	Die Angestellten sind überglücklich, dass ihr Vorgesetzter endlich entlassen wurde, weil sie alle unter seinen Wutausbrüchen liften.		1.8	/f/-/∅/
LP	F029D	Frau Geiger will bloß nicht alt werden und Falten bekommen, deswegen lässt sie sich jegliche Fältchen litten.		1.4	/f/-/∅/
HP	F030A	Celias kurzes Sommerkleid war genau richtig für die Hitzewelle, denn es war leicht und luftig	0.6	4.9	/f/-/s/
HP	F030B	Sauer macht lustig	0.9	4.6	/f/-/s/
LP	F030C	Celias kurzes Sommerkleid war genau richtig für die Hitzewelle, denn es war leicht und lustig.		2.3	/f/-/s/
LP	F030D	Sauer macht luftig.		1	/f/-/s/
HP	F031A	Wenn du Glück in der Liebe haben willst, dann küss deinen Partner unter dem Zweig einer Mistel	0.7	4.7	/s/-/∅/
HP	F031B	Um seine Geschäftsidee zu verwirklichen, fehlen Anton noch die finanziellen Mittel	1	4.8	/s/-/∅/
LP	F031C	Wenn du Glück in der Liebe haben willst, dann küss deinen Partner unter dem Zweig einer Mittel.		1.1	/s/-/∅/

Table A.1 continued from previous page

pred.	item#	item	cloze	plaus.	contrast
LP	F031D	Um seine Geschäftsidee zu verwirklichen, fehlen Anton noch die finanziellen Mistel.		1.1	/s/-/∅/
HP	F032A	Viele Menschen sind morgens schlecht gelaunt, also nennt man sie auch Muffel	0.9	4.5	/f/-/ʃ/
HP	F032B	Eliza findet am Strand eine schöne Muschel	0.9	5	/f/-/ʃ/
LP	F032C	Viele Menschen sind morgens schlecht gelaunt, also nennt man sie auch Muschel.		1.3	/f/-/ʃ/
LP	F032D	Eliza findet am Strand eine schöne Muffel.		1.2	/f/-/ʃ/
HP	F033A	In der griechischen Mythologie nennt man im Wasser lebende Frauen nicht Meerjungfrauen, sondern Nixen	0.5	4.7	/s/-/∅/
HP	F033B	Paul stimmte Laura zu mit einem kurzen, einfachen, stummen Nicken	0.8	4.3	/s/-/∅/
LP	F033C	In der griechischen Mythologie nennt man im Wasser lebende Frauen nicht Meerjungfrauen, sondern Nicken.		1.6	/s/-/∅/
LP	F033D	Paul stimmte Laura zu mit einem kurzen, einfachen, stummen Nixen.		1.1	/s/-/∅/
HP	F034A	Wenn es geregnet hat, gibt es auf der Autobahn häufig Aquaplaning wegen der Nässe	0.4	4.6	/f/-/s/
HP	F034B	Der Sohn meiner Schwester ist mein Neffe	0.9	4.6	/f/-/s/
LP	F034C	Wenn es geregnet hat, gibt es auf der Autobahn häufig Aquaplaning wegen der Neffe.		1.1	/f/-/s/
LP	F034D	Der Sohn meiner Schwester ist mein Nässe.		1	/f/-/s/
HP	F035A	Hannas Bauch war ganz rot, weil sie mit dem Bauch aufs Wasser platschte	0.3	3.9	/ʃ/-/ts/
HP	F035B	Der Knall war sehr laut als der Reifen platzte	0.7	4.6	/ʃ/-/ts/
LP	F035C	Hannas Bauch war ganz rot, weil sie mit dem Bauch aufs Wasser platzte.		2.1	/ʃ/-/ts/
LP	F035D	Der Knall war sehr laut als der Reifen platschte.		1.3	/ʃ/-/ts/
HP	F036A	Das Ruhrgebiet nennt man alternativ auch Kohlepott, Ruhrpott oder einfach nur Pott	0.7	4.9	/s/-/∅/
HP	F036B	Um den Brief als Einschreiben loszuschicken, ging Sarah zur Post	0.9	5	/s/-/∅/
LP	F036C	Das Ruhrgebiet nennt man alternativ auch Kohlepott, Ruhrpott oder einfach nur Post.		1.5	/s/-/∅/
LP	F036D	Um den Brief als Einschreiben loszuschicken, ging Sarah zur Pott.		1.1	/s/-/∅/
HP	F037A	Im Gottesdienst liest der Priester aus der Bibel einige Verse der beliebtesten Psalme	0.5	4.5	/s/-/∅/
HP	F037B	Am Karibikstrand liegt Margarita am liebsten unter einer Palme	1	4.6	/s/-/∅/
LP	F037C	Im Gottesdienst liest der Priester aus der Bibel einige Verse der beliebtesten Palme.		1.2	/s/-/∅/

Table A.1 continued from previous page

pred.	item#	item	cloze	plaus.	contrast
LP	F037D	Am Karibikstrand liegt Margarita am liebsten unter einer Psalme.		1	/s/-/∅/
HP	F038A	Beim Après-Ski trinken viele Leute alkoholische Getränke und sind schnell betrunken von kleinen, aus Fallfrüchten gebrannten Schnäpsen	0.6	4.2	/s/-/x/
HP	F038B	Barbaras roter Mantel war 50 % reduziert und deswegen ein richtiges Schnäppchen	1	4.6	/s/-/x/
LP	F038C	Beim Après-Ski trinken viele Leute alkoholische Getränke und sind schnell betrunken von kleinen, aus Fallfrüchten gebrannten Schnäppchen.		2.9	/s/-/x/
LP	F038D	Barbaras roter Mantel war 50 % reduziert und deswegen ein richtiges Schnäpsen.		1.3	/s/-/x/
HP	F039A	Auf dem Pausenhof brachten sich die Kinder gegenseitig bei, mit der Zunge zu schnalzen	0.6	4.3	/s/-/∅/
HP	F039B	Weil alles teurer geworden ist, mussten sie den Gürtel enger schnallen	0.9	4.6	/s/-/∅/
LP	F039C	Auf dem Pausenhof brachten sich die Kinder gegenseitig bei, mit der Zunge zu schnallen.		2	/s/-/∅/
LP	F039D	Weil alles teurer geworden ist, mussten sie den Gürtel enger schnalzen.		1.2	/s/-/∅/
HP	F040A	Siehst du, wie Irene die Hühner in den Stall scheucht	0.4	4.4	/ts/-/∅/
HP	F040B	Karls Partys sind bei seinen Freunden immer sehr beliebt, da er keine Kosten und Mühen scheut	0.7	4.5	/ts/-/∅/
LP	F040C	Siehst du, wie Irene die Hühner in den Stall scheut.		2	/ts/-/∅/
LP	F040D	Karls Partys sind bei seinen Freunden immer sehr beliebt, da er keine Kosten und Mühen scheucht.		1.3	/ts/-/∅/
HP	F041A	Der Professor ermahnte die müden Studenten, weil sie in der Vorlesung schliefen	0.7	5	/f/-/s/
HP	F041B	Um seine kleine Tochter nicht zu wecken, gab sich Max die größte Mühe, die Tür leise zu schließen	1	4.9	/f/-/s/
LP	F041C	Der Professor ermahnte die müden Studenten, weil sie in der Vorlesung schließen.		1.1	/f/-/s/
LP	F041D	Um seine kleine Tochter nicht zu wecken, gab sich Max die größte Mühe, die Tür leise zu schliefen.		1	/f/-/s/
HP	F042A	Nach der fünften Klasse kommt man in die sechste	0.7	4.3	/s/-/∅/
HP	F042B	Scientology ist eine Sekte	0.9	4.5	/s/-/∅/
LP	F042C	Nach der fünften Klasse kommt man in die Sekte.		1.2	/s/-/∅/
LP	F042D	Scientology ist eine sechste.		1	/s/-/∅/
HP	F043A	Leonie läuft es kalt den Rücken runter, wenn sie Schlangen sieht	0.9	5	/s/-/∅/
HP	F043B	In Deutschland ist es üblich, dass man ältere Personen nicht mit "du" anspricht, sondern sie siezt	0.9	4.8	/s/-/∅/

Table A.1 continued from previous page

pred.	item#	item	cloze	plaus.	contrast
LP	F043C	Leonie läuft es kalt den Rücken runter, wenn sie Schlangen siezt.		1.1	/s/-/∅/
LP	F043D	In Deutschland ist es üblich, dass man ältere Personen nicht mit "du" anspricht, sondern sie sieht.		1.2	/s/-/∅/
HP	F044A	Wespen haben hinten einen spitzen Stachel	1	4.5	/x/-/f/
HP	F044B	Hannah schaut gerade ihre Lieblingsserie und ist bei der letzten Folge der zweiten Staffel	1	4.7	/x/-/f/
LP	F044C	Wespen haben hinten einen spitzen Staffel.		1.6	/x/-/f/
LP	F044D	Hannah schaut gerade ihre Lieblingsserie und ist bei der letzten Folge der zweiten Stachel.		1	/x/-/f/
HP	F045A	Christian bewarb sich nach dem Studium auf viele Stellen	0.7	4.9	/s/-/∅/
HP	F045B	Die Clowns im Zirkus sind so groß, denn sie stolzieren auf Stelzen	1	3.7	/s/-/∅/
LP	F045C	Christian bewarb sich nach dem Studium auf viele Stelzen.		1.1	/s/-/∅/
LP	F045D	Die Clowns im Zirkus sind so groß, denn sie stolzieren auf Stellen.		1.2	/s/-/∅/
HP	F046A	Die Pfoten eines Tigers nennt man Tatzen	0.7	4.5	/ʃ/-/ts/
HP	F046B	Tee trinkt man normalerweise aus Tassen	0.9	4.6	/ʃ/-/ts/
LP	F046C	Die Pfoten eines Tigers nennt man Tassen.		1	/ʃ/-/ts/
LP	F046D	Tee trinkt man normalerweise aus Tatzen.		1.1	/ʃ/-/ts/
HP	F047A	Der Pfarrer muss diese Woche drei Babys taufen	0.9	5	/f/-/ʃ/
HP	F047B	Lisa gefiel Peters Geschenk besser und Peter gefiel Lisas Geschenk besser, also beschlossen sie zu tauschen	1	4.2	/f/-/ʃ/
LP	F047C	Der Pfarrer muss diese Woche drei Babys tauschen.		1	/f/-/ʃ/
LP	F047D	Lisa gefiel Peters Geschenk besser und Peter gefiel Lisas Geschenk besser, also beschlossen sie zu taufen.		1.3	/f/-/ʃ/
HP	F048A	Das Paar wollte beim Flug unbedingt nebeneinandersitzen, weswegen ein Passagier netterweise mit ihnen den Platz tauschte	0.9	4.8	/ʃ/-/∅/
HP	F048B	Marc und Stefan wollten heute Abend Fleisch aus ihrer Gefriertruhe essen, doch es war noch gefroren und sie mussten warten, bis das Fleisch taute	1	4	/ʃ/-/∅/
LP	F048C	Das Paar wollte beim Flug unbedingt nebeneinandersitzen, weswegen ein Passagier netterweise mit ihnen den Platz taute.		1.2	/ʃ/-/∅/
LP	F048D	Marc und Stefan wollten heute Abend Fleisch aus ihrer Gefriertruhe essen, doch es war noch gefroren und sie mussten warten, bis das Fleisch tauschte.		1.1	/ʃ/-/∅/
HP	F049A	Um den neuen Toaster ausprobieren zu können, brauchen wir unbedingt frischen Toast	0.7	4.7	/s/-/∅/

Table A.1 continued from previous page

pred.	item#	item	cloze	plaus.	contrast
HP	F049B	Weil Hannelore schon sehr alt ist, hat sie am meisten Angst vor dem Tod	0.8	4.7	/s/-/∅/
LP	F049C	Um den neuen Toaster ausprobieren zu können, brauchen wir unbedingt frischen Tod.		1.2	/s/-/∅/
LP	F049D	Weil Hannelore schon sehr alt ist, hat sie am meisten Angst vor dem Toast.		1.8	/s/-/∅/
HP	F050A	Viele Mädchen fangen ab einem bestimmten Alter an sich die Wimpern farbig zu tuschen	0.4	4.7	/pf/-/ʃ/
HP	F050B	Es ist wichtig, das Blut nicht schnell wegzuwischen, sondern vorsichtig mit einem Tuch von der Wunde zu tupfen	0.8	4.2	/pf/-/ʃ/
LP	F050C	Viele Mädchen fangen ab einem bestimmten Alter an sich die Wimpern farbig zu tupfen.		2.6	/pf/-/ʃ/
LP	F050D	Es ist wichtig, das Blut nicht schnell wegzuwischen, sondern vorsichtig mit einem Tuch von der Wunde zu tuschen.		2.5	/pf/-/ʃ/
HP	F051A	Nach dem Krieg gab es viele Menschen, die ihre zerstörte Heimat verließen	0.6	5	/ʃ/-/∅/
HP	F051B	Damit ihre Schwester nicht heimlich in ihrem Tagebuch lesen konnte, musste Lea es immer sicherheitshalber mit einem Schlüssel verschließen	0.8	4.8	/ʃ/-/∅/
LP	F051C	Nach dem Krieg gab es viele Menschen, die ihre zerstörte Heimat verschließen.		1.6	/ʃ/-/∅/
LP	F051D	Damit ihre Schwester nicht heimlich in ihrem Tagebuch lesen konnte, musste Lea es immer sicherheitshalber mit einem Schlüssel verließen.		1.9	/ʃ/-/∅/
HP	F052A	Daniel und Anne sind zu spät zur Vorlesung gekommen, weil sie den Wecker nicht hörten und deshalb verschliefen	0.9	4.7	/ʃ/-/∅/
HP	F052B	Ihr Handy ging aus und eine Karte hatten sie auch nicht dabei, weswegen sie sich verliefen	0.7	3.8	/ʃ/-/∅/
LP	F052C	Daniel und Anne sind zu spät zur Vorlesung gekommen, weil sie den Wecker nicht hörten und deshalb verliefen.		1.3	/ʃ/-/∅/
LP	F052D	Ihr Handy ging aus und eine Karte hatten sie auch nicht dabei, weswegen sie sich verschliefen.		1.2	/ʃ/-/∅/
HP	F053A	Wenn man von der Polizei festgenommen wird, bringen sie einen auf die Wache	0.7	4.5	/x/-/f/
HP	F053B	Die Polizisten mussten den Bankräuber in Notwehr erschießen, denn er bedrohte sie mit seiner gezückten Waffe	0.7	4.7	/x/-/f/
LP	F053C	Wenn man von der Polizei festgenommen wird, bringen sie einen auf die Waffe.		1.3	/x/-/f/

Table A.1 continued from previous page

pred.	item#	item	cloze	plaus.	contrast
LP	F053D	Die Polizisten mussten den Bankräuber in Notwehr erschießen, denn er bedrohte sie mit seiner gezückten Wache.		1.5	/x/-/f/
HP	F054A	Ich bade gerne in der großen Wanne	0.7	5	/s/-/∅/
HP	F054B	Auf der Mauer, auf der Lauer liegt ne kleine Wanze	0.7	4	/s/-/∅/
LP	F054C	Ich bade gerne in der großen Wanze.		1.2	/s/-/∅/
LP	F054D	Auf der Mauer, auf der Lauer liegt ne kleine Wanne.		1.3	/s/-/∅/
HP	F055A	Marco mag es überhaupt nicht, Wäsche zu waschen	0.9	5	/ʃ/-/x/
HP	F055B	Abends soll der Hund vor Einbrechern schützen und über das Haus wachen	1	4.3	/ʃ/-/x/
LP	F055C	Marco mag es überhaupt nicht, Wäsche zu wachen.		1.9	/ʃ/-/x/
LP	F055D	Abends soll der Hund vor Einbrechern schützen und über das Haus waschen.		1.1	/ʃ/-/x/
HP	F056A	In Herr Meiers Garten stehen nun so viele kleine Tannen, er hat schon fast sein eigenes kleines Wäldchen	0.6	4.8	/x/-/∅/
HP	F056B	Außerirdische wollen die Menschheit ausrotten in Steven Spielbergs Film "Krieg der Welten	0.7	4.4	/x/-/∅/
LP	F056C	In Herr Meiers Garten stehen nun so viele kleine Tannen, er hat schon fast sein eigenes kleines Welten.		1.2	/x/-/∅/
LP	F056D	Außerirdische wollen die Menschheit ausrotten in Steven Spielbergs Film "Krieg der Wäldchen."		2.6	/x/-/∅/
HP	F057A	Bei Stand-up-Comedy bringen Komiker das Publikum zum Lachen mit Geschichten und Witzen	0.6	3.8	/ʃ/-/ts/
HP	F057B	Professoren verdienen ihr Geld mit dem Vermitteln von Wissen	0.7	4.1	/ʃ/-/ts/
LP	F057C	Bei Stand-up-Comedy bringen Komiker das Publikum zum Lachen mit Geschichten und Wissen.		3.3	/ʃ/-/ts/
LP	F057D	Professoren verdienen ihr Geld mit dem Vermitteln von Witzen.		2.2	/ʃ/-/ts/
HP	F058A	Man kann ein DIN A4-Blatt mehrmals zusammenfalten	0.8	5	/f/-/h/
HP	F058B	In dieser schwierigen Zeit müssen die Freunde zusammenhalten	0.9	4.6	/f/-/h/
LP	F058C	Man kann ein DIN A4-Blatt mehrmals zusammenhalten.		1.2	/f/-/h/
LP	F058D	In dieser schwierigen Zeit müssen die Freunde zusammenfalten.		1.4	/f/-/h/
HP	F059A	Luisa hat so ein Glück im Leben und muss sich für nichts anstrengen - es wirkt so als ob ihr alles einfach zufällt	0.4	4.6	/f/-/h/
HP	F059B	Lars lacht zu laut, sodass Anna ihm den Mund zuhält	0.7	3.9	/f/-/h/
LP	F059C	Luisa hat so ein Glück im Leben und muss sich für nichts anstrengen - es wirkt so als ob ihr alles einfach zuhält.		1.2	/f/-/h/
LP	F059D	Lars lacht zu laut, sodass Anna ihm den Mund zufällt.		1.3	/f/-/h/
HP	F060A	Theo sah das Kind mit dem Messer rennen und es wunderte ihn, dass die Eltern das zuließen	0.3	4.9	/ʃ/-/∅/

Table A.1 continued from previous page

pred.	item#	item	cloze	plaus.	contrast
HP	F060B	Wenn man das Haus verlässt, sollte man die Türe immer zuschließen	0.8	4.6	/ʃ/-/∅/
LP	F060C	Theo sah das Kind mit dem Messer rennen und es wunderte ihn, dass die Eltern das zuschließen.		1.7	/ʃ/-/∅/
LP	F060D	Wenn man das Haus verlässt, sollte man die Türe immer zuließen.		1.8	/ʃ/-/∅/

Appendix B

Ordinal Regression Confidence Ratings (Exp 2.)

We conducted ordinal regressions of all analyses involving binarized confidence ratings in the manuscript. We find the same significant effects in the same directions as in the binarized analyses, with some exceptions. In the subset of quiet items, we find a negative effect of trial number, which we did not find before. Participants become less confident of their distractor responses as the experimental blocks went on. In the subset of wrong responses, we now find a significant effect of age, with lower confidence ratings for older participants. Finally, in the subset of target responses we find two new effects: We find a significant interaction of predictability and noise at -5SNR, leading to higher confidence ratings. Beyond the positive effect of the three-way interaction of predictability, age, and noise at 0SNR, we now also find a positive effect for this three-way interaction for noise at -5SNR.

The ordinal regressions were implemented in the ordinal package (Christensen, 2015) in R (R Core Team, 2022). We ran cumulative link mixed models (CLMM) that included the same fixed and random effects as the GLMMs with the binarized confidence ratings, which will be repeated below. First, we ran separate models on the three different response types (target, distractor, and wrong responses), and secondly, we ran separate models on the three subsets of distractor responses for the different noise types (quiet, 0SNR, -5SNR).

The model for the subset of target responses included fixed effects of Predictability (categorical predictor with two levels using dummy coding, mapping the High Pre-

Table B.1: Model Outcomes for the Confidence Rating Analysis (Target Subset)

	Estimate	*SE*	*z*-value	*p*-value	
Predictability LP	-2.50	0.24	-10.59	<.001	***
Noise 0SNR	-2.03	0.23	-8.86	<.001	***
Noise -5SNR	-4.27	0.26	-16.45	<.001	***
Age	0.15	0.23	0.65	.52	
Trial No	-0.05	0.04	-1.22	.22	
ContrastVP V	0.26	0.14	1.87	.06	.
Predictability LP : Noise 0 SNR	-0.02	0.24	-0.07	.94	
Predictability LP : Noise -5SNR	**1.21**	**0.28**	**4.33**	**<.001**	*******
Predictability LP : Age	-0.33	0.20	-1.65	.10	.
Noise 0SNR : Age	-0.75	0.20	-3.69	<.001	***
Noise -5SNR : Age	-0.80	0.22	-3.56	<.001	***
Predictability LP : Noise 0SNR : Age	0.75	0.22	3.46	<.001	***
Predictability LP : Noise -5SNR : Age	**0.84**	**0.26**	**3.30**	**<.001**	*******

Note. This table shows the analysis for the subset of target items. The response variable is the participants' confidence (high or low). Rows in bold denote new effects compared to the binomial analyses in Chapter 4.

dictability condition on the intercept), Noise (categorical predictor with three levels, mapping Quiet to the intercept), Age (scaled), Trial Number (scaled), ContrastVP (categorical predictor with two levels using dummy coding, mapping Plosive to the intercept), as well as the three-way interaction of Predictability, Noise, and Age. A by-Participant random intercept was included with random slopes for Noise and Predictability, and a by-Item random intercept with a random slope for Predictability. The model revealed a significant effect of Predictability, with lower confidence in LP versus HP (β = -2.50, SE = 0.24, z = -10.59, $p < .001$). The model revealed lower confidence in Noise compared to Quiet (β = -2.03, SE = 0.23, z = -8.86, $p < .001$ for 0SNR, and β = -4.26, SE = 0.26, z = -8.86, $p < .001$ for -5SNR). We find a significant interaction for Predictability and Noise, but only for the -5SNR condition, with higher confidence ratings in LP items (β = 1.21, SE = 0.28, z = 4.33, $p < .001$). The interaction of Noise and Age was significant, with lower confidence for older participants in noise (β = -0.75, SE = 0.20, z = -3.69, $p < .001$ for 0SNR, and β = -0.80, SE = 0.22, z = -3.56, $p < .001$ for -5SNR). Finally, the three-way interaction of Predictability, Noise, and Age was significant, with higher confidence ratings with age in LP in noise (β = 0.75, SE = 0.22, z = 3.46, $p < .001$ for 0SNR, β = 0.84, SE = 0.26, z = 3.30, $p < .001$ for -5SNR). The other effects were not significant (all p-values $> .06$), all effects can be found in Table B.1.

The model for the subset of distractor responses included the same fixed effects as the model on the subset of target responses. Non-significant interactions were re-

Table B.2: Model Outcomes for the Confidence Rating Analysis (Distractor Subset)

	Estimate	SE	z-value	p-value	
Predictability LP	0.47	0.44	1.07	.29	
Noise 0SNR	0.04	0.19	0.22	.83	
Noise -5SNR	-1.29	0.20	-6.57	<.001	***
Age	1.07	0.21	0.52	.60	
Trial No	-0.16	0.06	-2.70	<.01	**
ContrastVP V	-0.47	0.18	-2.66	<.01	**
Noise -5SNR : Age	-0.57	0.19	-3.07	<.01	**
Noise 0SNR : Age	-0.35	0.19	-1.87	.06	.

Note. This table shows the analysis for the subset of distractor items. The response variable is the participants' confidence (high or low).

moved, so that only the interaction of Noise and Age was included. A by-Participant random intercept was included, as well as a by-Item random intercept with a random slope of Predictability. The model revealed a significant effect of background noise, but only for -5SNR (β = -1.29, SE = 0.20, z = -6.57, $p < .001$). There is a significant effect of Vowel/Plosive contrast (β = -0.47, SE = 0.18, z = -2.66, $p < .01$), suggesting that participants were less confident about their answers on items that had a vowel contrast, rather than those with a plosive contrast. Additionally, there was a significant effect of Trial Number, where participants are less confident in later trials (β = -0.16, SE = 0.06, z = -2.70, $p < .01$). The interaction of Noise and Age was significant, but only for the -5SNR condition (β = -0.57, SE = 0.19, z = -3.07, $p < .01$), suggesting that for -5SNR confidence was lower with increasing age. The other effects were not significant (all p-values $> .06$). All effects can be seen in Table B.2.

The model for the subset of wrong answer items included the same fixed effects as the previous two models, except that this model did not include any interaction effects. A by-Participant random intercept was included, as well as a by-Item random intercept. The model revealed a significant effect for both noise conditions, with lower confidence ratings in noise (β = -1.47, SE = 0.28, z = -5.28, $p < .001$ for 0SNR, and β = -2.71, SE = 0.28, z = -9.84, $p < .001$ for -5SNR). Additionally, the model showed a significant effect of Age, with lower confidence ratings for older participants (β = -0.37, SE = 0.12, z = -3.06, $p < .01$). None of the other effects were significant (all p-values $> .54$), and all effects are presented in Table B.3.

The model on the subset of 0SNR trials included fixed effects of Predictability, Age, Trial Number, and ContrastVP (all coded and scaled as before). The model also included random intercepts for Subject and Item. The model showed signifi-

Table B.3: Model Outcomes for the Confidence Rating Analysis (Wrong Subset)

	Estimate	SE	z-value	p-value	
Predictability LP	-0.09	0.17	-0.54	.54	
Noise 0SNR	-1.47	0.28	-5.28	<.001	***
Noise -5SNR	-2.71	0.28	-9.84	<.001	***
Age	**-0.37**	**0.12**	**-3.06**	**<.01**	****
Trial No	0.01	0.08	0.13	.90	
ContrastVP V	-0.11	0.21	-0.55	.58	

Note. This table shows the analysis for the subset of wrong items. The response variable is the participants' confidence (high or low). Rows in bold denote new effects compared to the binomial analyses in Chapter 4.

cantly lower confidence as the trials went on (β = -0.24, SE = 0.10, z = -2.49, $p <$ .05). Additionally, confidence ratings were significantly lower for items with a Vowel contrast compared to items with a Plosive contrast (β = -0.67, SE = 0.24, z = -2.77, $p <$ .01). The other effects were not significant (all p-values $>$ .17).

The model on the subset of -5SNR trials consisted of the same fixed and random effects as the 0SNR model. We find only a significant effect of Age, where older participants are less confident of their responses than younger adults (β = -0.51, SE = 0.16, z = -3.18, $p <$ .01). None of the other effects were significant (all p-values $>$ .19).

The model on the quiet subset of the data again included the same fixed and random effects as the previous two models. We only find a significant effect of Trial (β = -0.45, SE = 0.20, z = -2.24, $p <$ .05). Other effects were not significant (all p-values $>$ .16). All outcomes from these three GLMMs are presented in Table 4.

The outcomes from this ordinal regression do not change our conclusions based on the binarized confidence ratings reported in Chapter 4. If anything, the additional significant effect of trial number in the quiet subset suggests that participants learned to trust their distractor responses less, while the negative effect of age in the wrong responses shows that older adults were less confident of their responses than younger adults. This is the opposite of what we would expect for false hearing. We do find positive effects on confidence ratings in the subset of target responses, but as these are correct responses, they do not signify false hearing.

Table B.4: Model Outcomes for the False Hearing Analysis

	Quiet subset				
	Estimate	SE	z-value	p-value	
Predictability LP	1.75	1.24	1.42	.16	
Age	0.24	0.25	0.93	.35	
Trial No	**-0.45**	**0.20**	**-2.24**	**<.05**	*****
ContrastVP V	-0.31	0.37	-0.84	.40	
	0SNR subset				
	Estimate	SE	z-value	p-value	
Predictability LP	0.32	0.81	0.39	.70	
Age	-0.22	0.16	-1.38	0.17	
Trial No	-0.24	0.10	-2.49	<.05	*
ContrastVP V	-0.67	0.24	-2.77	<.01	**
	-5SNR subset				
	Estimate	SE	z-value	p-value	
Predictability LP	0.71	0.69	1.02	.31	
Age	-0.51	0.16	-3.18	<.01	**
Trial No	0.03	0.09	0.28	.78	
ContrastVP V	-0.31	0.24	-1.32	.19	

Note. This table shows the analysis for the subset of distractor items in quiet, 0SNR, and -5SNR. The response variable is the participants' confidence (high or low). Rows in bold denote new effects compared to the binomial analyses in Chapter 4.

Appendix C

Stimuli Experiment 3

This Appendix presents the stimuli that were used in Experiment 3 (see Chapter 5). Table C.1 shows the sentences that were used in the listening task. The items are grouped per similar-sounding word pair and both word order versions of the sentence are given (SI = target word occurs in the beginning of the sentence; SF = the target word occurs towards the end of the sentence), as well as the surprisal values for the target word. Generally, the surprisal for the SI sentence is higher than for the SF sentence.

Table C.2 presents the items that were tested in the recognition memory test. For each target, the corresponding semantic lure is given. The right-most column presents the new items. For all items, the CELEX frequency (log per million words) is listed.

Finally, Table C.3 lists all filler items with their corresponding comprehension questions and the correct response.

Table C.1: Overview of the sentences used in the listening task in Experiment 3

item#	target	order	item	surprisal
1A	fatal	SI	Fatale Folgeschäden verursachte der Autounfall bei Sarah.	12.89
1B		SF	Der Autounfall verursachte bei Sarah fatale Folgeschäden.	11.91
1A	fötal	SI	Die fötale Entwicklung ist ab der sechsten Schwangerschaftswoche besonders ausgeprägt.	15.95
1B		SF	Besonders ausgeprägt ist ab der sechsten Schwangerschaftswoche die fötale Entwicklung.	9.18

Table C.1 continued from previous page

item#	target	order	item	surprisal
2A	Abszess	SI	Ein Abszess im Gesicht bildet sich oft bei schwer entzündeter Akne.	13.02
2B		SF	Bei schwer entzündeter Akne bildet sich im Gesicht oft ein Abszess.	6.10
2A	Prozess	SI	Ein Prozess gegen den Sexualstraftäter wird vor Gericht ausgetragen.	8.04
2B		SF	Vor Gericht wird gegen den Sexualstraftäter ein Prozess ausgetragen.	4.77
3A	Konfession	SI	Je nach Konfession werden die Schüler dem katholischen oder evangelischen Religionsunterricht zugeteilt.	8.85
3B		SF	Die Schüler werden dem katholischen oder evangelischen Religionsunterricht zugeteilt, je nach Konfession.	2.21
3A	Depression	SI	Die Depression gehört zu den häufigsten psychischen Erkrankungen.	10.85
3B		SF	Zu den häufigsten psychischen Erkrankungen gehört die Depression.	3.43
4A	Frisur	SI	Die Frisur wurde verändert, indem ein gutes Stück der Haare abgeschnitten wurde.	11.24
4B		SF	Indem ein gutes Stück der Haare abgeschnitten wurde, wurde die Frisur verändert.	4.56
4A	Fissur	SI	Eine Fissur entwickeln viele Menschen bei trockener, brüchiger Haut an der Ferse.	15.98
4B		SF	An der Ferse entwickeln viele Menschen bei trockener, brüchiger Haut eine Fissur.	7.71
5A	abrupfen	SI	Abrupfen sollte Lea die Blüten der frisch gesetzten Tulpen nicht.	29.35
5B		SF	Lea sollte die Blüten der frisch gesetzten Tulpen nicht abrupfen.	15.52
5A	abtupfen	SI	Abtupfen und desinfizieren sollte man eine Schürfwunde, um eine Entzündung zu verhindern.	19.06
5B		SF	Um eine Entzündung zu verhindern, sollte man eine Schürfwunde desinfizieren und abtupfen.	12.61
6A	Esche	SI	Die Esche zählt zu den wichtigsten heimischen Laubnutzhölzern in Mitteleuropa.	13.10
6B		SF	In Mitteleuropa zählt zu den wichtigsten heimischen Laubnutzhölzern die Esche.	4.30
6A	Wäsche	SI	Viel Wäsche sammelt sich an, wenn man jeden Tag die Kleidung wechselt.	11.30
6B		SF	Wenn man jeden Tag die Kleidung wechselt, sammelt sich viel Wäsche an.	7.25
8A	Vakzine	SI	Vakzine wurden zum Schutz gegen Viren entwickelt.	16.64

Table C.1 continued from previous page

item#	target	order	item	surprisal
8B		SF	Zum Schutz gegen Viren wurden Vakzine entwickelt.	10.56
8A	Magazine	SI	Magazine liegen häufig in Wartezimmern beim Arzt aus.	13.30
8B		SF	In Wartezimmern beim Arzt liegen häufig Magazine aus.	9.57
9A	Dialyse	SI	Die Dialyse übernimmt beim Versagen der Nieren deren Blutreinigungsfunktion.	11.96
9B		SF	Beim Versagen der Nieren übernimmt deren Blutreinigungsfunktion die Dialyse.	7.25
9A	Analyse	SI	Zur Analyse zieht der Wissenschaftler qualitative oder quantitative Methoden heran	7.42
9B		SF	Der Wissenschaftler zieht quantitaive oder qualitative Methoden zur Analyse heran.	2.67
10A	Addition	SI	Addition heißt in der Mathematik das Summieren von Zahlen.	12.67
10B		SF	Das Summieren von Zahlen heißt in der Mathematik Addition.	6.80
10A	Absorption	SI	Zu Absorption und Reflexion von Licht kommt es bei dessen Auftreten auf einer Oberfläche.	15.79
10B		SF	Bei Auftreten von Licht auf einer Oberfläche kommt es zu dessen Reflexion und Absorption.	4.57
12A	Rast	SI	Eine Rast erleichtert oft lange und eintönige Autofahrten.	10.65
12B		SF	Lange und eintönige Autofahrten erleichtert oft eine Rast.	9.05
12A	Mast	SI	Der Mast eines Bootes ist die Stange, an der das Großsegel hängt.	9.62
12B		SF	Die Stange eines Bootes, an der das Großsegel hängt, ist der Mast.	3.35
13A	Barmherzigkeit	SI	Barmherzigkeit ist ein Begriff, der oft im Zusammenhang mit Gott und Religion verwendet wird.	14.34
13B		SF	Ein Begriff, der oft im Zusammenhang mit Gott und Religion verwendet wird, ist Barmherzigkeit.	9.91
13A	Warmherzigkeit	SI	Ihrer Warmherzigkeit und Freundlichkeit wegen wird Celia von jedem gemocht.	15.53
13B		SF	Celia wird von jedem gemocht wegen ihrer Freundlichkeit und Warmherzigkeit.	8.21
14A	Gastronome	SI	Gastronome sind Leiter von Restaurants mit besonderen Kenntnissen über Essen und Bewirtung von Gästen.	17.03
14B		SF	Leiter von Restaurants mit besonderen Kenntnissen über Essen und Bewirtung von Gästen sind Gastronome.	13.68

Table C.1 continued from previous page

item#	target	order	item	surprisal
14A	Hämatome	SI	Hämatome und schwere Rippenprellungen hatte Tim nachdem er von der Leiter fiel.	15.38
14B		SF	Nachdem er von der Leiter fiel, hatte Tim schwere Rippenprellungen und Hämatome.	4.14
15A	Veganer	SI	Veganer verzichten auf Fleisch und jegliche sonstige tierische Produkte.	15.83
15B		SF	Auf Fleisch und jegliche sonstige tierische Produkte verzichten Veganer.	10.93
15A	Nirwana	SI	Das Nirwana bezeichnet im Buddhismus den Austritt aus dem Leiden und der Wiedergeburt.	14.44
15B		SF	Den Austritt aus dem Leiden und der Wiedergeburt bezeichnet im Buddhismus das Nirwana.	7.17
16A	Melone	SI	Die Melone ist eine kugelförmige Frucht der gleichnamigen subtropischen Pflanze.	12.77
16B		SF	Eine kugelförmige Frucht der gleichnamigen subtropischen Pflanze ist die Melone.	7.91
16A	Gallone	SI	Die Gallone ist eine Raumeinheit des angloamerikanischen Maßsystems.	15.31
16B		SF	Eine Raumeinheit des angloamerikanischen Maßsystems ist die Gallone.	11.45
17A	Insolvenz	SI	Insolvenz heißt der Zustand der Zahlungsunfähigkeit eines Unternehmens oder einer Privatperson.	12.00
17B		SF	Der Zustand der Zahlungsunfähigkeit eines Unternehmens oder einer Privatperson heißt Insolvenz.	4.75
17A	Audienz	SI	Eine Audienz ist ein offizieller Empfang bei Königen oder kirchlichen Persönlichkeiten.	10.04
17B		SF	Ein offizieller Empfang bei Königen oder kirchlichen Persönlichkeiten ist eine Audienz.	8.80
18A	Anagramm	SI	Ein Anagramm erhält man durch das Umstellen von Buchstaben innerhalb eines Wortes.	12.30
18B		SF	Durch das Umstellen von Buchstaben innerhalb eines Wortes erhält man ein Anagramm.	6.99
18A	Diagramm	SI	Diagramm nennt man eine bildliche Darstellung von Zahlen und Fakten.	12.91
18B		SF	Eine bildliche Darstellung von Zahlen und Fakten nennt man Diagramm.	10.57
19A	Konflikt	SI	Ein Konflikt ist eine Auseinandersetzung mehrerer Staaten, die häufig Waffengewalt beinhaltet.	8.97
19B		SF	Eine Auseinandersetzung mehrerer Staaten, die häufig Waffengewalt beinhaltet, ist ein Konflikt.	3.27
19A	Distrikt	SI	Die Distrikte gliedern sich nach den jeweiligen Stadtteilen.	11.82

Table C.1 continued from previous page

item#	target	order	item	surprisal
19B		SF	Nach den jeweiligen Stadtteilen gliedern sich die Distrikte.	5.59
20A	Augenoperation	SI	Eine Augenoperation braucht Lara, weil sie so schlecht sieht.	11.18
20B		SF	Weil sie so schlecht sieht, braucht Lara eine Augenoperation.	9.01
20A	Approximation	SI	Die Approximation ist ein bekanntes mathematisches Verfahren.	19.21
20B		SF	Ein bekanntes mathematisches Verfahren ist die Approximation.	6.43
21A	Manipulation	SI	Durch Manipulation beeinflusste Markus das Verhalten anderer Menschen.	9.92
21B		SF	Markus beeinflusste das Verhalten anderer Menschen durch Manipulation.	8.38
21A	Stipulation	SI	Die Stipulation ist ein Verbalvertrag, bei dem eine Partei eine formelhafte Frage formuliert, die von der anderen Partei bejaht wird.	17.23
21B		SF	Ein Verbalvertrag, bei dem eine Partei eine formelhafte Frage formuliert, die von der anderen Partei bejaht wird, ist die Stipulation.	12.36
22A	Glukose	SI	Glukose und Fruktose finden sich neben Haushaltszucker in unseren Lebensmitteln.	14.50
22B		SF	In unseren Lebensmitteln finden sich neben Haushaltszucker Fruktose und Glukose.	2.65
22A	Hose	SI	Die Hose sitzt durch das elastische Material wie angegossen.	10.36
22B		SF	Wie angegossen sitzt durch das elastische Material die Hose.	6.52
23A	Spasmus	SI	Spasmus nennt man einen Krampf im Muskel.	18.78
23B		SF	Einen Krampf im Muskel nennt man Spasmus.	8.00
23A	Sarkasmus	SI	Sarkasmus ist beißender, verletzender Spott, der oft unangebracht ist.	14.17
23B		SF	Beißender, verletzender Spott, der oft unangebracht ist, ist Sarkasmus.	12.84
24A	Typologie	SI	In der Typologie werden Objekte aufgrund gemeinsamer Eigenschaften klassifiziert.	11.93
24B		SF	Aufgrund gemeinsamer Eigenschaften werden Objekte in der Typologie klassifiziert.	12.68
24A	Theologie	SI	Die Theologie lehrt Inhalte eines spezifischen religiösen Glaubens.	10.74
24B		SF	Inhalte eines spezifischen religiösen Glaubens lehrt die Theologie.	4.43

Table C.1 continued from previous page

item#	target	order	item	surprisal
25A	Transposition	SI	Transposition heißt, etwas in eine andere Position zu verschieben.	16.67
25B		SF	Etwas in eine andere Position zu verschieben heißt Transposition.	14.56
25A	Tradition	SI	Tradition hat bei Familie Müller das Aufstellen eines Weihnachtsbaumes an Weihnachten.	9.51
25B		SF	An Weihnachten hat bei Familie Müller das Aufstellen eines Weihnachtsbaums Tradition.	4.56
26A	Folie	SI	Die Folie dient zum Schutz des Handys, damit der Bildschirm nicht zerkratzt wird.	10.66
26B		SF	Damit der Bildschirm nicht zerkratzt wird, dient zum Schutz des Handys die Folie.	5.67
26A	Magnolie	SI	Die Magnolie blüht häufig bereits früh im Frühjahr.	14.80
26B		SF	Früh im Frühjahr blüht häufig bereits die Magnolie.	7.10
27A	Silikon	SI	Mit Silikon ließ sich Lea vom Schönheitschirurgen die Brüste vergrößern.	11.87
27B		SF	Vom Schönheitschirurgen ließ sich Lea die Brüste mit Silikon vergrößern.	6.76
27A	Silikat	SI	Als Silikat bezeichnet man das Salz der Kieselsäure.	18.00
27B		SF	Das Salz der Kieselsäure bezeichnet man als Silikat.	14.74
28A	Melatonin	SI	Melatonin ist ein Hormon, das unseren Tag-Nacht-Rhythmus steuert.	13.50
28B		SF	Ein Hormon, das unseren Tag-Nacht-Rhythmus steuert ist Melatonin.	7.71
28A	Melancholie	SI	In Melancholie verfiel Luka als er an den Tod seiner Oma dachte.	15.86
28B		SF	Als er an den Tod seiner Oma dachte, verfiel Luka in Melancholie.	7.60
29A	Gelatine	SI	Gelatine ist in Gummibärchen und Wackelpudding enthalten.	14.28
29B		SF	In Gummibärchen und Wackelpudding ist Gelatine enthalten.	10.24
29A	Gardine	SI	Die Gardinen müssen noch am Fenster aufgehängt werden.	12.07
29B		SF	Am Fenster aufgehängt werden müssen noch die Gardinen.	8.09
30A	Collagen	SI	Collagen entstehen durch das Aufkleben verschiedener Bilder auf ein Plakat.	16.03
30B		SF	Durch das Aufkleben verschiedener Bilder auf ein Plakat entstehen Collagen.	5.93
30A	Passagen	SI	Passagen sind kleine überdachte Ladenstraßen für Fußgänger.	14.46

Table C.1 continued from previous page

item#	target	order	item	surprisal
30B		SF	Kleine überdachte Ladenstraßen für Fußgänger sind Passagen.	11.42
31A	Falsett	SI	Das Falsett erlaubt Männern, in einer hohen Stimmlage zu singen.	14.65
31B		SF	In einer hohen Stimmlage zu singen, erlaubt Männern das Falsett.	10.13
31A	Ballett	SI	An Ballett gefallen Hanna am besten das Tutu und Pirouetten drehen.	14.81
31B		SF	Das Tutu und Pirouetten drehen gefallen Hanna am besten an Ballett.	12.70
32A	Jambus	SI	Der Jambus bezeichnet die Wortbetonung auf der zweiten Silbe.	16.33
32B		SF	Die Wortbetonung auf der zweiten Silbe bezeichnet der Jambus.	12.54
32A	Bambus	SI	Bambus ist eine schnellwachsende Pflanze und das Hauptnahrungsmittel von Pandas.	12.93
32B		SF	Eine schnellwachsende Pflanze und das Hauptnahrungsmittel von Pandas ist Bambus.	5.63
33A	Vitamin	SI	Lebensnotwendige Vitamine finden sich in Obst und Gemüse.	7.86
33B		SF	In Obst und Gemüse finden sich lebensnotwendige Vitamine.	2.75
34A	Exempel	SI	Ein Exempel nimmt sich die kleine Schwester an ihrem älteren Bruder, ihrem Vorbild.	10.08
34B		SF	An ihrem älteren Bruder, ihrem Vorbild, nimmt sich die kleine Schwester ein Exempel.	8.53
34A	Stempel	SI	Im Postamt erhielten früher alle Briefe einen Stempel.	9.64
34B		SF	Einen Stempel erhielten früher alle Briefe im Postamt.	1.27
35A	Zuckerdose	SI	Die Zuckerdose steht beim Kaffee und Kuchen auf dem Tisch.	16.24
35B		SF	Beim Kaffee und Kuchen steht auf dem Tisch die Zuckerdose.	9.72
35A	Zoonose	SI	Eine Zoonose ist das Coronavirus, das vermutlich von Fledermäusen auf den Menschen übertragen wurde.	15.02
35B		SF	Das Coronavirus, das vermutlich von Fledermäusen auf den Menschen übertragen wurde, ist eine Zoonose.	4.51
36A	Demos	SI	Bei Demos setzen sich Aktivisten für ihre Rechte ein.	12.58
36B		SF	Für ihre Rechte setzen sich Aktivisten bei Demos ein.	7.98
36A	Memos	SI	Memos können anstelle von Textnachrichten auf Whats App verschickt werden.	14.79

Table C.1 continued from previous page

item#	target	order	item	surprisal
36B		SF	Anstelle von Textnachrichten können auf Whats App Memos verschickt werden.	10.32
37A	Pragmatik	SI	Mit Pragmatik und logischen Schlussfolgerungen widmete Tom sich der Lösung des Problems.	16.47
37B		SF	Der Lösung des Problems widmete sich Tom mit logischen Schlussfolgerungen und Pragmatik.	9.44
37A	Informatik	SI	Informatik studiert Max, weil er sich schon immer für das Programmieren interessiert hat.	11.35
37B		SF	Weil er sich schon immer für das Programmieren interessiert hat, studiert Max Informatik.	11.26
38A	Gnom	SI	Der Gnom ist ein menschenähnliches, kleinwüchsiges Fabelwesen, das auch als Kobold bekannt ist.	12.30
38B		SF	Ein menschenähnliches, kleinwüchsiges Fabelwesen, das auch als Kobold bekannt ist, ist der Gnom.	6.62
38A	Strom	SI	Der Strom fiel wegen des Unwetters schon wieder aus.	8.30
38B		SF	Wegen des Unwetters fiel schon wieder der Strom aus.	2.81
39A	Tide	SI	Die Tide beschreibt das Steigen und Fallen des Wassers der Ozeane.	13.40
39B		SF	Das Steigen und Fallen des Wassers der Ozeane beschreibt die Tide.	9.42
39A	Friede	SI	Damit Friede einkehren kann, müssen sich die beiden Geschwister wieder miteinander versöhnen.	13.61
39B		SF	Die beiden Geschwister müssen sich wieder miteinander versöhnen, damit Friede einkehren kann.	11.41
40A	Harmonie	SI	Harmonie ist besonders wichtig für einen friedlichen und ausgeglichenen Haushalt.	8.92
40B		SF	Besonders wichtig für einen friedliche und ausgeglichene Beziehung ist Harmonie.	3.82
40A	Pneumonie	SI	Pneumonie nennt der Arzt eine akut oder chronisch verlaufende Entzündung des Lungengewebes.	16.06
40B		SF	Eine akut oder chronisch verlaufende Entzündung des Lungegewebes nennt der Arzt Pneumonie.	8.30
41A	Palliativ	SI	Ein Palliativ gibt man Sterbenden zur Linderung der Schmerzen.	13.17
41B		SF	Zur Linderung der Schmerzen gibt man Sterbenden ein Palliativ.	11.59
41A	Stativ	SI	Auf das Stativ stellt man die Kamera, damit man schöne, unverwackelte Fotos machen kann.	10.76
41B		SF	Damit man schöne, unverwackelte Fotos machen kann, stellt man die Kamera auf das Stativ.	7.06

Table C.1 continued from previous page

item#	target	order	item	surprisal
42A	Drainage	SI	Eine Drainage wird zum Ableiten von überflüssigem Wasser gelegt, damit der Rasen nicht überwässert wird.	11.93
42B		SF	Damit der Rasen nicht überwässert wird, wird zum Ableiten von überflüssigem Wasser eine Drainage gelegt.	3.74
42A	Massage	SI	Eine Massage dient zur Entspannung der Muskeln.	9.91
42B		SF	Zur Entspannung der Muskeln dient eine Massage.	2.04
43A	Wunder	SI	Ein Wunder war die Geburt der Zwillinge, da die Frau 6 Jahre lang als unfruchtbar galt.	7.36
43B		SF	Da die Frau 6 Jahre lang als unfruchtbar galt, war die Geburt der Zwillinge ein Wunder.	4.74
43A	Zunder	SI	Zunder heißt leicht brennbares Material, das zum Feuer anzünden verwendet wird.	14.98
43B		SF	Leicht brennbares Material, das auch zum Feuer anzünden verwendet wird, heißt Zunder.	8.28
44A	Prognose	SI	Eine Prognose wird erstellt, um Entwicklungen einer Krankheit vorauszusagen.	7.58
44B		SF	Um Entwicklungen einer Krankheit vorauszusagen, wird eine Prognose erstellt.	3.22
44A	Franzose	SI	Der Franzose hatte wie immer Croissants und Camembert für das Buffet mitgebracht.	7.37
44B		SF	Croissants und Camembert für das Buffet hatte wie immer der Franzose mitgebracht.	5.50
45A	Demografie	SI	Die Demografie beschäftigt sich mit den Veränderungen von Geburtenverhalten, Alterung und Migration.	12.99
45B		SF	Mit den Veränderungen von Geburtenverhalten, Alterung und Migration beschäftigt sich die Demografie.	9.09
45A	Demokratie	SI	Demokratie und freie Meinungsäußerung herrschen in allen Staaten der EU.	11.19
45B		SF	In allen Staaten der EU herrschen freie Meinungsäußerung und Demokratie.	4.40
46A	Suizide	SI	Suizide werden von psychisch kranken Menschen begangen, wenn sie aus Verzweiflung keinen anderen Ausweg mehr sehen.	16.91
46B		SF	Wenn sie aus Verzweiflung keinen anderen Ausweg mehr sehen, begehen psychisch kranke Menschen Suizide.	10.66
46A	Sulfide	SI	Als Sulfide bezeichnet man Salze des Schwefelwasserstoffes.	16.76
46B		SF	Salze des Schwefelwasserstoffes bezeichnet man als Sulfide.	3.71

Table C.1 continued from previous page

item#	target	order	item	surprisal
47A	Histamin	SI	Histamin ist ein Stoff, der in leicht verderblichen tierischen Lebensmitteln vorkommt und oft Unverträglichkeiten verursacht.	14.57
47B		SF	Ein Stoff, der in leicht verderblichen tierischen Lebensmitteln vorkommt und oft Unverträglichkeiten verursacht, ist Histamin.	15.19
48A	Quaddel	SI	Eine Quaddel entsteht manchmal bei Nesselsucht oder Insektenstichen.	16.44
48B		SF	Bei Nesselsucht oder Insektenstichen entsteht manchmal eine Quaddel.	8.35
48A	Paddel	SI	Ein Paddel brauchen sie, um mit dem Kanu im See fahren zu können.	11.66
48B		SF	Um mit dem Kanu im See fahren zu können, brauchen sie ein Paddel.	4.14
49A	Plastik	SI	Durch Plastik werden Ozeane verschmutzt und Lebensräume zerstört.	11.54
49B		SF	Ozeane werden verschmutzt und Lebensräume zerstört durch Plastik.	4.98
49A	Stochastik	SI	Die Stochastik beschäftigt sich mit der Beschreibung und Untersuchung von Zufallsexperimenten.	12.82
49B		SF	Mit der Beschreibung und Untersuchung von Zufallsexperimenten beschäftigt sich die Stochastik.	10.68
50A	Klausur	SI	Eine Klausur oder eine Hausarbeit stehen am Ende jeder Vorlesung.	10.77
50B		SF	Am Ende jeder Vorlesung stehen eine Hausarbeit oder eine Klausur.	5.83
50A	Kreatur	SI	Eine Kreatur nennt man auch einen bedauernswerten oder verachtenswerten Menschen.	10.28
50B		SF	Einen bedauernswerten oder verachtenswerten Menschen nennt man auch eine Kreatur.	7.08
52A	Friseur	SI	Zum Friseur wollte Lina, weil ihre Haare zu lang und spröde waren.	10.50
52B		SF	Weil ihre Haare zu lang und spröde waren, wollte Lina zum Friseur.	3.29
52A	Malheur	SI	Ein Malheur passierte dem Mann, als er unglücklicherweise auf den Rock der Frau neben sich trat.	12.25
52B		SF	Als er unglücklicherweise auf den Rock der Frau neben sich trat, passierte dem Mann ein Malheur.	8.29
53A	Broschüren	SI	Die Broschüren zum Aussuchen von Hotels und Flügen liegen gut sortiert im Reisebüro.	11.21
53B		SF	Gut sortiert liegen im Reisebüro zum Aussuchen von Hotels und Flügen die Broschüren.	6.13

Table C.1 continued from previous page

item#	target	order	item	surprisal
53A	Allüren	SI	Robins Allüren und Launen nerven seine Freundin schon lange.	16.69
53B		SF	Schon lange nerven seine Freundin Robins Launen und Allüren.	8.91
54A	Rezidenz	SI	Eine Residenz als zweiten Wohnsitz kaufte Oliver auf Mallorca, um dort über den Sommer zu leben.	11.19
54B		SF	Um dort über den Sommer zu leben, kaufte Oliver auf Mallorca als zweiten Wohnsitz eine Residenz.	6.94
54A	Präzedenz	SI	Als Präzedenz taugt Tinas Fall, da er für zukünftige Fälle als Muster dienen kann.	12.09
54B		SF	Da Tinas Fall für zukünftige Fälle als Muster dienen kann, taugt er als Präzedenz.	7.39
56A	Emphatie	SI	Empathie heißt die Bereitschaft und Fähigkeit, sich in die Gefühlswelt anderer Menschen einzufühlen.	13.96
56B		SF	Die Bereitschaft und Fähigkeit, sich in die Gefühlswelt anderer Menschen einzufühlen, heißt Empathie.	7.98
56A	Antipathie	SI	Über Antipathie und Sympathie entscheidet oft der Klang der Stimme.	18.65
56B		SF	Der Klang der Stimme entscheidet oft über Sympathie und Antipathie.	7.73
57A	Parole	SI	Eine aufhetzende Parole gegen den Politiker wurde mit Graffiti an die Wand gesprüht.	6.73
57B		SF	An die Wand wurde mit Graffiti eine gegen den Politiker aufhetzende Parole gesprüht.	3.57
57A	Pistole	SI	Die Pistole richtete der Verbrecher auf Hans, um ihn zu erpressen.	10.22
57B		SF	Um ihn zu erpressen, richtete der Verbrecher die Pistole auf Hans.	1.80
60A	Hamburg	SI	Hamburg ist eine Hafenstadt im Norden Deutschlands, in die Lars schon immer mal verreisen wollte.	8.79
60B		SF	Eine Hafenstadt im Norden Deutschlands, in die Lars schon immer mal verreisen wollte, ist Hamburg.	4.31
60A	Humbug	SI	Völliger Humbug sind absurde Werbeaktionen, die versuchen einem ein Produkt anzudrehen.	8.83
60B		SF	Absurde Werbeaktionen, die versuchen einem ein Produkt anzudrehen, sind völliger Humbug.	6.94
61A	Autismus	SI	Autismus äußert sich in Schwierigkeiten im sozialen Umgang und der Kommunikation mit anderen Menschen.	14.15
61B		SF	In Schwierigkeiten im sozialen Umgang und der Kommunikation mit anderen Menschen äußert sich Autismus.	10.85

Table C.1 continued from previous page

item#	target	order	item	surprisal
61A	Altruismus	SI	Auf dem Altruismus der Bürger basiert zu Notzeiten die Stabilität der demokratischen Gemeinschaft.	18.44
61B		SF	Die Stabilität der demokratischen Gemeinschaft basiert zu Notzeiten auf dem Altruismus der Bürger.	11.74
62A	Jugend	SI	Tims Jugend und Kindheit waren unbeschwert, weil seine Eltern ihn immer umsorgt haben.	9.66
62B		SF	Weil ihn seine Eltern immer umsorgt haben, waren Tims Kindheit und Jugend unbeschwert.	0.44
62A	Tugend	SI	Eine Tugend ist eine erstrebenswerte Eigenschaft von Menschen, wie zum Beispiel die Geduld.	10.41
62B		SF	Eine erstrebenswerte Eigenschaft von Menschen, wie zum Beispiel die Geduld, ist eine Tugend.	3.79
63A	Banane	SI	Die Banane ist eine tropische Frucht, die in Deutschland gerne verzehrt wird.	11.89
63B		SF	Eine tropische Frucht, die in Deutschland gerne verzehrt wird, ist die Banane.	5.70
63A	Kleptomane	SI	Weil er ein Kleptomane ist, kann Tom nicht aufhören zu stehlen.	12.94
63B		SF	Tom kann nicht aufhören zu stehlen, weil er ein Kleptomane ist.	12.89
64A	Million	SI	Eine Million Euro ist der Hauptgewinn einer beliebten Fernsehshow mit Günther Jauch.	6.58
64B		SF	Der Hauptgewinn einer beliebten Fernsehshow mit Günther Jauch ist eine Million Euro.	3.13
64A	Mitigation	SI	Die Mitigation bezeichnet Maßnahmen, um Gefahren des Klimawandels abzuschwächen.	16.59
64B		SF	Maßnahmen, um Gefahren des Klimawandels abzuschwächen, bezeichnet die Mitigation.	12.54
65A	Fiktion	SI	Reine Fiktion war die Geschichte, sie entsprach nicht der Realität.	6.58
65B		SF	Die Geschichte entsprach nicht der Realität, sie war reine Fiktion.	2.19
65A	Kontradiktion	SI	Zur Kontradiktion kam es als beide Zeugen eine unterschiedliche Aussage trafen.	15.78
65B		SF	Als beide Zeugen eine unterschiedliche Aussage trafen kam es zur Kontradiktion.	15.80
66A	Parlament	SI	Das Parlament beschließt in Deutschland neue Gesetze.	5.70
66B		SF	Neue Gesetze beschließt in Deutschland das Parlament.	3.31
66A	Pergament	SI	Auf Pergament wurden früher sehr wertvolle Bücher geschrieben.	12.45

Table C.1 continued from previous page

item#	target	order	item	surprisal
66B		SF	Sehr wertvolle Bücher wurden früher auf Pergament geschrieben.	5.93
67A	Bühne	SI	Die Bühne dient zur Präsentation von Theateraufführungen und Konzerten.	8.72
67B		SF	Zur Präsentation von Theateraufführungen und Konzerten dient die Bühne.	5.10
67A	Sühne	SI	Eine Sühne ist eine Wiedergutmachung, die man leistet, wenn man etwas Unrechtes getan hat.	13.10
67B		SF	Eine Wiedergutmachung, die man leistet, wenn man etwas Unrechtes getan hat, ist eine Sühne.	9.61

Table C.2: Overview of the memory test items in Experiment 3

Item#	Target	Freq	Lure	Freq	New	Freq
1	fatal	0.699	katastrophal	0.9542	Becher	0.699
1	fötal	0	embryonal	0	Brand	1.4472
2	Abszess	0	Geschwür	0.301	Brikole	0
2	Prozess	2.0294	Verfahren	1.8633	Chat	0
3	Depression	0	Burn-out	0	Dynamit	0
3	Konfession	0.8451	Glaubensbekenntnis	0.301	Elster	0.301
4	Fissur	0	Hautriss	0	Feigling	0.4771
4	Frisur	0.301	Haarschnitt	0	Format	1.0414
5	abrupfen	0	pflücken	0	Geschwindigkeit	1.6228
5	abtupfen	0	säubern	0.699	Hilfestellung	4.771
6	Esche	0	Buche	0.6021	Klarinette	0.301
6	Wäsche	1.0414	Waschgänge	0	Klaustrophobie	0
8	Magazine	0.9542	Zeitschriften	1.699	Kleiderbügel	0
8	Vakzine	0	Impfung	0.4771	Kopfkissen	0
9	Analyse	1.3979	Studie	1.1461	Lokal	1.2304
9	Dialyse	0	Blutwäsche	0	Minute	2.3802
10	Absorption	0	Aufnahme	1.7782	Mistkäfer	0
10	Addition	0.301	Subtraktion	0	Mülleimer	0
12	Mast	0.4771	Pfosten	0.4771	Namenstag	0
12	Rast	0.301	Pause	1.6335	Ödland	0
13	Barmherzigkeit	0.4771	Vergebung	0.699	Paramilitär	0
13	Warmherzigkeit	0	Güte	1.1461	Paramimie	0
14	Gastronome	0	Wirte	1.2304	Park	1.5441
14	Hämatome	0	Bluterguss	0	Plakette	0
15	Nirwana	0	Jenseits	0.6021	Postsendung	0
15	Veganer	0	Vegetarier	0	Sauna	1.0414
16	Gallone	0	Unze	0.4771	Sicherheitsgurt	0
16	Melone	0.4771	Orange	0	Tarantel	0
17	Audienz	0.301	Bankett	0.4771	Toilette	1
17	Insolvenz	0.301	Bankrott	0	Torpedo	0
18	Anagramm	0	Schüttelwort	0	Trabant	0.301
18	Diagramm	0.301	Schaubild	0	Vanille	0.6021
19	Distrikt	0.301	Bezirk	2.0334	Verfasser	1.301
19	Konflikt	1.6628	Krieg	2.5024	Zeitangabe	0
20	Approximation	0	Annäherung	1.2788	Ziegelstein	0.4771
20	Augenoperation	0	Augenlasern	0		
21	Manipulation	0.8451	List	1.0414		
21	Stipulation	0	Übereinkommen	0.7782		
22	Glukose	0	Traubenzucker	0		
22	Hose	1.3617	Jeans	0		
23	Sarkasmus	0	Zynismus	0.6021		
23	Spasmus	0	Zuckung	0		

Table C.2 continued from previous page

Item#	Target	Freq	Lure	Freq
24	Theologie	0.9031	Religionswissenschaft	0
24	Typologie	0	Typenlehre	0
25	Tradition	1.6435	Brauch	1
25	Transposition	0	Transfer	0.301
26	Folie	0.6021	Anklebeband	0
26	Magnolie	0	Forsythie	0
27	Silikat	0	Bleicherde	0
27	Silikon	0	Plastik	1.1139
28	Melancholie	0.699	Trübsal	0
28	Melatonin	0	Kortisol	0
29	Gardine	0.6021	Vorhang	1.2553
29	Gelatine	0	Agar-Agar	0
30	Collagen	0	Gemälde	1.2304
30	Passagen	0.9542	Durchgang	1.0792
31	Ballett	0.9542	Jazztanz	1.4472
31	Falsett	0	Tenor	0.6021
32	Bambus	0	Eukalyptus	0
32	Jambus	0	Versfuß	0
33	Vitamin	0.9031	Mineralstoff	0
34	Exempel	0.4771	Beispiel	2.5575
34	Stempel	0.9031	Siegel	0.4771
35	Zoonose	0	Infektionskrankheit	0.301
35	Zuckerdose	0	Milchkännchen	0
36	Demos	0	Proteste	0
36	Memos	0	Sprachaufnahmen	0
37	Informatik	0	Computerwissenschaft	0
37	Pragmatik	0	Sachbezogenheit	0
38	Gnom	0.301	Wichtel	0
38	Strom	1.7404	Elektrizität	0.9031
39	Friede	2.0374	Ruhe	1.8692
39	Tide	0	Gezeiten	0
40	Harmonie	1	Einklang	1.0414
40	Pneumonie	0	Bronchitis	0
41	Palliativ	0	Schmerzmittel	0
41	Stativ	0	Ständer	0
42	Drainage	0	Entwässerung	0.301
42	Massage	0.699	Akupunktur	0
43	Wunder	1.6628	Sensation	1.0414
43	Zunder	0	Lunte	0
44	Franzose	1.699	Pariser	1.1761
44	Prognose	1.1761	Vorhersage	0.699
45	Demografie	0	Bevölkerungslehre	0
45	Demokratie	1.9912	Volksherrschaft	0

Table C.2 continued from previous page

Item#	Target	Freq	Lure	Freq
46	Suizide	0	Selbstmorde	1.3424
46	Sulfide	0	Mineral	0.6021
47	Histamin	0	Gewebehormon	0
48	Paddel	0	Ruder	0.6021
48	Quaddel	0	Schwellung	0
49	Plastik	1.1139	Kunststoff	0
49	Stochastik	0	Statistik	1.3979
50	Klausur	0	Prüfung	1.7076
50	Kreatur	0.301	Gestalt	1.8513
52	Friseur	0	Coiffeur	0
52	Malheur	0	Missgeschick	0.4771
53	Allüren	0	Marotten	0
53	Broschüren	0	Infoblätter	0
54	Präzedenz	0	Priorität	0.9542
54	Rezidenz	0	Villa	1.301
56	Antipathie	0	Widerwille	0
56	Emphatie	0	Einfühlungsvermögen	0
57	Parole	1.0792	Slogan	0.301
57	Pistole	1.1461	Gewehr	1.2553
60	Hamburg	2.7152	Bremen	1.699
60	Humbug	0	Unsinn	1.2553
61	Altruismus	0	Selbstlosigkeit	0
61	Autismus	0	Asperger	0
62	Jugend	2.1553	Pubertät	0
62	Tugend	1.1139	Moral	1.3222
63	Banane	0.7782	Ananas	0.301
63	Kleptomane	0	Dieb	1.2041
64	Million	2.5944	Vermögen	1.4472
64	Mitigation	0	Eindämmung	0.4771
65	Fiktion	0.9542	Einbildung	0.301
65	Kontradiktion	0	Widerspruch	1.699
66	Parlament	1.9191	Bundestag	0
66	Pergament	0	Schriftrolle	0
67	Bühne	1.7559	Podium	0.9542
67	Sühne	0.4771	Entschädigung	1.1139

Table C.3: Overview of the filler items with comprehension questions and correct responses

Item#	Filler	Answer
1	Berühmte Exponate werden im Museum ausgestellt.	
	Werden im Museum berühmte Exponate ausgestellt?	ja
2	Da Lina gerne Briefe verschickte, hatte sie eine große Briefmarkensammlung.	
	Verschickt Lina gerne Pakete?	nein
3	Das Armband konnte Marie nicht umtauschen, weil sie den Kassenbon weggeworfen hatte.	
	Konnte Marie das Armband umtauschen?	nein
4	Den Raum optisch aufwerten sollten die neuen Pflanzen.	
	Sollten die neuen Pflanzen den Raum optisch aufwerten?	ja
5	Der Marienkäfer krabbelte den Stiel entlang, um Blattläuse zu essen.	
	Wollte der Marienkäfer Blattläuse essen?	ja
6	Die Erstklässlerin war stolz auf ihr neues Mäppchen, in dem eine Menge Buntstifte waren.	
	Waren in dem neuen Mäppchen eine Menge Filzstifte?	nein
7	Die Ladung sicherte Felix mit einem Expander, damit nichts herunterfiel.	
	Sicherte Felix die Ladung mit einem Expander?	ja
8	Die richtige Buchseite musste Hannah suchen, weil sie ihr Lesezeichen verloren hatte.	
	Hatte Hannah ihr Buch verloren?	nein
9	Ein Geodreieck benutzten die Schüler, um die Winkel zu messen.	
	Haben die Schüler ein Geodreieck benutzt, um die Winkel zu messen?	ja
10	Ein neues Schuhgeschäft hat in dem Einkaufszentrum geöffnet.	
	Hat in dem Einkaufszentrum ein neues Eiscafé geöffnet?	nein
11	Eine E-Mail schrieb die Studentin der Professorin, um sich für den folgenden Tag krank zu melden.	
	Schrieb die Professorin der Studentin eine E-Mail?	nein
12	Einen Eisbären hat Tim gestern im Zoo gesehen.	
	Hat Tim gestern im Zoo einen Tiger gesehen?	nein
13	Einen großen Blumenstrauß schenkte Jan seiner Freundin am Valentinstag.	
	Schenkte Jan seiner Frau einen großen Blumenstrauß?	nein
14	Einen Muffin kaufte Elena sich an der Autobahnraststätte.	
	Kaufte Elena sich an der Autobahnraststätte ein Sandwich?	nein
15	Einen neuen Ausweis beantragte Nicole, da ihrer bereits seit drei Monaten abgelaufen war.	

Table C.3 continued from previous page

Item#	Filler	Answer
	Hat Nicole einen neuen Ausweis beantragt?	ja
16	Emmas neuer Staubsauger ging nach nur zwei Wochen Benutzung kaputt.	
	Ging Emmas Staubsauger nach zwei Wochen Benutzung kaputt?	ja
17	Frisches Brot bietet der Supermarkt jeden Montag an.	
	Bietet der Supermarkt jeden Montag frisches Brot an?	ja
18	Für ihren Auftritt auf dem Stadtfest kaufte sich Lara eine neue Gitarre.	
	Kaufte Lara für ihren Auftritt auf dem Stadtfest eine neue Geige?	nein
19	Für seinen Umzug kaufte sich Paul komplett neue Möbel.	
	Kaufte Paul sich für seinen Umzug gebrauchte Möbel?	nein
20	Im Schlossgarten ging Sarah mit ihrer Tante spazieren.	
	Ging Sarah mit ihrer Tante im Schlossgarten spazieren?	ja
21	In einem Restaurant arbeitete Tom neben seinem Studium, um sich Geld für einen Auslandsaufenthalt anzusparen.	
	Arbeitete Tom neben seinem Studium in einem Restaurant?	ja
22	In ihrem Poesiealbum sammelte Sophie ihre Lieblingsgedichte.	
	Sammelte Sophie ihre Lieblingsgedichte in ihrem Tagebuch?	nein
23	In kalten Jahreszeiten greift Sophie gerne zum Nahrungsergänzungsmittel Vitamin D.	
	Greift Sophie im Winter gerne zum Vitamin D?	ja
24	Jedes Jahr zu Ostern bekommt Leon von seiner Oma eine Geschenkkarte mit fünfzig Euro.	
	Bekommt Leon von seiner Oma zu Ostern eine Geschenkkarte mit 50 Euro?	ja
25	Lange Predigten hält der Pfarrer dieses Wochenende.	
	Hält der Pfarrer dieses Wochenende lange Predigten?	ja
26	Lara war noch nie am Meer und reiste nun zum ersten mal an die Nordsee.	
	Reiste Lara zum ersten Mal an den Pazifik?	nein
27	Maja wollte Spanisch lernen, also besuchte sie einen Sprachkurs.	
	Wollte Maja Spanisch lernen?	ja
28	Mit einem Tacker heftete die Assistentin die Unterlagen zusammen.	
	Heftete die Assistentin die Unterlagen mit einem Tacker zusammen?	ja
29	Mit Lampions und Lichterketten verschönerte Emily ihre Terasse.	

Table C.3 continued from previous page

Item#	Filler	Answer
	Verschönerte Emily ihre Terasse mit Lichterketten und Lampions?	ja
30	Mit Leckerlis belohnt Sarah ihren Hund, wenn er einen neuen Trick gelernt hat.	
	Belohnt Sarah ihre Katze mit Leckerlis?	nein
31	Nach einem anstregenden Arbeitstag, gönnte sich Anna eine Auszeit in der Badewanne.	
	Hatte Anna einen entspannten Arbeitstag?	nein
32	Nach einem Taschentuch fragte Chiara ihren Freund, weil sie erkältet war.	
	Fragte Chiara ihre Mutter nach einem Taschentuch?	nein
33	Nach Geld fragte Franz seinen Vater, um ins Kino zu gehen.	
	Fragte Franz seine Mutter nach Geld, um ins Kino zu gehen?	nein
34	Nach Lavendel riecht die Duftkerze, die Klara gekauft hat.	
	Riecht die Duftkerze, die Klara gekauft hat nach Lavendel?	ja
35	Nachdem die Waschmaschine 15 Jahre in Gebrauch war, gab sie den Geist auf.	
	Gab der Trockner nach 15 Jahren den Geist auf?	nein
36	Nina freute sich auf den Winter, weil sie dann schlittschuhlaufen konnte.	
	Freut sich Nina auf den Winter, weil sie dann schlittschuhlaufen kann?	ja
37	Verschiedene Aspekte des Klimawandels stellte Lisa in ihrer Präsentation vor.	
	Stellte Lisa in ihrer Präsentation verschiedene Aspekte der Tierhaltung vor?	nein
38	Während Lena auf ihren Bus wartete, suchte sie in Online-Shops nach neuen Stiefeln.	
	Suchte Lena nach einer neuen Bluse?	nein
39	Weil das Müsli Walnüsse und Früchte enthält, ist es gesund.	
	Enthält das Müsli Erdnüsse?	nein
40	Zum Geburtstag ihrer Tochter backte die Mutter eine Torte.	
	Backte die Mutter ihrer Tochter eine Torte zum Geburtstag?	ja
41	Zum Optiker ging Celine, da sie merkte, dass ihre Sehkraft nachgelassen hat.	
	Ging Celine zum Optiker, weil sie merkte, dass ihre Sehkraft nachgelassen hat?	ja

Appendix D

Experimental Instructions

In this Appendix we will give an overview of the instructions that were used in the online experiments. They are divided into the following sections. In Section D.1, we will present the general instructions that were used in multiple or all experiments. Section D.2 gives the instructions used in Experiments 1 and 2, as well as their stimuli creation. Finally, the instructions for the third experiment and related tasks are given in Section D.3.

D.1 General Instructions

Some parts of the instructions were used before multiple experiments. This is the case for the informed consent paragraph that participants had to agree with. For each audio experiment we made sure to test the participant's audio settings with a series of instructions and a test sound.

Consent

Rechtliche Hinweise: Das Experiment wird im Rahmen eines laufenden Forschungsprojektes an der Universität des Saarlandes durchgeführt. Bitte lesen Sie sich Folgendes sorgfältig durch, bevor Sie fortfahren. Um teilnehmen zu können, müssen Sie mindestens 18 Jahre alt sein. Ihre Teilnahme an der Studie ist freiwillig. Die Teilnahme an der Studie hat für Sie weder Nachteile noch Vorteile. Sie können die Beantwortung jeder der folgenden Fragen ablehnen. Sie können die Teilnahme jederzeit abbrechen,

ohne dass dies negative Konsequenzen für Sie hat. Alle Daten werden vor der Auswertung anonymisiert. Wenn Sie mit der Teilnahme einverstanden sind, klicken Sie bitte auf “Next”.

Audio Settings Test

Bitte machen Sie das Browserfenster nicht größer als 1 Blatt DinA 4 (also wie ein normales Blatt Druckerpapier).

Stellen Sie bitte sicher, dass Ihre Kopfhörer bzw. Lautsprecher mit dem Computer verbunden sind!

Klicken Sie auf "Weiter" um fortzufahren.

Stellen Sie nun die gewünschte Lautstärke ein. Drücken Sie anschließend auf das Lautsprechersymbol. Sie hören nun Hintergrundgeräusche. Wenn Sie diese nicht hören können, kontrollieren Sie bitte ihren Lautsprecheranschluss sowie die Lautstärke, und probieren Sie es noch einmal.

Zum Testen drücken Sie bitte auf das Lautsprechersymbol.

Das Rauschen wird gerade abgespielt.

Wenn Sie die Geräusche gehört haben, klicken Sie bitte auf "Weiter".

Wenn Sie die Hintergrundgeräusche nicht hören können, überprüfen Sie Ihre Lautsprechereinstellungen und stellen Sie sicher, dass Sie Google Chrome als Browser verwenden.

Wenn Sie auf "Weiter" klicken, hören Sie ein paar Übungssätze.

D.2 Experiment 1 & 2

Experiments 1 (see Chapter 3) and 2 (see Chapter 4) used the same experimental design and instructions, which are given below. Before we could run the experiments, we pretested the stimuli in a cloze ratings task. We also collected plausibility ratings for all items. These instructions are presented below. As mentioned above, before they started with an experiment, participants were asked to provide consent and when applicable, were asked to test their audio settings.

Cloze Test

Ihre Aufgabe wird es sein, Sätze zu vervollständigen.

Vervollständigen Sie die Satzanfänge, indem Sie **ein** Wort einfügen, von dem Sie denken, dass dieses am ehesten zum Satzanfang passt.

Beispiel: Er keltert den...

Mögliche Antwort: ... Wein.

Hierbei gibt es keine richtigen oder falschen Antworten; es geht nur um Ihre persönliche Einschätzung. Geben Sie Ihre Antworten möglichst so schnell es geht ab und denken Sie nicht allzu lange über die Sätze nach.

Insgesamt werden Ihnen 67 Sätze präsentiert, die Bearbeitung wird etwa 10 Minuten dauern.

Mit einem Klick auf 'Next' fängt der Test an.

Plausibility Test

Ihre Aufgabe besteht darin, die Ihnen im Folgenden präsentierten Sätze aufmerksam durchzulesen und sie hinsichtlich ihrer Natürlichkeit auf einer Skala von 1 bis 5 Sternen zu bewerten.

Dabei sind Sätze, die sich überhaupt nicht natürlich anhören, mit einem Stern zu bewerten und jene, die sich sehr natürlich anhören, mit fünf Sternen. Als "natürlich" gelten z.B. jene Sätze, die Ihnen plausibel vorkommen und sich für Sie als Muttersprachler/in gut anhören.

Insgesamt werden Ihnen 65 Sätze präsentiert, die Bearbeitung wird etwa 10 Minuten dauern.

Geben Sie Ihre Wertungen möglichst so schnell es geht ab und denken Sie nicht allzu lange über die Sätze nach.

Mit einem Klick auf 'Next' fängt der Test an.

Main Experiment

Herzlich Willkommen zu diesem Experiment!

Im Folgenden werden Ihnen Sätze präsentiert, die gut oder weniger gut verständlich sein können. Hören Sie genau zu, was gesagt wird.

Nachdem der Satz dargeboten wurde, möchte ich Sie bitten, nur das letzte Wort des Satzes, den Sie gehört haben, auf der PC-Tastatur einzugeben.

Beispiel: Wenn Sie den Satz " Wenn sie einen Ausritt machen, reiten sie am liebsten durch den Wald." hören, tippen Sie bitte "Wald" auf der Tastatur ein. Anschließend werden Sie noch gefragt, wie sicher Sie sich bei Ihrer Antwort waren. Bei Ihrer Antwort können Sie aus vier Optionen auswählen (von unsicher/geraten bis absolut sicher).

Zunächst können Sie sich im Übungsdurchgang mit der Aufgabe vertraut machen. Sollten Sie sich bei Ihrer Antwort unsicher sein, so hören Sie auf Ihr Bauchgefühl und geben die Antwort, die Ihnen als erstes einfällt. Klicken Sie auf "Weiter" um fortzufahren.

D.3 Experiment 3

For the third experiment (see Chapter 5), we conducted several pretests before running the main experiment. These were a pretest to determine the appropriate background noise level, and a test to check whether participants were familiar with the items. The instructions of these two tasks are given below.

Finally, the main experiment consisted of several tasks as well: the listening task for exposure, the memory test to assess participants' recognition memory, and three individual differences measures: the Raven's Progressive Matrices; a backwards digit span task, and a vocabulary test. All instructions are presented in Section D3.1. The tasks followed upon each other in the experiment.

Noise Level Pretest

Lieber Teilnehmer,

Danke für Ihre Teilnahme an dieser Studie!

In diesem Test werden Sie sich Sätze anhören. In den Aufnahmen sind Hintergrundgeräusche zu hören. Hören Sie genau zu, was gesagt wird. Nachdem der Satz dargeboten wurde, möchten wir Sie bitten, den gehörte Satz einzugeben. Anschließend werden Sie noch gefragt, wie gut Sie alle Wörter des Satzes verstehen konnten. Ihre Antwort können Sie dann auf einer Skala von 1 bis 5 Sternen geben. Wenn Sie keines der Wörter verstehen konnten, bewerten Sie den Satz bitte mit einem

Stern. Wenn Sie alle Wörter verstehen konnten, bewerten Sie den Satz bitte mit fünf Sternen.

Sie werden insgesamt 30 Sätze hören. Zuletzt muss noch ein kurzer Fragebogen ausgefüllt werden. Insgesamt wird das Experiment 20 Minuten dauern.

Word Knowledge Pretest

Herzlich Willkommen zum Experiment!

In diesem Experiment werden Sie 20 Wörter sehen und wir bitten Sie, diese Wörter zu definieren und zu bewerten wie gut Sie das Wort kennen.

Falls Sie das Wort nicht kennen, schreiben Sie das in das Textfeld. Das ist gar kein Problem!

Das Experiment beginnt, wenn Sie „Next“ klicken.

D3.1 Main Experiment

Listening Task

Herzlich Willkommen zum Experiment!

Im folgenden Experiment werden Sie sich Sätze anhören. Mal sind in den Aufnahmen Hintergrundgeräusche zu hören, mal nicht. Im Anschluss an jeden Satz werden Sie gefragt, wie gut Sie alle Wörter des Satzes verstehen konnten. Ihre Antwort können Sie dann auf einer 5-Punkte-Skala geben. Wenn Sie keines der Wörter verstehen konnten, bewerten Sie den Satz bitte mit einem Stern. Wenn Sie alle Wörter verstehen konnten, bewerten Sie den Satz bitte mit fünf Sternen. Es ist wichtig, dass Sie sich die Sätze sehr aufmerksam anhören, da Sie manchmal Fragen zu einem Satz beantworten müssen, den Sie zuvor gehört haben.

Sie werden insgesamt 60 Sätze hören, was etwa 20 Minuten in Anspruch nehmen wird. Nachdem Sie sich alle Sätze angehört haben, werden Sie noch andere Aufgaben durchführen, die Ihren Wortschatz, Ihren IQ und Ihr Arbeitsgedächtnis testen sollen. Zuletzt muss noch ein kurzer Fragebogen ausgefüllt werden. Insgesamt wird das Experiment 45 Minuten dauern.

Bitte vergewissern Sie sich, dass Ihr Ton funktioniert, und verwenden Sie nach Möglichkeit Kopfhörer. Das Experiment funktioniert am besten in Google Chrome, und leider nicht in Safari.

Auf der nächsten Seite können Sie testen, ob Ihr Ton funktioniert. Bitte beachten Sie, dass die Sätze automatisch abgespielt werden und nur einmal angehört werden können.

Backwards Digit Span

In diesem Experiment werden Ihnen Ziffernfolgen vorgelegt, die Sie in umgekehrter Reihenfolge nacheinander eingeben müssen. Wenn Ihnen beispielsweise die Ziffernfolge "123" angezeigt wurde, geben Sie bitte "321" in das Textfeld ein, ohne Leerzeichen oder Satzzeichen dazwischen. Wenn Sie sich an eine Ziffer nicht mehr erinnern können, geben Sie stattdessen ein Fragezeichen ein, also z.B. "3?1" für das obige Beispiel.

Während des Experiments wird die Länge der präsentierten Sequenzen zunehmen.

Für unsere Forschung ist es sehr wichtig, dass Sie das Auswendiglernen und Umkehren im Kopf machen und nichts aufschreiben. Dies ist eine anspruchsvolle Aufgabe und wir erwarten nicht, dass Sie alles richtig machen.

Sie werden zwei Übungssequenzen machen und dann beginnt die eigentliche Aufgabe. Insgesamt sollte das Experiment nicht länger als 5 Minuten dauern.

Wenn Sie fertig sind, klicken Sie bitte auf "Weiter", um fortzufahren.

Übungsfrage: Bitte geben Sie die Ziffern in umgekehrter Reihenfolge ein.

Nun beginnt das eigentliche Experiment. Bitte schreiben Sie nichts auf und seien Sie nicht frustriert, wenn Sie nicht alles richtig machen – das erwarten wir auch nicht!

Raven's Progressive Matrices

In diesem Test werden Sie ein Bild mit einem Muster sehen. Das Muster ist unvollständig: Es fehlt ein Teil des Puzzles.

Ihre Aufgabe besteht darin, aus einer Reihe von sechs bis acht Optionen das Puzzleteil auszuwählen, dass zum Muster des Bildes passt. Sie treffen Ihre Auswahl, indem Sie auf das Puzzleteil klicken, das Ihrer Meinung nach zum Muster passt. Die Muster werden im Laufe der Studie immer schwieriger.

Hinweis: Manchmal dauert es einige Sekunden, bis alle Bilder richtig angezeigt werden.

Diese Studie umfasst 12 Elemente und dauert etwa 5 Minuten.

Damit Sie die Dauer des Tests besser einschätzen können, haben wir eine Uhr am oberen Rand des Bildschirms angebracht. Allerdings gibt es keine negativen Folgen, sollten Sie bei diesem Test die Zeit überschreiten.

Memory task

Wir möchten nun Ihr Gedächtnis mit einer kurzen Aufgabe testen. Sie sehen ein Wort und werden gebeten anzugeben, ob Sie dieses Wort in den Sätzen der ersten Aufgabe gehört haben. Außerdem müssen Sie auf einer Skala von "sehr unsicher" bis "sehr zuversichtlich" einschätzen, wie sicher Sie sich sind. Sie können beginnen, indem Sie auf "Weiter" klicken.

Vocabulary Test

In diesem Experiment werden Sie 55 Sätze sehen, die Sie mit einer von fünf Antwortmöglichkeiten vervollständigen müssen. Es ist immer nur eine Antwort richtig. Einige der Wörter können schwer verständlich sein. Wenn Sie die Antwort nicht wissen, vertrauen Sie ihrer Intuition; oft wissen Sie unterbewusst mehr, als Sie gedacht hätten. Bitte schlagen Sie keine Wortbedeutungen nach.

Dies ist der letzte Teil des Experiments. Vielen Dank für Ihre Teilnahme!